Studies in Systems, Decision and Control

Volume 109

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Systems, Decision and Control" (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control- quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/13304

Kyandoghere Kyamakya · Wolfgang Mathis
Ruedi Stoop · Jean Chamberlain Chedjou
Zhong Li
Editors

Recent Advances in Nonlinear Dynamics and Synchronization

With Selected Applications in Electrical Engineering, Neurocomputing, and Transportation

Springer

Editors

Kyandoghere Kyamakya
Institute of Smart Systems Technologies
Alpen-Adria-Universität Klagenfurt
Klagenfurt
Austria

Jean Chamberlain Chedjou
Institute of Smart Systems Technologies
Alpen-Adria Universität Klagenfurt
Klagenfurt
Austria

Wolfgang Mathis
Institut für Theoretische Elektrotechnik
Leibniz Universität Hannover
Hannover
Germany

Zhong Li
Philipp-Reis-Gebäude
Fernuniversität Hagen
Hagen
Germany

Ruedi Stoop
Institute of Neuroinformatics
University of Zürich and ETH Zürich
Zürich
Switzerland

ISSN 2198-4182 ISSN 2198-4190 (electronic)
Studies in Systems, Decision and Control
ISBN 978-3-319-58995-4 ISBN 978-3-319-58996-1 (eBook)
DOI 10.1007/978-3-319-58996-1

Library of Congress Control Number: 2017941481

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Since several decades nonlinear dynamics has been and remains a very hot research field of high relevance in many areas of science and technology and it is a precious source of inspiration for cutting-edge novel ideas. The contributions of this book do analyze a series of interesting issues and aspects related to nonlinear dynamics and/or synchronization in the context of either practical or theoretical applications in selected areas, namely in theoretical electrical engineering, signal processing & communications engineering, neuro-computing, transportation, and computational intelligence. Besides some additional external contributions, most of the chapters of this book are extended versions of either papers or plenary talks presented at the 4th International Workshop on Nonlinear Dynamics and Synchronization (INDS'2015) held July 30th–31st, 2015 in Klagenfurt, Austria (at Alpen-Adria-Universität Klagenfurt).

This book is structured in 5 parts covering a total of 16 chapters:

- Part I: Nonlinear dynamics—Fundamentals, with 3 chapters
- Part II: Nonlinear dynamics—Selected applications, with 2 chapters
- Part III: Transportation, with 4 chapters
- Part IV: Signal processing & communications engineering, with 3 chapters
- Part V: Computational Intelligence, with 4 chapters

Part I "Nonlinear Dynamics—Fundamentals", contains three contributions. The first chapter of Part I, "On the Construction of Dissipative Polynomial Nambu Systems with Limit Cycles", by Richard Mathis and Wolfgang Mathis, does study nonlinear Nambu systems with canonical dissipation in four dimensions where prescribed limit cycles arise. Hereby, some possibilities are discussed to apply concepts from the geometry of 4-dimensional Euclidean spaces in order to construct a desired intersection by a set of planes and surfaces. An analysis concept based on CD Nambu systems for calculating non-isolated zeros of nonlinear equations is introduced. The discussed considerations are illustrated by means of two examples of canonical dissipative (CD) Nambu systems where limit cycles occur.

In the second chapter of Part I, "On the Dynamics of Chaotic Systems with Multiple Attractors: A case Study", by J. Kengne, A. Nguomkam Negou, D. Tchiotsop, V. Kamdoum Tamba and G.H. Kom, the dynamics of chaotic systems with multiple co-existing attractors is analyzed by using the well-known Newton–Leipnik system as prototype. In the parameters' space, regions of multistability (where the system exhibits up to four disconnected attractors) are depicted by performing both forward and backward bifurcation analysis of the model. A suitable electrical circuit (i.e. analog simulator) is designed and used for the experimental investigations.

Then, in the third and last chapter of Part I, "Multivaluedness Aspects in Self-Organization, Complexity and Computations Investigations by Strong Anticipation", by Alexander Makarenko, new examples of discrete dynamical systems with anticipation are considered. Hereby, the mathematical formulation of problems, possible analytical formulas for solutions and numerical examples of presumable solutions are proposed. As well-known, one of the most interesting properties in such systems is the presumable multivaluedness of the solutions. Overall, it can be considered from the point of view of dynamical chaos and complex behavior. The chapter presents examples of periodic and complex solutions, attractor's properties and presumable applications in self-organization.

Part II "Nonlinear Dynamics—Selected Applications", contains two contributions. In the first chapter of Part II, "Nonlinear Modeling of Continuous-Wave Spin Detection Using Oscillator-Based ESR-on-a-Chip Sensors", by Jens Anders, an advanced nonlinear energy-based modeling of LC tank oscillators used as sensors for ensembles of electron or nuclear spins is presented. The chapter starts with a description of the experimental setup of the oscillator-based spin detection approach, which is somewhat different from that used for conventional resonator-based detection. The interaction between the nonlinear electrical oscillator and the spin ensemble is modeled using the solution of the Bloch equation in the steady-state, which models the dynamics of the spin ensemble, and the magnetic energy associated with the inductor of the LC tank oscillator. The final model of the LC tank oscillator is used to find analytical expressions for the limit of detection of frequency-sensitive oscillator-based spin detectors. Finally, experimental results from a prototype realization are used to validate the accuracy of the derived signal and noise models.

Then in the second chapter of Part II, "Effect of Non-Linearity and Boiler Dynamics in Automatic Generation Control of Multi-area Thermal Power System with Proportional-Integral-Derivative and Ant Colony Optimization Technique", by K. Jagatheesan, B. Anand, K. Baskaran, N. Dey, A.S. Ashour, and V.E. Balas, the Automatic Generation Control (AGC) of a multi-area interconnected power system is presented. The occurrence of sudden load disturbance in the interconnected power generating unit does significantly affect both system performance (consistency in system frequency and voltage) and system stability. In order to moderate the last mentioned negative effects, a PID controller is introduced as a secondary controller. An Artificial Intelligence (AI) based Ant Colony Optimization (ACO) technique is considered for tuning of the controller parameters. Furthermore, both non-linearity and boiler dynamics effects are considered to evaluate the performance of the investigated power system.

Part III "Transportation", contains four contributions. The first chapter of Part III, "A Review of Traffic Light Control Systems and Introduction of a Control Concept Based on Coupled Nonlinear Oscillators", by Jean Chamberlain Chedjou and Kyandoghere Kyamakya, presents an in-depth overview of the state-of-the- art on traffic light control and optimization. Several traditional traffic control and simulation methods, concepts and tools are described whereby their related pros and cons are discussed. Further, the chapter develops a system of coupled nonlinear oscillators, which is used for traffic light control and optimization both at isolated junctions (i.e. local control) and in a network of coupled traffic junctions (i.e. area control). The system developed is viewed as a modified version of the self-organized Kuramoto model for traffic light control due to some important features, which are common to both systems (i.e. the traditional Kuramoto model and the new concept developed).

In the second chapter of Part III, "Neural-Network-Based Calibration of Macroscopic Traffic Flow Models", by Nkiediel Alain Akwir, Jean Chamberlain Chedjou, and Kyandoghere Kyamakya, a neural network based calibration concept of macroscopic traffic flow models expressed in the form of nonlinear partial differential equations (PDEs) is proposed. The calibration scheme developed aims at improving both accuracy and stability of the nonlinear PDE models in order to make them becoming more realistic. The calibration scheme is used to dynamically optimize all outputs of the nonlinear "PDE"-model in order to obtain a realistic set of parameters, which can be later used by the PDE-model to describe the real/realistic dynamics of traffic flow.

In the third chapter of Part III, "Travelers in the Second Modernity: Where Technological and Social Dynamic Complexity Meet Each Other", by Oana Mitrea, a comprehensive essay is presented which mainly draws its inspiration from the findings of both systems theory and sociology about non-linearity and individualization with particular application to mobility systems. Its first part analyzes the societal system of the second modernity as an open system with high non-linearity. The second part applies the reflections from the construction of hypotheses about the usage dynamics of an intelligent concept of social interaction on the move.

Finally, in the fourth and last chapter of Part III, "COMPRAM Assessment and System Dynamics Modeling and Simulation of Car-Following Model for Degraded Roads", by A.K. Kayisu, M.K. Joseph, and K. Kyamakya, the complex societal problem related to the consequence on traffic management of potholes in road in analyzed. Potholes in roads is a complex societal reality in developing countries and leads to situations like congestion, chaotic driving and an acceleration of road degradation. This complex phenomenon—traffic congestion in the town of Kinshasa (in DR Congo) linked to degraded roads is closely analyzed with the help of the COMPRAM methodology. The results of the quantitative simulations demonstrate that in the presence of the pothole at microscopic level, speed and travel time are negatively impacting road capacity at macroscopic level.

Part IV "Signal Processing & Communications Engineering", contains three contributions. The first chapter of Part IV, "Design of a Chaotic Pulse-Position Modulation Circuit", by Junying Niu, Zhong Li, Yuhong Song, and Wolfgang A. Halang, addresses an analogue chaotic modulation circuit based on Chua's circuit, which is proposed to generate a chaotic pulse-position signal. The circuit is designed with the standard electronic components, and the parameters of the generated signals, including the pulse period, the modulation range of the pulse-position, even the probability distribution of the pulse-position, can be adjusted flexibly.

The second chapter of Part IV, "Chaos-Based Digital Communication Systems with Low Data-Rate Wireless Applications", by Nguyen Xuan Quyen and Kyandoghere Kyamakya, presents a study on the modeling and performance evaluation of chaos-based coherent and non-coherent systems, i.e., chaotic direct-sequence code-division multiple-access (CDS-CDMA) and differential chaos-shift keying (DCSK), for low data-rate applications in wireless communications. Simulated performance results are shown and compared with the corresponding estimated ones, where the effects of the ratio Eb/N0, spreading factor, number of users, sample rate, and the number of transmission paths on the BER are fully evaluated. Overall, the obtained results do show that the low-rate chaos-based systems can exploit the multipath nature of wireless channels in order to improve their BER performances.

And the third chapter of Part IV, "Nonlinear Programming Approach for Design of High Performance Sigma-Delta Modulators", by Valeri Mladenov and Georgi Tsenov, presents a nonlinear programming approach for the design of a third order Sigma-Delta modulator with respect to maximization of the signal to noise ratio, taking into account the modulator's stability. The results are compared with optimal third order modulator design provided by DStoolbox.

Part V "Computational Intelligence", contains four contributions. In the first chapter of Part V, "Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review", by Mouhannad Ali, Ahmad Haj Mosa, Fadi Al Machot, and Kyandoghere Kyamakya, a comprehensive review of the state-of-the-art methodologies for emotion recognition based on physiological changes and speech is presented. In particular, one investigate the potential of physiological signals and driver's speech for emotion recognition and their requirements for ADAS (advanced driver assistance systems). All steps of an automatic recognition system are explained: emotion elicitation, data preprocessing such as noise and artifacts removal, the features extraction and selection, and finally classification.

In the second chapter of Part V, "A Hybrid Reasoning Approach for Activity Recognition Based on Answer Set Programming and Dempster-Shafer Theory", by Fadi Al Machot, Heinrich C. Mayr and Suneth Ranasinghe, one discusses a promising approach for multi-sensor based activity recognition in smart homes. The research is originated in the domain of Active and Assisted Living, particularly in the field of supporting people in mastering their daily life activities. The proposed fusion concept combines Answer Set Programming and Dempster-Shafer theory. In order to check the overall performance, this approach has been tested by using the HBMS dataset on an embedded platform.

In the third chapter of Part V, "Estimation of Infection Force of Hepatitis C Virus Among Drug Users in France", by Selain Kasereka, Yann Le Strat and Lucie Léon, the spread of diseases as a dynamic and complex phenomenon is analyzed. The aim of this work is to estimate the force of infection of hepatitis C from two national cross-sectional epidemiological surveys conducted in 2004 and 2011 by The French Institute for Public Health Surveillance and its partners in a drug user population in France. The force of infection has been modeled according to a SIS (Susceptible-Infected-Susceptible) compartmental model using differential equations (ODE) and as being a function of the derivative function of the prevalence depending on age, time, HIV serological status and having injected at least once in their life or not, from 2000 to 2010. Overall, this work provides guidance for researchers to compare several cross-sectional epidemiological surveys among drug users and proposes an alternative method to estimate the force of infection among drug users from cross-sectional surveys in the absence of a cohort.

Finally, the fourth and last chapter of Part V, "Neurocomputing-Based Matrix Inversion: A critical Review of the Related State-of-the-Art", by Vahid Tavakkoli, Jean Chamberlain Chedjou, and Kyandoghere Kyamakya, does provide a comprehensive overview of both traditional and neuro-computing based methods for solving the matrix inversion problem (i.e.: analytical, heuristics, dynamic system based methods, etc.). These methods are compared to each other based on some important criteria like convergence, parallelizability/scalability, accuracy and applicability to time-varying matrices. Finally, we propose a new concept based on Neurocomputing for solving the matrix inversion problem. The main advantage of this proposed concept is the possibility of efficiently fulfilling all criteria at stake.

Guest Editors

Contents

Part IV Signal processing & Communications Engineering

Part V Computational Intelligence

Part I
Nonlinear Dynamics—Fundamentals

On the Construction of Dissipative Polynomial Nambu Systems with Limit Cycles

Richard Mathis and Wolfgang Mathis

Abstract In this chapter we study nonlinear Nambu systems with canonical dissipation in four dimensions in which prescribed limit cycles arise. For this purpose we discuss some possibilities to apply concepts from the geometry of four-dimensional Euclidean spaces in order to construct a desired intersection by a set of planes and surfaces. Since scalar Nambu functions are needed for the construction of Nambu systems, the relationship between these functions and hypersurfaces will be discussed. We illustrate our considerations by means of two examples of canonical dissipative (CD) Nambu systems in which limit cycles occur. Whereas we considered the synthesis problem of limit cycle circuits with the CD Nambu approach previously, we introduce in this paper an analysis concept based on CD Nambu systems for calculating nonisolated zeros of nonlinear equations.

1 Introduction

Since Nambu [1] generalized Hamiltonian mechanics in 1973, his concept as received increasing interest. In contrast to Hamiltonian systems, the state (phase) space can be odd, so that a much larger class of dynamical systems can be included in our considerations. In the following years, the properties of this new type of mechanics and some applications in classical and quantum mechanics have been studied by many authors, e.g., [2–7]. It turned out that not only the rigid rotator, which was considered by Nambu, but also the equations of motion of other physical systems can be reformulated as Nambu systems [8].

More recently, Nambu systems have discussed in which different kinds of dissipation terms are included. In 2008, Bihlo [9] introduced a metriplectic generalization of Nambu systems in which the gradient part describes the dissipation. Then in 2010,

R. Mathis (✉)
Technical University of Braunschweig, Braunschweig, Germany
e-mail: r.mathis@tu-bs.de

W. Mathis
TET, Leibniz Universität Hannover, Hanover, Germany
e-mail: mathis@tet.uni-hannover.de

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_1

Axenides and Floratos [2, 10] decomposed the Lorenz equation into a Nambu part and a linear dissipation part [11]. More general Nambu systems with linear dissipation were considered later on by Roupas [12]. A decomposition of the Lorenz equation into a conservative and a dissipative part without the Nambu framework was discussed by Haken and Wunderlin [13] in 1977. In 2010, Frank introduced canonical dissipation [14] into Nambu mechanics [15] and used this concept in biophysical applications [16, 17]. Then Yamaleev's oscillator was investigated by Mongkolsakulvong, Chaikhan, and Frank [18] in great detail, and also a stochastic part was added. Motivated by the Frank group's paper on Yamaleev's oscillator, Mathis, Stahl, and Mathis [19] extended previous results of the Hannover group; cf. Thiessen [20] and Mathis, et al. [21] about canonical dissipative (CD) systems to CD Nambu systems. In their paper [19], they studied the Lorenz system from [10] and added canonical dissipation; cf. also Mathis, Mathis [22]. In both the CD Yamaleev and CD Lorenz systems, the limit cycle behavior [23] was the main objective of research. In particular, the papers of [19, 22] were dedicated to a new approach to the design of electrical oscillators in which a limit cycle can be prescribed. In a recent paper [24] we gave a sketch of a system of differential equations of order 4 with a limit cycle that is constructed by the framework of CD Nambu mechanics.

In this article we will give a comprehensive overview of Nambu systems with canonical dissipation and illustrate it with some examples. We organized the article in the following manner: In Sect. 2, some mathematical details of CD Nambu systems are presented. In Sect. 3, four-dimensional CD Nambu systems with limit cycles are considered. For this purpose the geometric construction of intersections of surfaces in four-dimensional Euclidean spaces are discussed in Sect. 3.1, where we restrict ourselves to polynomials in four variables in order to use the symbolic manipulator Mathematica; cf., e.g., [25]. In Sect. 3.2, an example of a CD Nambu system with a single limit cycle is presented, and we discuss some dynamical properties of this system. In Sect. 3.3, a CD Nambu system with several limit cycles is presented on the basis of a Clebsch cubic. It is shown in Sects. 2 and 3 not only that the CD Nambu approach can be applied to the synthesis of circuits and systems with limit cycles, cf. our previous paper [22]), but also that it is useful for the analysis of nonlinear systems with nonisolated zero sets. The paper ends with a summary.

2 Canonical Dissipative Nambu Systems

Basically, the vector field of an n-dimensional Nambu system is constructed by $n-1$ Nambu functions $H_i : \mathbb{R}^n \to \mathbb{R}$ with $i = 1, \ldots, n-1$. Using these functions, we obtain the equations of motion for the n-dimensional Nambu system, cf. Nambu [1],

$$\begin{aligned}\frac{dx_i}{dt} &= \sum_{j_1,\ldots,j_{n-1}} \epsilon_{i,j_1,\ldots,j_{n-1}} \frac{\partial H_1}{\partial x_{j_1}} \cdots \frac{\partial H_{n-1}}{\partial x_{j_{n-1}}} =: \\ &=: \frac{\partial\left(x_i, H_1, \ldots, H_{n-1}\right)}{\partial\left(x_1, \ldots, x_n\right)},\end{aligned} \tag{1}$$

for $i = 1, \ldots, n$, where $\epsilon_{i,j_1,\ldots,j_{n-1}}$ is the n-dimensional Levi-Civita tensor.

From a geometric point of view, Nambu functions can be characterized by their level surfaces $H_i(x_1, \ldots, x_n) = E_i$ $(i = 1, \ldots, n)$ in the n-dimensional space $\mathbb{R}^n$, which can be interpreted as hypersurfaces. If the level values E_i are determined by the initial conditions of the system, a solution of (1) runs in the intersection set of these level surfaces. Furthermore, Steeb [6] has proved that all polynomial functions are invariants of the motion.

In the following, we consider only polynomial functions for which the level functions are algebraic sets. Using results from real algebraic geometry [26, 27], we know that the intersection of algebraic sets is also algebraic.

In order to obtain an asymptotic behavior that is independent on the initial values, we note that this property corresponds to a so-called limit cycle, where nonlinear dissipation has to be in the system; cf. [23]. It was shown by Ebeling and Sokolov [14] that the concept of canonical dissipation is an elegant way to introduce dissipation in a Hamilton mechanics. Inspired by Frank [15], we use canonical dissipation for constructing Nambu systems with dissipation. We have already shown [19, 22] that this concept can be useful for the design of electronic oscillator circuits.

Based on the Nambu functions $H_1, H_2, \ldots, H_{n-1}$, Frank [15] constructed generalized canonical dissipation terms for $i = 1, \ldots, n$:

$$G_i(H_1, \ldots, H_{n-1}) := \sum_{j=1}^{n-1} \frac{\partial H_j}{\partial x_i} \left(H_j - E_j\right), \tag{2}$$

where the E_j are the values of the energy levels. Using these terms and combining them with Nambu equations (1), we obtain the equations of motion for a more general Nambu system with canonical dissipation:

$$\frac{dx_i}{dt} = \frac{\partial\,(x_i, H_1, \ldots, H_{n-1})}{\partial\,(x_1, \ldots, x_n)} - \varepsilon G_i(H_1, \ldots, H_{n-1}), \tag{3}$$

where $i = 1, \ldots, n$ and ε is the damping coefficient.

We will show by means of two representative examples that the canonical dissipative terms $G_i(H_1, \ldots, H_{n-1})$ go asymptotically to zero, and therefore the dynamics corresponds to the dynamics of the Nambu system without dissipation.

In summary, we are able to construct CD Nambu systems in which each trajectory that starts near the intersection of the hypersurfaces of Nambu functions converges to this intersection set. If the intersection set is a one-dimensional closed curve, we obtain a dynamical system with a limit cycle. Furthermore, if the level surfaces are algebraic sets, then the limit cycle is algebraic. In the next section we consider a CD Nambu system in four dimensions and study its asymptotic behavior.

3 Four-Dimensional Canonical Dissipative Nambu Systems with Limit Cycles

3.1 Construction of Closed Curves in Four Dimensions

Since the level surfaces of Nambu functions in a 4-dimensional space are 3-dimensional hypersurfaces, we need more understanding of geometry in four dimensions. At first, we restrict our attention to hyperplanes in a 4-dimensional real vector space $\mathbb{R}^4$. In this case, a basis of four linear independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_4$ exists such that an arbitrary vector $\mathbf{x} \in \mathbb{R}^4$ can be decomposed into $\mathbf{x} = \sum_{i=1}^{4} x_i \, \mathbf{b}_i$, where $x_i \in \mathbb{R}$ $(i = 1, \ldots, 4)$. Therefore, each $\mathbf{x}$ can be represented by a 4-tuple of real numbers x_i in $\mathbb{R}^4$. Curves and surfaces in $\mathbb{R}^4$ can be defined as 1-dimensional and 2-dimensional objects using one real parameter ξ and a two real parameters (ξ, η), respectively, such that $\mathbf{x}(\xi)$ for $\xi \in T \subset \mathbb{R}$ as well as $\mathbf{x}(\xi, \eta)$ for $(\xi, \eta) \in D \subset \mathbb{R}^2$. A 1-dimensional straight line in $\mathbb{R}^4$ is defined by $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^4)$

$$\mathbf{x}(\xi) = \mathbf{x}_0 + \xi \, \mathbf{x}_1, \quad \xi \in T, \tag{4}$$

whereas a 2-dimensional plane can be defined by

$$\mathbf{x}(\xi, \eta) = \mathbf{x}_0 + \xi \, \mathbf{x}_1 + \eta \, \mathbf{x}_2, \quad (\xi, \eta) \in D. \tag{5}$$

Obviously, 3-dimensional objects $\mathbf{x}(\xi, \eta, \vartheta)$ can be defined with three real parameters $(\xi, \eta, \vartheta) \in D \subset \mathbb{R}^3$ that can be identified with hypersurfaces in $\mathbb{R}^4$. Algebraically, a hypersurface corresponds to the solution set of a scalar equation $H(\mathbf{x}) = 0$ with $H : \mathbb{R}^4 \to \mathbb{R}$. For example, a hyperplane is defined by $H(\mathbf{x}) = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 = r$ with $a_1, \ldots, a_4, r \in \mathbb{R}$.

In order to measure distances, $\mathbb{R}^4$ has to be endowed with a Euclidean scalar product $\langle \cdot, \cdot \rangle$.

The definition of the Euclidean vector space $\mathbb{R}^4$ is straightforward, but the spatial positioning of the above-defined objects and their intersections is slightly more involved. For example, it seems strange that two (2-dimensional) planes intersect only in a point. Moreover, a plane intersects a (3-dimensional) hyperplane in a straight line; that is, a (2-dimensional) knife is not useful for cutting a loaf of bread in four dimensions; cf., e.g., Weizenböck [28]. In n-dimensional spaces $\mathbb{R}^n$, the dimension s of the intersection set can be calculated by the following formula:

$$s = l + k - n, \tag{6}$$

where l and k are the dimensions of the hyperplanes; cf. Aleksandrov et al. [29]. That is, if two hyperplanes have one point in common and $l + k \geq n$, then they intersect in a hyperplanes of dimension not less than s.

To gain further insight into the problem, we discuss it from the algebraic point of view. If two hyperplanes intersect, the intersection is a 2-dimensional plane in generic cases. This can be proved using the following algebraic descriptions of hyperplanes in the coordinates $x_1, \dots, x_4$:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = r_1, \tag{7}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = r_2, \tag{8}$$

where the solution space is a 2-dimensional affine subspace of $\mathbb{R}^4$, a plane in general position, when the coefficient matrix is of full rank. The hyperplanes are parallel if the intersection plane is at infinity. It can be verified in the same manner that the intersection of two planes consists of a point $x \in \mathbb{R}^4$ in generic cases. However, if the planes are included in the same hyperplane, then the intersection is a straight line in generic cases. Furthermore, the intersection of a plane and a hyperplane and three hyperplanes is generically a straight line in generic cases.

If the surfaces in $\mathbb{R}^4$ are more general and nonlinear, it is not easy to determine whether the intersection of three surfaces corresponds to a one-dimensional (closed) curve. A hypersurface in $\mathbb{R}^4$ can be defined implicitly by a scalar function $f : \mathbb{R}^4 \to \mathbb{R}$ in the variables $(x, y, z, w)^T \in \mathbb{R}^4$. Obviously, the sets of zeros of $H_i(x, y, z, w) - E_i = 0$, $i = 1, 2, 3$, correspond to hypersurfaces associated with the Nambu functions H_i $(i = 1, 2, 3)$. The common set of zeros corresponds geometrically to the desired intersection of the level sets of the Nambu functions. Since possibly complex zeros arise, algebraic geometry considers mostly complex manifolds instead of real manifolds, in which the number of zeros does not vary if the parameters are changed; cf. Hulek [26]. The results from real algebraic geometry [30] are rare, so that there is no systematic way to construct intersections of level surface of three Nambu functions so that we obtain a closed curve in $\mathbb{R}^4$ or higher dimensions.

In the following we discuss two examples of CD Nambu systems in $\mathbb{R}^4$ defined by three suitable Nambu functions H_1, H_2, H_3. Gürses et al. [31] defined an autonomous system as super integrable if it has $n - 1$ functionally independent first integrals (constants of motion), and Steeb [6] proved that polynomial Nambu functions are integrals of motion. Since polynomial Nambu functions can be treated more simply than nonpolynomial functions, if a symbol manipulator (e.g., Mathematica) is used, we restrict attention in the following to Nambu systems in $\mathbb{R}^4$ with polynomial Nambu functions.

3.2 Canonical Dissipative Nambu System with a Single Limit Cycle

In this section we construct a CD Nambu system with a single limit cycle in four dimensions and discuss its properties. For this purpose we need three Nambu functions H_1, H_2, H_3 whose corresponding hypersurfaces intersect in a single closed curve. As mentioned above, we restrict attention to CD Nambu systems in $\mathbb{R}^4$ with polynomial Nambu functions. We choose the following Nambu functions:

$$H_1(x, y, z, w) = \frac{1}{2}\left(\frac{w^2}{2} + x^2 + y^2 + z^2\right) - 1, \tag{9}$$

$$H_2(x, y, z, w) = wx^2 + wy^2 + z^2, \tag{10}$$

$$H_3(x, y, z, w) = x + y - 1, \tag{11}$$

where the hypersurfaces are defined by the implicit equations $H_i(x, y, z, w) - E_i = 0, i = 1, 2, 3$. The intersection set is not empty, since one point in the intersection is $\left(\frac{1}{2} + \frac{\sqrt{3}}{2}, \frac{1}{2} - \frac{\sqrt{3}}{2}, 0, 0\right)$, and with formula (6), the local generic dimension can be calculated as $s = 3 + 3 - 4 = 2$ for the intersection of two linearized hypersurfaces, and as $s = 2 + 3 - 4 = 1$ for the intersection of all three linearized hypersurfaces.

Using (1), we obtain the corresponding Nambu system

$$\dot{x} = N_1(x, y, z, w) := wz - x^2z - y^2z, \tag{12}$$

$$\dot{y} = N_2(x, y, z, w) := -wz + x^2z + y^2z, \tag{13}$$

$$\dot{z} = N_3(x, y, z, w) := -w^2x + x^3 + w^2y - x^2y + xy^2 - y^3, \tag{14}$$

$$\dot{w} = N_4(x, y, z, w) := -2xz + 2wxz + 2yz - 2wyz, \tag{15}$$

where $\mathbf{N}(x, y, z, w) := (N_1, N_2, N_3, N_4)^T$ is the Nambu vector field.

In addition, the dissipation terms $G_1(x, y, z, w)$, $G_2(x, y, z, w)$, $G_3(x, y, z, w)$, and $G_4(x, y, z, w)$ have to be constructed with (2). We obtain

$$\begin{aligned} G_1(x, y, z, w) = -\Bigg(& x\left(-E_1 + \frac{1}{2}\left(\frac{w^2}{2} + x^2 + y^2 + z^2\right) - 1\right) \\ & +2wx\left(-E_2 + wx^2 + wy^2 + z^2\right) \\ & -E_3 + x + y - 1\Bigg), \end{aligned} \tag{16}$$

$$\begin{aligned} G_2(x, y, z, w) = -\Bigg(& y\left(-E_1 + \frac{1}{2}\left(\frac{w^2}{2} + x^2 + y^2 + z^2\right) - 1\right) \\ & +2wy\left(-E_2 + wx^2 + wy^2 + z^2\right) \\ & -E_3 + x + y - 1\Bigg), \end{aligned} \tag{17}$$

$$\begin{aligned} G_3(x, y, z, w) = -\Bigg(& z\left(-E_1 + \frac{1}{2}\left(\frac{w^2}{2} + x^2 + y^2 + z^2\right) - 1\right) \\ & +2z\left(-E_2 + wx^2 + wy^2 + z^2\right)\Bigg), \end{aligned} \tag{18}$$

$$\begin{aligned} G_4(x, y, z, w) = -\Bigg(& \frac{1}{2}w\left(-E_1 + \frac{1}{2}\left(\frac{w^2}{2} + x^2 + y^2 + z^2\right) - 1\right) \\ & + \left(x^2 + y^2\right)\left(-E_2 + wx^2 + wy^2 + z^2\right)\Bigg). \end{aligned} \tag{19}$$

where E_1, E_2, E_3 are the energy constants of (2) and ϵ is the damping coefficient.

Finally, we obtain the equations of motion of the CD Nambu system:

$$\dot{x} = N_1(x, y, z, w) + \varepsilon G_1(x, y, z, w), \tag{20}$$

$$\dot{y} = N_2(x, y, z, w) + \varepsilon G_2(x, y, z, w), \tag{21}$$

$$\dot{z} = N_3(x, y, z, w) + \varepsilon G_3(x, y, z, w), \tag{22}$$

$$\dot{w} = N_4(x, y, z, w) + \varepsilon G_4(x, y, z, w), \tag{23}$$

or

$$\dot{\mathbf{X}} = \mathbf{N}(\mathbf{X}) + \varepsilon \mathbf{G}(\mathbf{X}), \tag{24}$$

with $\mathbf{X} := (x, y, z, w)^T$ and $\mathbf{G} := (G_1, G_2, G_3, G_4)^T$. Now we choose $E_1 = E_2 = E_3 = 1$ for the constants and initial conditions and solve these differential equations with a standard numerical solver of Mathematica. In order to represent solutions graphically, we project them into a suitable 3-dimensional subspace. Depending on the value of ε, the dissipation changes, and we obtain trajectories with different behaviors. In Fig. 1, two trajectories of the CD Nambu system are shown, where both trajectories start from their initial values and converge to the limit cycle: 1. Blue curve: $(x(0), y(0), z(0), w(0)) = (-10, -5, -5, -5)$, $\varepsilon = 0.4$. 2. Red curve: $(x(0), y(0), z(0), w(0)) = (5, 5, 5, 5)$, $\varepsilon = 0.04$. The limit cycle is prescribed by the intersection of hypersurfaces that correspond to the Nambu functions.

It can be shown by numerical calculations that the dissipation terms $\mathbf{G}(x(t), y(t), z(t), w(t))$ in the blue and red solution trajectories $(x(t), y(t), z(t), w(t))_{blue,red}$ converge to zero. If we consider the Nambu functions of the red and blue solutions in Fig. 2, we find that these functions converge to the chosen values $E_1 = E_2 = E_3 = 1$; see Fig. 2.

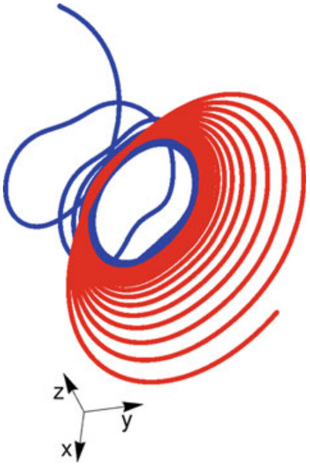


Fig. 1 CD Nambu system of Example 1

We already mentioned in the introduction of this article that CD Nambu systems can be used for the synthesis of nonlinear systems with a limit cycle whose path arises as intersection of hypersurfaces that correspond to the chosen Nambu functions. If the vector field $\mathbf{N} + \varepsilon\mathbf{G}$ can be realized by electronics subsystems (analog adders and multipliers), an oscillator circuit can be developed with a prescribed limit cycle; cf. [22].

However, it is also possible to use CD Nambu systems for the analysis of mathematical problems. It is obvious from its construction procedure and our numerical results that the damping term $\mathbf{G}$ converges to zero, because the Nambu functions H_i $(i = 1, 2, 3)$ converge to the prescribed constants E_1, E_2, E_3; see Fig. 2. The transient processes are very fast, so that a trajectory of the remaining system of differential equations approximates the intersection set of the hypersurfaces. Generically, the local dimension of the intersection set is one, since we have $n - 1$ equations generated by $n - 1$ Nambu functions, which depend on n variables. Therefore, the CD Nambu concept can be interpreted as an embedding method whereby at least parts of the solution set of a system of nonlinear equations $H_i(x, y, z, w) - E_i = 0$, $i = 1, 2, 3$, can be calculated by means of a dynamical process. In fact, if we consider the zero set of a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ consisting generically of isolated zeros, $\mathbf{f}$ can be interpreted as the vector field of the system of differential equations

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{25}$$

with initial values $\mathbf{x}(\mathbf{0}) = \mathbf{x_0}$. Obviously, the zeros of $\mathbf{f}$ correspond to the equilibrium points of the associated differential equation (25). If the initial values $\mathbf{x_0}$ are chosen within the basin of attraction of an equilibrium point, the corresponding trajectory converges to this point. Depending on the stability of the equilibrium points, we have to proceed in the positive or negative direction of the variable t. The CD Nambu approach generalizes this idea to nonisolated zeros of a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^{n-1}$.

It is known from the literature in numerics that almost no algorithms are available for zero set problems with nonisolated zeros; cf. Keller [32] for further details. The construction of a corresponding CD Nambu system can be an interesting concept for the class of numerical problems for which the set of nonisolated zeros of a system of nonlinear equations has generic dimension one, that is, the zero set problem is defined by $n - 1$ equations in n unknown variables. In the next section we discuss a problem in which the intersection set consists of multiply connected parts for which the CD

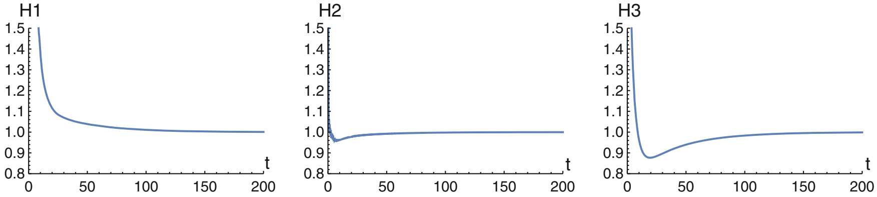


Fig. 2 Nambu functions of the solutions $x(t)$, $y(t)$, $z(t)$, $w(t)$ of Example 1

Nambu system can also be interpreted in the sense of an embedding method. It will be shown that the use of weak damping leads to a probing of the multiply connected parts of a zero set, so that a certain part can be obtained by finding suitable initial values. Therefore, we suggest a new concept for a difficult class of mathematical problems.

3.3 CD Nambu System with Several Limit Cycles

Inspired by C. Nehrkorn [33], we studied a CD Nambu system based on the Clebsch cubic hypersurface in $\mathbb{R}^4$ generated by the Nambu function

$$H_3(x, y, z, w) = x^3 + y^3 + z^3 + w^3 - (w + x + y + z)^3. \tag{26}$$

This hypersurface is intersected by two hypersurfaces that correspond to

$$H_1(x, y, z, w) = \left(2 - w^2\right)\left(x^2 + y^2 + w^2 + z^2 - 2\right), \tag{27}$$

$$H_2(x, y, z, w) = w^2z^2 - w^2 + wx^2 + wy^2. \tag{28}$$

Using these Nambu functions, the procedure (1) results in the following Nambu system (without dissipation):

$$\begin{aligned}
\dot{x} = {} & \Big(-8wyz + 8w^2yz + 4w^3yz - 4w^4yz\Big)\,(3w^2 - 3(w + x + y + z)^2) \\
& +(-8wz - 12w^3z + 8w^5z + 4x^2z - 2w^2x^2z + 4w^3x^2z \\
& +4y^2z - 2w^2y^2z + 4w^3y^2z + 8wz^3)(3y^2 - 3(w + x + y + z)^2) \\
& -(-8wy - 16w^2y + 4w^3y + 8w^4y + 4x^2y + 2w^2x^2y + 4y^3 + 2w^2y^3 \\
& +8wyz^2 + 4w^2yz^2 - 4w^3yz^2)(3z^2 - 3(w + x + y + z)^2) =: N_1(x, y, z, w),
\end{aligned} \tag{29}$$

$$\begin{aligned}
\dot{y} = {} & -(-8wxz + 8w^2xz + 4w^3xz - 4w^4xz)(3w^2 - 3(w + x + y + z)^2) \\
& -(-8wz - 12w^3z + 8w^5z + 4x^2z - 2w^2x^2z + 4w^3x^2z + 4y^2z - 2w^2y^2z \\
& +4w^3y^2z + 8wz^3)(3x^2 - 3(w + x + y + z)^2) + (-8wx - 16w^2x + 4w^3x \\
& +8w^4x + 4x^3 + 2w^2x^3 + 4xy^2 + 2w^2xy^2 + 8wxz^2 + 4w^2xz^2 \\
& -4w^3xz^2)\Big(3z^2 - 3(w + x + y + z)^2\Big)) =: N_2(x, y, z, w),
\end{aligned} \tag{30}$$

$$\begin{aligned}
\dot{z} = {} & (-8wy - 16w^2y + 4w^3y + 8w^4y + 4x^2y + 2w^2x^2y + 4y^3 + 2w^2y^3 \\
& +8wyz^2 + 4w^2yz^2 - 4w^3yz^2)\Big(3x^2 - 3(w + x + y + z)^2\Big) \\
& -(-8wx - 16w^2x + 4w^3x + 8w^4x + 4x^3 + 2w^2x^3 + 4xy^2 + 2w^2xy^2 \\
& +8wxz^2 + 4w^2xz^2 - 4w^3xz^2)\Big(3y^2 - 3(w + x + y + z)^2\Big)) =: N_3(x, y, z, w),
\end{aligned} \tag{31}$$

$$\begin{aligned}
\dot{w} = {} & -(-8wyz + 8w^2yz + 4w^3yz - 4w^4yz)(3x^2 - 3(w + x + y + z)^2) + (-8wxz \\
& +8w^2xz + 4w^3xz - 4w^4xz)\Big(3y^2 - 3(w + x + y + z)^2\Big)) =: N_4(x, y, z, w),
\end{aligned}$$

where $\mathbf{N}(x, y, z, w) := (N_1, N_2, N_3, N_4)^T$ is the Nambu vector field. Now the canonical damping term $\mathbf{G}$ is added to the Nambu vector field $\mathbf{N}$ using the procedure (2), and the corresponding CD Nambu system is obtained. The components of $\mathbf{G}$ consist of many summands, so it is not convenient to include them in this paper. If necessary, these components can be constructed by (2). Again choosing $E_1 = E_2 = E_3 = 1$ and damping coefficient $\varepsilon = 0.005$, we obtain for the initial point $(1/2, 1, 0, 1)$ the trajectory in Fig. 3. We find that after a fast but complex transient, the limit cycle is reached in a long relaxation process. Starting from the bottom, the trajectory runs over the displayed surface, which corresponds to the projection of the intersection of the three hypersurfaces constructed above onto the xyz-coordinate plane. Eventually, the trajectory converges to a limit cycle at the front of the graphical object, where the trajectory covers a part of the surface very densely. Furthermore, it can be observed that no trajectories exist in some areas of the graphical object.

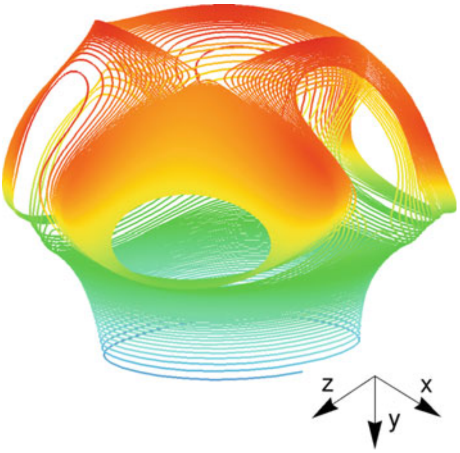


Fig. 3 Trajectory of Example 2 with weak damping $\varepsilon = 0.005$

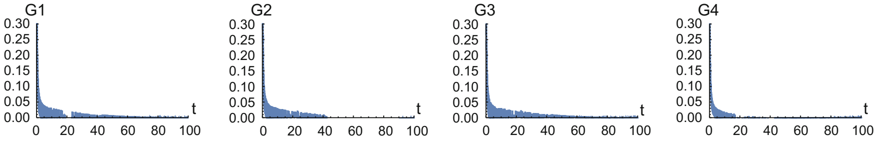


Fig. 4 Dissipative terms of Example 2 with weak damping $\varepsilon = 0.005$

In Fig. 4, the four damping terms G_i $(i = 1, 2, 3)$ are shown with respect to the time steps. Obviously, the damping terms converge to zero very rapidly, and the dynamics is determined by the remaining Nambu system. In this case, the Nambu functions converge to the values of energy levels $E_1 = E_2 = E_3 = 1$; see Fig. 5.

In Fig. 6, three trajectories are shown with different initial values and damping coefficients. It is easy to recognize that the CD Nambu has two different limit cycles, colored in blue and green. The convergence of the trajectory of the weak damped CD Nambu system is already shown in Fig. 3, but in Fig. 6, the calculation time is reduced in order to get a clear graphical representation. It was already mentioned that CD Nambu systems can be used for the construction of nonlinear systems with prescribed limit cycles. Example 2 was constructed on the basis of the Clebsch cubic, where the intersection decomposes into multiply connected parts. Therefore, our approach can used to construct nonlinear dynamical systems with prescribed limit cycles or for solving nonlinear equations with a set of nonisolated zeros that decomposes into multiply connected parts. To the best of our knowledge, there is no known algorithm for the calculation of such solution sets of nonlinear systems.

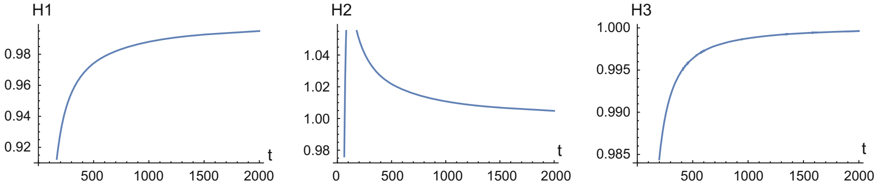


Fig. 5 Nambu functions of Example 2 with weak damping $\varepsilon = 0.005$

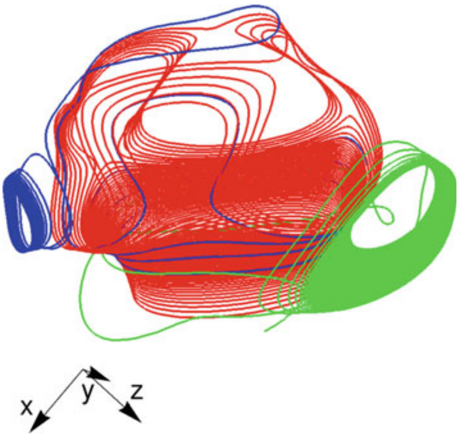


Fig. 6 Example 2: *Red* $\varepsilon = 0.005$, *blue* and *green* $\varepsilon = 0.05$

4 Conclusion

In this article we discussed the properties of CD Nambu systems in four-dimensional state spaces. It was argued that it is rather difficult to find systematic methods for constructing Nambu functions (e.g., in algebraic geometry) whose corresponding hypersurfaces intersect in a closed curve. Therefore, we have presented two examples and discuss their properties. In fact, it was our original thought to deal with CD Nambu systems for the synthesis of oscillators with a prescribed limit cycle. However, we found that the CD Nambu approach can be used for the analysis of nonlinear equations with nonisolated zeros. If we consider a system of $n - 1$ equations in n variables, it is underdetermined and has nonisolated zeros. Now we interpret the equations as Nambu functions and construct an associated CD Nambu system. If we use weak damping, ε is "small," then the simply or multiply connected zero set can be approximated by a suitable trajectory of the CD Nambu system. It has to be proven by further studies whether this approach is suitable for practical applications.

Acknowledgements We would like to thank M. Sc. Daniel Stahl and Dipl.- Ing. Marco Reit for their support with respect to Mathematica and LaTeX.

References

1. Nambu, Y.: Generalized hamiltonian dynamics. Phys. Rev. D **7**, 2403–2412 (1973)
2. Axenidesa, M., Floratos, E.: Strange attractors in dissipative Nambu mechanics: classical and quantum aspects. J. High Energy Phys. **2010**(4), 1–32 (2010)
3. Fecko, M.: On symmetries and conserved quantities in Nambu mechanics. J. Math. Phys. **54**, 102901 (2013)
4. Mukunda, N., Sudarshan, E.C.G.: Relation between Nambu and Hamiltonian mechanics. Phys. Rev. D **13**(10), 2403–2412 (1976)
5. Modin, K.: Time transformation and reversibility of Nambu-poisson systems. J. Gen. Lie Theory Appl. **3**(1), 39–52 (2009)
6. Steeb, W.-H., Euler, N.: A note on Nambu mechanics. Il Nuovo Cimento B **106**(3), 263–272 (1991)
7. Wade, A.: Nambu-Dirac structures for Lie algebroids. Lett. Math. Phys. **61**, 85–99 (2002)
8. Takhtajan, L.: On foundation of the generalized Nambu mechanics. Commun. Math. Phys. **160**, 295–315 (1994)
9. Bihlo, A.: Rayleigh-Bénard convection as a Nambu-metriplectic problem. J. Phys. A: Math. Theor. **41**, 292001 (2008)
10. M. Axenides, E. Floratos, Scaling properties of the lorenz system and dissipative nambu mechanics. arXiv:1205.3462v2 [nlin.CD]. Accessed 19 June 2012
11. Névir, P., Blender, R.: Hamiltonian and Nambu representation of the non-dissipative Lorenz equations. Beitr. Phys. Atmosph. **67**(2), 133–140 (1994)
12. Roupas, Z.: Phase space geometry and chaotic attractors in dissipative Nambu mechanics. arXiv:1110.0766v3 [nlin.CD]. Accessed 25 Apr 2012
13. Haken, H., Wunderlin, A.: New interpretation and size of strange attractor of the Lorenz model of turbulence. Phys. Lett. **62A**, 133–134 (1977)
14. Ebeling, W., Sokolov, I.M.: Statistical Thermodynamics and Stochastic Theory of Nonequilibrium Systems. World Scientific Publ. C. Pte. Ltd., Singapore (2005)

15. Frank, T.D.: Active systems with Nambu dynamics: with applications to rod wielding for haptic length perception and self-propagating systems on two-spheres. Eur. Phys. J. B **74**, 195–203 (2010)
16. Frank, T.D.: Fokker-Planck approach to canonical-dissipative Nambu systems: with an application to human motor control during dynamic haptic perception. Phys. Lett. A **374**, 3136–3142 (2010)
17. Frank, T.D.: Unifying mass-action kinetics and Newtonian mechanics by means of Nambu brackets. J. Biol. Phys. **37**, 375–385 (2011)
18. Mongkolsakulvong, S., Chaikhan, P., Frank, T.D.: Oscillatory nonequilibrium Nambu systems: the canonical-dissipative Yamaleev oscillator. Eur. Phys. J. B, 85–90 (2012)
19. Mathis, W., Stahl, D., Mathis, R.: Oscillator synthesis based on Nambu Mechanics with Canonical Dissipative Damping. In: Proceedings of the 21st European Conference on Circuit Theory and Design (ECCTD), Dresden, Germany, 18–12 Sept, 2013
20. Thiessen, T., Mathis, W.: On noise analysis of oscillators based on statistical mechanics. Int. J. Electron. Telecommun. **56**, 357–366 (2010)
21. Mathis, W., Richter, F., Mathis, R.: Stochastic behavior of dissipative hamiltonian systems with limit cycles. In: Proceedings of the MATHMOD 2012, Vienna, Austria, 15–17 February, 2012
22. Mathis, W., Mathis, R.: Dissipative Nambu systems and oscillator circuit design. IEICE Nonlinear Theory Appl. **5**(3), 259–271 (2014)
23. Andronov, A.A.: Les cycles limites de Poincaré et la théorie des oscillations autoentretenues. Comptes Rendus **189**, 559 (1929)
24. Mathis, R., Mathis, W.: 4-dimensional polynomial dynamical systems with prescribed algebraic limit cycles using Nambu brackets. In: Proceedings of the Fourth International Workshop on Nonlinear Dynamics and Synchronisation (INDS'15), Klagenfurt, Austria, 31 July 2015
25. Grozin, A.: Introduction to Mathematica for Physicists. Springer, Cham (2014)
26. Hulek, K.: Elementary Algebraic Geometry. Student Mathematical Library. American Mathematical Society, Providence (2003)
27. Odani, K.: The limit cycle of the van der pol equations is not algebraic. J. Differ. Equ. **115**, 146–152 (1995)
28. Weizenöck, R.W.: Der vierdimensionale Raum. Springer Basel AG, Basel (1956)
29. Aleksandrov, A.D., Kolmogorov, A.N., Lavrent'ev, M.A. (eds.): Mathematics - Its Content, Methods, and Meaning, vol. 3. The MIT Press Massachusetts Institute of Technology Cambridge, Massachusetts (1963)
30. Smith, K.E., Kahanpää, L., Kekäläinen, P., Treves, W.: An Invitation to Algebraic Geometry. Springer, New York (2000)
31. Gürses, M., Guseinov, G.S., Zheltukhin, K.: Dynamical systems and poisson structure. J. Math. Phys. **50**, 112703 (2009)
32. Keller, H.B.: Geometrically isolated nonisolated solutions and their approximation. SIAM J. Numer. Anal. **18**, 822–838 (1981)
33. Nehrkorn, C.M.: Die 27 Geraden auf einer glatten Kubik (English translation: The 27 straight lines on a smooth cubic), diploma work, University of Freiburg (2010)
34. Christopher, C.: Polynomial vector fields with prescribed algebraic limit cycles. Geom. Dedicata. **88**, 255–258 (2001)

On the Dynamics of Chaotic Systems with Multiple Attractors: A Case Study

J. Kengne, A. Nguomkam Negou, D. Tchiotsop, V. Kamdoum Tamba and G.H. Kom

Abstract In this chapter, the dynamics of chaotic systems with multiple coexisting attractors is addressed using the well-known Newton–Leipnik system as prototype. In the parameters space, regions of multistability (where the system exhibits up to four disconnected attractors) are depicted by performing forward and backward bifurcation analysis of the model. Basins of attraction of various coexisting attractors are computed, showing complex basin boundaries. Owing to the fractal structure of basin boundaries, jumps between coexisting attractors are predicted in experiment. A suitable electrical circuit (i.e., analog simulator) is designed and used for the investigations. Results of theoretical analysis are verified by laboratory experimental measurements. In particular, the hysteretic behavior of the model is observed in experiment by monitoring a single control resistor. The approach followed in this chapter shows that by combining both numerical and experimental techniques, one can gain deep insight into the dynamics of chaotic systems exhibiting multiple attractor behavior.

1 Introduction

It is well known that nonlinear dynamical systems can develop various forms of complexity such as bifurcation, chaos, hyperchaos, and intermittency, just to name a few. The occurrence of two or more asymptotically stable equilibrium points or attracting sets (e.g., period-n limit cycle, torus, chaotic attractor) as the system parameters are being monitored represents another striking and complex behavior observed in nonlinear systems. In a system developing coexisting attractors, the trajectories selec-

J. Kengne (✉) · A. Nguomkam Negou · D. Tchiotsop · G.H. Kom
Laboratoire d'Automatique et Informatique Appliquée (LAIA), Department of Electrical Engineering, IUT-FV Bandjoun, University of Dschang, Dschang, Cameroon
e-mail: kengnemozart@yahoo.fr

A. Nguomkam Negou · V. Kamdoum Tamba
Laboratory of Electronics and Signal Processing, Department of Physics, University of Dschang, 67, Dschang, Cameroon

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_2

tively converge to either of the attracting sets depending on the initial state of the system. Correspondingly, the basin of attraction of an attractive set is defined as the set of initial points whose trajectories converge to the given attractor. The boundary separating each basin of attraction can be a smooth boundary or riddled basin with no clear demarcation (i.e., fractal). This striking and interesting phenomenon has been encountered in various nonlinear systems including lasers [1], biological systems [2, 3], chemical reactions [4], Lorenz systems [5], Newton–Leipnik systems [6], and electrical circuits [7–11]. Such a phenomenon is connected primarily to the system symmetry and may be accompanied by some special effects such as symmetry-breaking bifurcation, symmetry-restoring crisis, coexisting bifurcations, and hysteresis [12–15]. In practice, the coexistence of multiple attractors implies that an attractor may suddenly jump to a different attractor, the situation in which coexisting attractors possess a fractal or intermingled basin of attraction being the most intriguing. In this case, due to noise, the observed signal may be the result of random switching of the system trajectory between two or more concurrent coexisting attractors. This chapter deals with the dynamics of the Newton–Leipnik equation considered as a prototypal dynamical system with multiple coexisting attractors. First of all, let us review some interesting works related to the analysis and control of this particular system. The mathematical model of the so-called Newton–Leipnik system was introduced by Newton and Leipnik [6] in 1981. The Euler rigid-body equations were modified with the addition of linear feedback. A system of three quadratic differential equations was obtained that for certain feedback gains develops two strange attractors. The attractor for an orbit was determined by the location of the initial point for that orbit. In [16], Wang and Tian consider the bifurcation analysis and linear control of the Newton–Leipnik system as a prototypal dynamic system with two strange attractors. The static and dynamic bifurcations of the model are studied. Chaos controlling is performed by a linear controller, and numerical simulation of the control is supplied. Further results on the dynamics and bifurcations of the Newton–Leipnik equation were provided by Lofaro [17]. The authors used numerical computations and local stability calculations to suggest that the dynamics of the Newton–Leipnik equations are related to the dynamics and bifurcations of a family of odd symmetric bimodal maps. The article [18] also studies the dynamical behavior of the Newton–Leipnik system and its trajectory-transformation control problem to multiple attractors. A simple linear state feedback controller for the Newton–Leipnik system based on Lyapunov stability theory and application of the inverse optimal control strategy is designed. Chaotic attractors are stabilized asymptotically to unstable equilibria of the systems, so that the transformation of one attractor to another for the trajectory of the Newton–Leipnik system is realized. In [19], it is shown how a chaotic system with more than one strange attractor can be controlled. Issues in controlling multiple (coexisting) strange attractors as stabilizing a desired motion within one attractor as well as taking the system dynamics from one attractor to another are addressed. Realization of these control objectives is demonstrated using as a numerical example the Newton–Leipnik equation. Motivated by the above-mentioned results, this chapter proposes a methodological analysis of the Newton–Leipnik equation considered as a prototypal dynamical system with multiple coexisting attractors. Regions of multiple attractor behavior (i.e., hysteretic dynamics) are illustrated using bifurcation diagrams computed based on suitable techniques. Furthermore, basins of attraction of various coexisting attractors are also computed to visualize how the various coexisting attractors magnetize the state space. Owing to the fast computation speed of analog computers, we suggest that such an apparatus can be advantageously exploited to investigate nonlinear systems with multiple attractors such as the Newton–Leipnik equation. The rest of the chapter is structured as follows. Section 2 describes the mathematical model of the Newton–Leipnik system. Some basic properties of the model are underlined with practical implications for the occurrence of multiple attractors. In Sect. 3, the bifurcation structures of the system are investigated numerically showing period-doubling and symmetry-recovering crisis phenomena. Regions of the parameters space corresponding to the occurrence of multiple coexisting attractors are depicted. Correspondingly, basins of attraction of various coexisting solutions are computed showing complex basin boundaries. A suitable electrical circuit (i.e., analog computer) that can be exploited for the analysis of the Newton–Leipnik equation is proposed in Sect. 4. Finally, some concluding remarks are presented in Sect. 5.

2 Description and Analysis of the Model

2.1 The Model

The mathematical model of the Newton–Leipnik equation [6] considered in this chapter is expressed by the following set of three coupled first-order nonlinear differential equations:

$$\begin{aligned} \dot{x}_1 &= -ax_1 + x_2 + bx_2x_3, \\ \dot{x}_2 &= -x_1 - ax_2 + 5x_1x_3, \\ \dot{x}_3 &= cx_3 - 5x_1x_2, \end{aligned} \tag{1}$$

where x_1, x_2, and x_3 are the state variables; a, b, and c are three positive real constants. It can be seen that the model possesses three quadratic nonlinearities in which are involved the three state variables (x_1, x_2, and x_3).

The presence of this nonlinearity is responsible for the complex behaviors exhibited by the whole system. Obviously, system (1) is invariant under the transformation $(x_1, x_2, x_3) \Leftrightarrow (-x_1, -x_2, x_3)$. Therefore, if (x_1, x_2, x_3) is a solution of system (1) for a specific set of parameters, then $(-x_1, -x_2, x_3)$ is also a solution for the same parameter set. The fixed point $E_0(0, 0, 0)$ is a trivial symmetric static solution. Also, attractors in state space have to be symmetric by reflection in the x_3-axis; otherwise, they must appear in pairs to restore the exact symmetry of the model equations. This exact symmetry could serve to explain the presence of several coexisting attractors in state space [20, 21]. Furthermore, it represents a good way to check the scheme used for numerical analysis. It is important to note that for typical parameters values $a = 0.4$, $b = 10$, $c = 0.175$, system (1) has five equilibrium points, which are all unstable [18, 19]. Also, for these parameter values, the system experiences self-excited oscillations [22, 23].

2.2 Dissipation and Existence of Attractors

Preliminary insights related to the existence of attractors in Newton–Leipnik systems can be gained by evaluating the volume contraction rate [20, 21] of the model. Briefly recall that the volume contraction rate of a continuous-time dynamical system described by $\dot{x} = \varphi(x)$, where x=$(x_1, x_2, x_3)^T$ and $\varphi(x)$=$(\varphi_1(x), \varphi_2(x), \varphi_3(x))^T$, is given by

$$\Lambda = \nabla.\varphi(x) = \frac{\partial \varphi_1}{\partial x_1} + \frac{\partial \varphi_2}{\partial x_2} + \frac{\partial \varphi_3}{\partial x_3}. \tag{2}$$

We note that if Λ is a constant, then the time evolution in phase space is determined by $V(t) = V_0 e^{\lambda(t)}$, where $V_0 = V(t = 0)$. Thus, a negative value of Λ leads to a fast exponential shrinkage (i.e., damped) of the volume in state space; the dynamical system is dissipative and can experience or develop attractors. For $\Lambda = 0$, phase space volume is conserved, and the dynamical system is conservative. If Λ is positive, the volume in phase space expands, and hence there exist only unstable fixed points or cycles or possibly chaotic repellors [20, 21]; in other words, the dynamics diverge as the system evolves (i.e., for $t \to \infty$) if the initial conditions do not coincide exactly with one of the fixed points or stationary states. Referring to the model in (1), it can easily be shown that $\Lambda = c - 2a < 0$ independently of the position $(x_1, x_2, x_3)^T$ in state space; hence system (1) is dissipative, and thus can support attractors.

3 Numerical Study

3.1 Numerical Methods

In order to examine the rich variety of dynamical behaviors that can be observed in a Newton–Leipnik system, we solve numerically system (1) using the classical fourth-order Runge–Kutta integration algorithm. For each set of parameters used in this chapter, the time step is always $\Delta t = 0.005$ and the calculations are carried out using variables and constant parameters in extended mode. For each parameter setting, the system is integrated for a sufficiently long time and the transient is canceled. Two indicators are used to identify the type of scenario giving rise to chaos. The bifurcation diagram represents the first indicator, the second indicator being the graph of the largest Lyapunov exponent (λ_{max}). Concerning the latter case, the dynamics of the system is classified using its Lyapunov exponent, which is computed numerically using the algorithm described by Wolf and collaborators [24]. In particular, the sign of the largest Lyapunov exponent determines the rate of almost all small perturbations to the system's state, and consequently, the nature of the underlined dynamical attractor. For $\lambda_{max} < 0$, all perturbations vanish, and trajectories starting sufficiently close to each other converge to the same stable equilibrium point in state space; for $\lambda_{max} = 0$, initially close orbits remains close but distinct, corresponding to oscillatory dynamics on a limit cycle or torus; and finally, for $\lambda_{max} > 0$, small perturbations grow exponentially, and the system evolves chaotically within the folded space of a strange attractor.

3.2 *Route to Chaos*

To investigate the sensitivity of the system with respect to a single control parameter, we fix $a = 0.6$, $b = 5$ and vary c in the range $0.13 \leq c \leq 0.15$. In monitoring the control parameter, it is found that the Newton–Leipnik system under consideration can experience very rich and striking bifurcation scenarios. Sample results showing bifurcation diagrams for varying c and the corresponding spectrum of largest Lyapunov exponent are provided in Fig. 1a and b respectively.

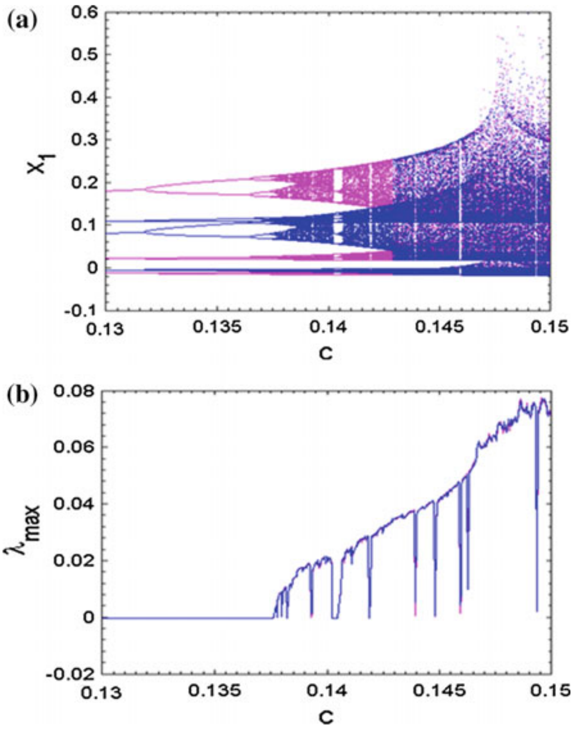


Fig. 1 Bifurcation diagram **a** showing local maxima of the coordinate x_1 versus c and the corresponding graph **b** of largest Lyapunov exponent (λ_{max}) plotted in the range $0.13 \leq c \leq 0.15$. A window of hysteretic dynamics can be noticed for lower values of c. *Magenta* and *blue* diagrams correspond respectively to increasing and decreasing values of c. The positive value of λ_{max} is the signature of chaotic motion: Parameter values $(a, b) = (0.6, 5)$

The bifurcation diagram is obtained by plotting local maxima of the coordinate x_1 in terms of the control parameter, which is increased (or decreased) in tiny steps in the range $0.13 \leq c \leq 0.15$. The final state at each iteration of the control parameter serves as the initial state for the next iteration. In the graph of Fig. 1a, two sets of data corresponding to increasing (blue) and decreasing (magenta) values of c are superimposed. This strategy, known as forward and backward continuation, represents a simple way to localize the window in which the system develops multiple coexisting attractor behaviors (see Sect. 4). In light of Fig. 1a and b, the following bifurcation sequence emerges when the control parameter c is slowly increased. For values of c under the critical value $c_c = 0.11$, the system exhibits periodic oscillations (i.e., period-3 limit cycle). On increasing the control parameter c past this critical value, a stable period-3 limit cycle born from the Hopf bifurcation undergoes a series of period-doubling bifurcations, culminating in a single-band spiraling chaotic attractor. On further increasing c up to $c_{cr1} = 0.1402$, a periodic window suddenly appears in which the system displays a period-8 and then a single-band chaotic attractor. Past the critical value $c_{cr2} = 0.143$, the single-band chaotic attractor suddenly changes to a double-band chaotic attractor following a symmetry-recovering crisis. Also note the presence of many tiny windows of periodic motions sandwiched in the chaotic bands. It can be seen that the bifurcation diagram coincides well with the spectrum of the Lyapunov exponent. With the same parameter settings in Fig. 1, various

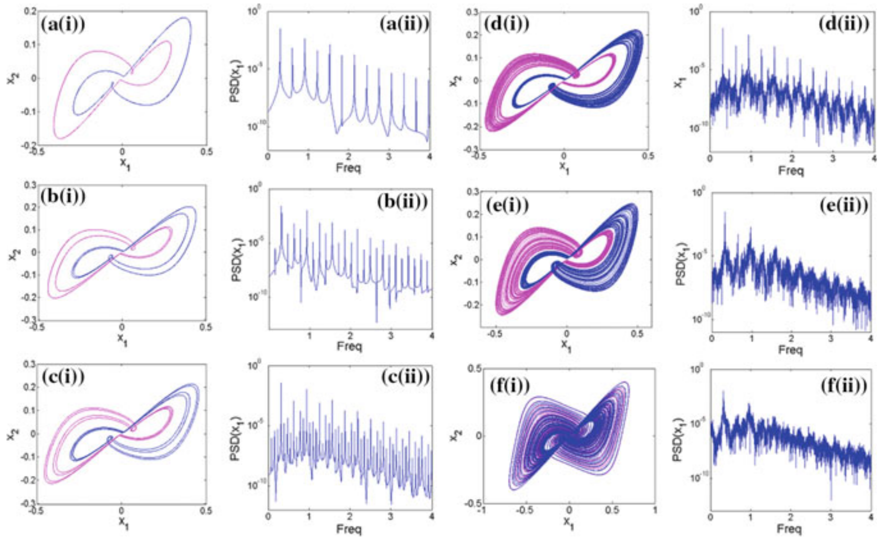


Fig. 2 Two-dimensional views of the attractor projected onto the (x_1, x_2)-plane (*left*) of the system, showing routes to chaos (in terms of the control parameter c) and corresponding power spectra (*right*): **a** period-1 for $c = 0.13$, **b** period-2 for $c = 0.135$, **c** period-4 for $c = 0.1373$, **d** single-band chaos for $c = 0.1388$, **e** single-band chaos for $c = 0.1328$, **f** double-band for $c = 0.145$. The parameters are those of Fig. 2

numerical phase portraits and corresponding frequency spectra were obtained, confirming different bifurcation sequences depicted previously (see Fig. 2).

It should be noted that for periodic motion, all spikes in the power spectrum are harmonically related to the fundamental, whereas a broadband power spectrum is characteristic of a chaotic mode of oscillations. Briefly recall that the periodicity of an attractor is deduced by counting the number of spikes located at the left-hand side of the highest spike of the spectrum (the latter being included).

3.3 Occurrence of Multiple Attractors

To observe the phenomenon of multiple attractors, two prototypes were considered. The first is obtained by setting $(a, b) = (0.6, 5)$. With reference to the bifurcation diagram of Fig. 1, a window of hysteretic dynamics (i.e., multiple stability) can be identified in the range $0.13 \leq c \leq 0.14$ (see Fig. 3).

For values of c within this range, the long-term behavior of the system depends on the initial state, thus leading to the interesting and striking phenomenon of coexisting multiple attractor behaviors. Up to four different attractors can be obtained depending solely on the selection of initial conditions. For $c = 0.142$, four different chaotic attractors are presented in a phase portrait of (Fig. 4) using different initial conditions $(x_1(0), x_2(0), x_3(0)) = (0, \pm 0.1, \pm 0.01)$.

The second approach is obtained using $(a, b) = (0.73, 10)$. We plot the corresponding bifurcation diagram versus the control parameter c, as well as the corresponding graph of the Lyapunov exponent (see Fig. 5a and b).

In the diagram of Fig. 5a, a window of hysteretic dynamics (and thus multistability) can be identified in the range $0.13 \leq c \leq 0.16$. For values of c within this range, the long-term behavior of the system depends on the initial state; hence the system develops the interesting phenomenon of coexisting multiple attractor behaviors. For

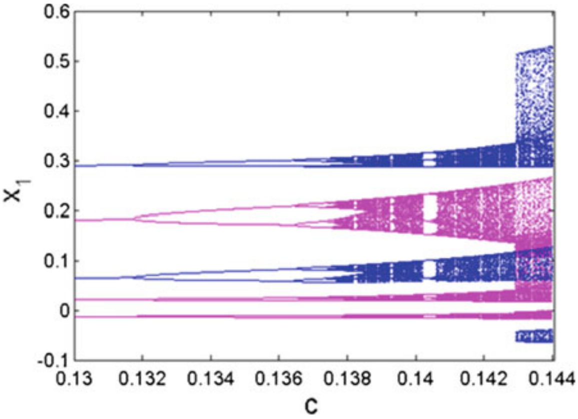


Fig. 3 Enlargement of the bifurcation diagram of Fig. 2 showing the region in which the system exhibits multiple coexisting attractors. This region corresponds to values of c in the range $0.13 \leq c \leq 0.144$. Two sets of data corresponding to increasing (*magenta*) and decreasing (*blue*) values of the control parameter are superimposed

Fig. 4 Coexistence of four different attractors consisting of two pairs of single-band chaotic attractors projected onto the (x_1, x_2)-plane for $(a, b) = (0.6, 5)$ and $c = 0.142$. Initial conditions are $(x_1(0), x_2(0), x_3(0)) = (0, \pm 0.1, \pm 0.01)$

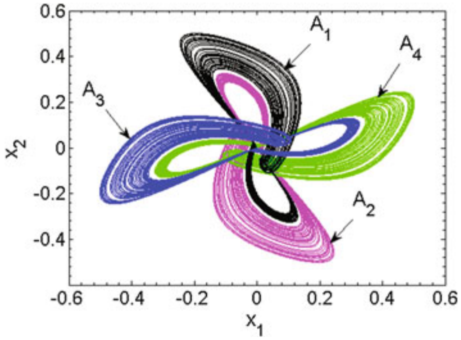
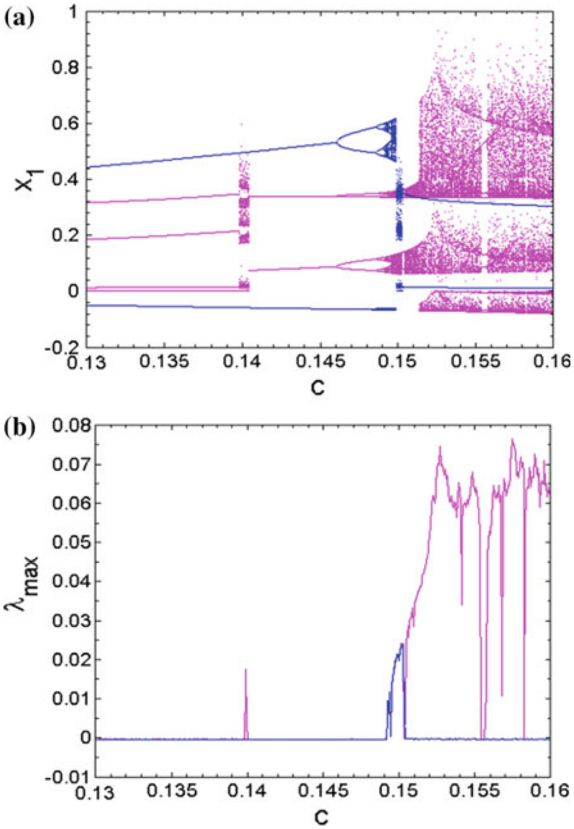


Fig. 5 Bifurcation diagram **a** showing local maxima of the coordinate x_1 versus c and the corresponding graph **b** of the largest Lyapunov exponent λ_{max} plotted in the range $0.13 \leq c \leq 0.16$. A window of hysteretic dynamics can be noticed for some values of c. *Magenta* and *blue* diagrams correspond respectively to increasing and decreasing values of c. The positive value of λ_{max} is the signature of chaotic motion: parameter values $(a, b) = (0.73, 10)$



$c = 0.151$, four different attractors (see Fig. 6d) can be obtained depending only on the selection of initial conditions.

For instance, the pair of period-1 phase portraits and a pair of chaotic attractors of Fig. (6d) can be obtained under the initial conditions $(x_1(0), x_2(0), x_3(0)) = (0, \pm 0.1, \pm 0.01)$. We also see that for $c = 152$, three different attractors (see Fig. 7d) can be obtained depending only on the selection of initial conditions, a pair of

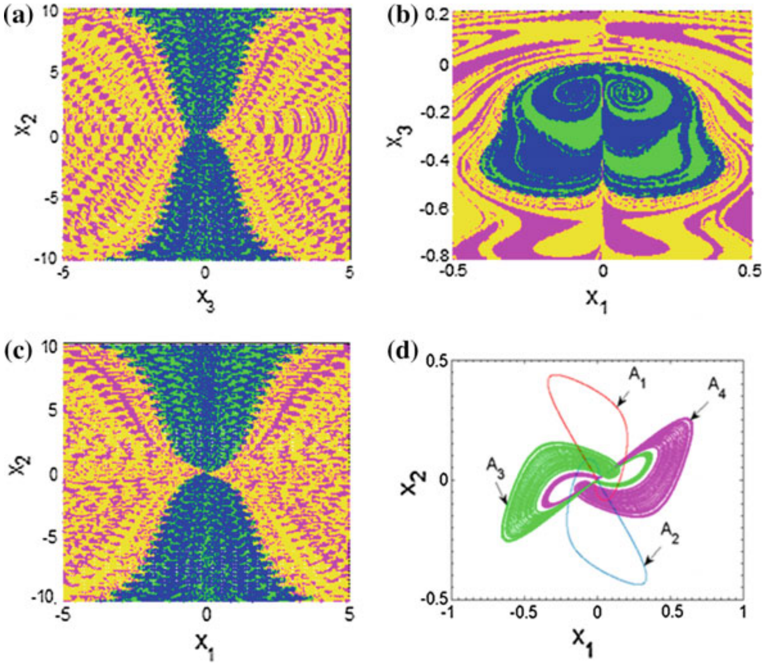


Fig. 6 Cross sections of the basin of attraction **a**, **b**, **c** for $x_1(0) = 0$, $x_2(0) = 0$, and $x_3(0) = 0$, and $x_3(0) = 0$ respectively and two-dimensional views of four different attractors consisting of a pair of single-band chaotic attractors with a pair of asymmetric period-1 attractors **d** projected onto the (x_1, x_2)-plane for $(a, b) = (0.73, 10)$ and $c = 0.151$. Initial conditions are $(x_1(0), x_2(0), x_3(0)) = (0, \pm 0.1, \pm 0.01)$. *Yellow* and *green* regions correspond to the period-1 pair of attractors, while the *magenta* and *blue* regions are associated with the chaotic solutions obtained for $(a, b, c) = (0.73, 5, 0.151)$

period-1 phases portraits and a double-band chaotic attractor for $(x_1(0), x_2(0), x_3(0)) = (0, \pm 0.1, \pm 0.01)$. Therefore, considering the parameters in Fig. 6d and carrying out a scan of initial conditions (see Fig. 6a–c), we have defined the domain of initial conditions in which each attractor can be found. Figure 6a–c show cross sections of the basin of attraction respectively for $x_1(0) = 0$, $x_2(0) = 0$, and $x_3(0) = 0$ corresponding to the symmetric pair of limit cycles (blue and green) and the pair of chaotic attractors (magenta and yellow). Likewise, considering the parameter setting in Fig. 7d, we have defined the domain of initial conditions in which each attractor can be found.

Figure 7a–c show the structure of the sections of the basin of attraction respectively for $x_1(0) = 0$, $x_2(0) = 0$, and $x_3(0) = 0$. Green and yellow lead to a pair of period-1 limit cycle while magenta regions are associated to the double-band chaotic solution. It should be mentioned that multiple attractor behavior (involving at least

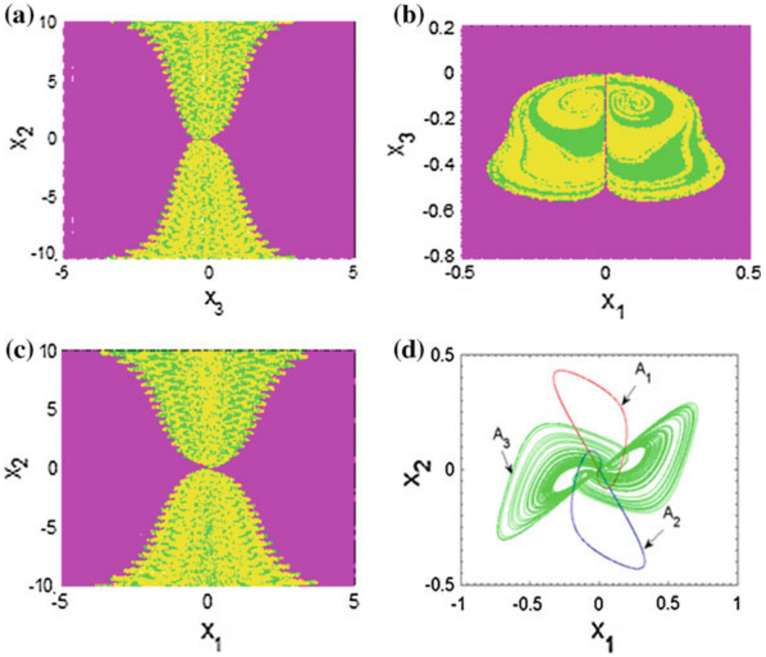


Fig. 7 Cross sections of the basin of attraction **a**, **b**, **c** for $x_1(0) = 0$, $x_2(0) = 0$, and $x_3(0) = 0$ respectively and two-dimensional views of three different attractors consisting of a double-band chaotic attractor with a pair of asymmetric period-1 ones **d** projected onto the (x_1, x_2)-plane. Initial conditions are $(x_1(0), x_2(0), x_3(0)) = (0, \pm 0.1, \pm 0.01)$. *Green* and *yellow* regions correspond to the period-1 pair of attractors, while *magenta* regions are associated with the chaotic solutions obtained for $(a, b, c) = (0.73, 10, 0.152)$

four nonstatic disconnected attractors) is common in various nonlinear systems (see, e.g., Sect. 1). Very recently, Hens and collaborators considered the case of coexistence of infinitely many attractors, also referred to as extreme multistability, in coupled dynamical systems [25]. It is obvious that the occurrence of multiple attractors is an additional source of randomness in chaotic systems that may be exploited for chaos-based secure communication. However, in many other cases, this singular type of behavior is not desirable and it justifies the need for control. Detailed study along this line is beyond the scope of this work; also, interested readers are referred to the review work on control of multistability by Pisarshik and collaborators [26].

4 The Analog Computer Approach

It is well known that even with a very fast computer, scanning the parameter space can turn out to be very time-consuming. Furthermore, there is no rigorous method for selecting the integration step used for the numerical integration as well as the duration of the transient time. These difficulties (as well as many others) faced in performing numerical computation can be overcome by adopting the analog computer approach [27, 28]. One of the merits of the analog computer is the possibility of exploring wide ranges of dynamic behaviors by simply monitoring, for instance, a single control resistor. Nevertheless, the accuracy of the results of analog computation strongly depends on the quality of the electronic components used in the construction of the analog computer. Also, by combining the fast computation speed of an analog simulator and the precision of a digital computer, one can gain deep insight into the dynamics of a given nonlinear process such as the Newton–Leipnik system. Our goal in this section is to design and implement an appropriate analog simulator that can be exploited for the analysis of the model defined in system (1). A circuit diagram of the proposed electronic simulator is provided in Fig. 8. Compared to the

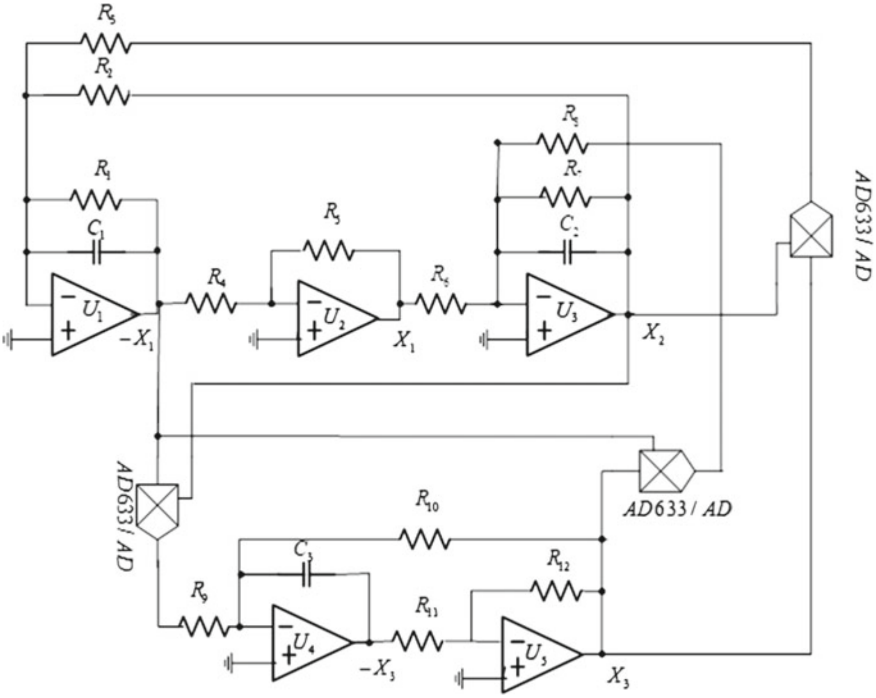


Fig. 8 Electronic circuit realization of a Newton–Leipnik system with three quadratic interactions using $R_1 = R_7 = 16666\Omega$, $R_2 = R_4 = R_5 = R_6 = R_{11} = R_{12} = 10K\Omega$, $R_{10} = 0 - 100K\Omega$, and $C = 10nF$

circuit proposed in [28] (utilizing up to twenty resistors and nine operational (op) amplifiers), the analog simulator shown in Fig. 8 involves a minimum number of electronic components.

The electronic multipliers are the analog devices AD633JN, versions of the AD633 four-quadrant voltage multipliers chips used to implement the nonlinear terms of our model. They operate over a dynamic range of ±1 V with a typical error less than 1%. They also have a built-in divide-by-ten feature. The signal W at the output depends on the signals at inputs $X_1(+)$, $X_2(-)$, $Y_1(+)$, $Y_2(-)$, and $W = (X_1 - X_2)(Y_1 - Y2)/10 + Z$. The operational amplifiers and associated circuitry implement the basic operations of addition, subtraction, and integration. By adopting an appropriate time scaling, the simulator outputs can be displayed directly on an oscilloscope by connecting the output voltage of X_1 to the X input and the output voltage of X_2 to the Y input. By applying the Kirchhoff current and voltage laws to the circuit in Fig. 8, it can be shown that the voltages X_1, X_2, and X_3 satisfy the set of three coupled first-order nonlinear differential equations

$$\begin{aligned}
\frac{dX_1}{dt} &= -\frac{X_1}{R_1C_1} + \frac{X_2}{R_2C_1} + \frac{X_2X_3}{10R_3C_1}, \\
\frac{dX_2}{dt} &= -\frac{X_2}{R_7C_2} - \frac{X_1}{R_6C_2} + \frac{X_1X_3}{10R_8C_2}, \\
\frac{dX_3}{dt} &= \frac{X_3}{R_{10}C_3} - \frac{X_1X_2}{10R_9C_3}.
\end{aligned} \tag{3}$$

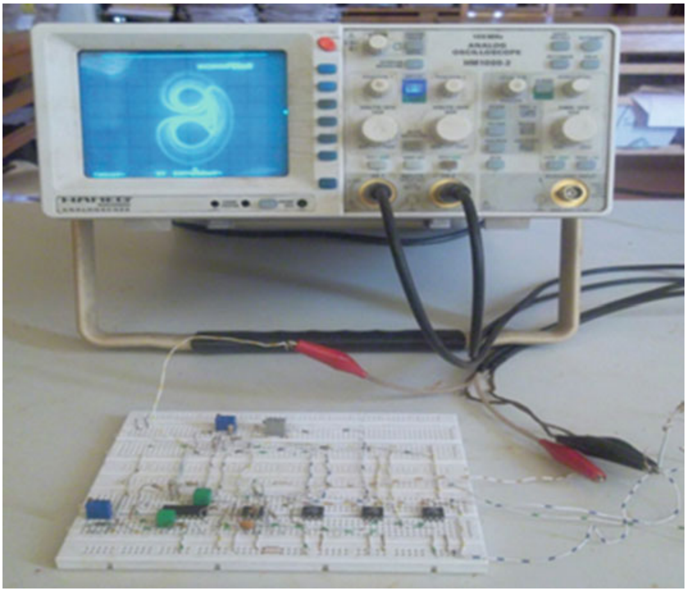


Fig. 9 The experimental Newton–Leipnik simulator in operation

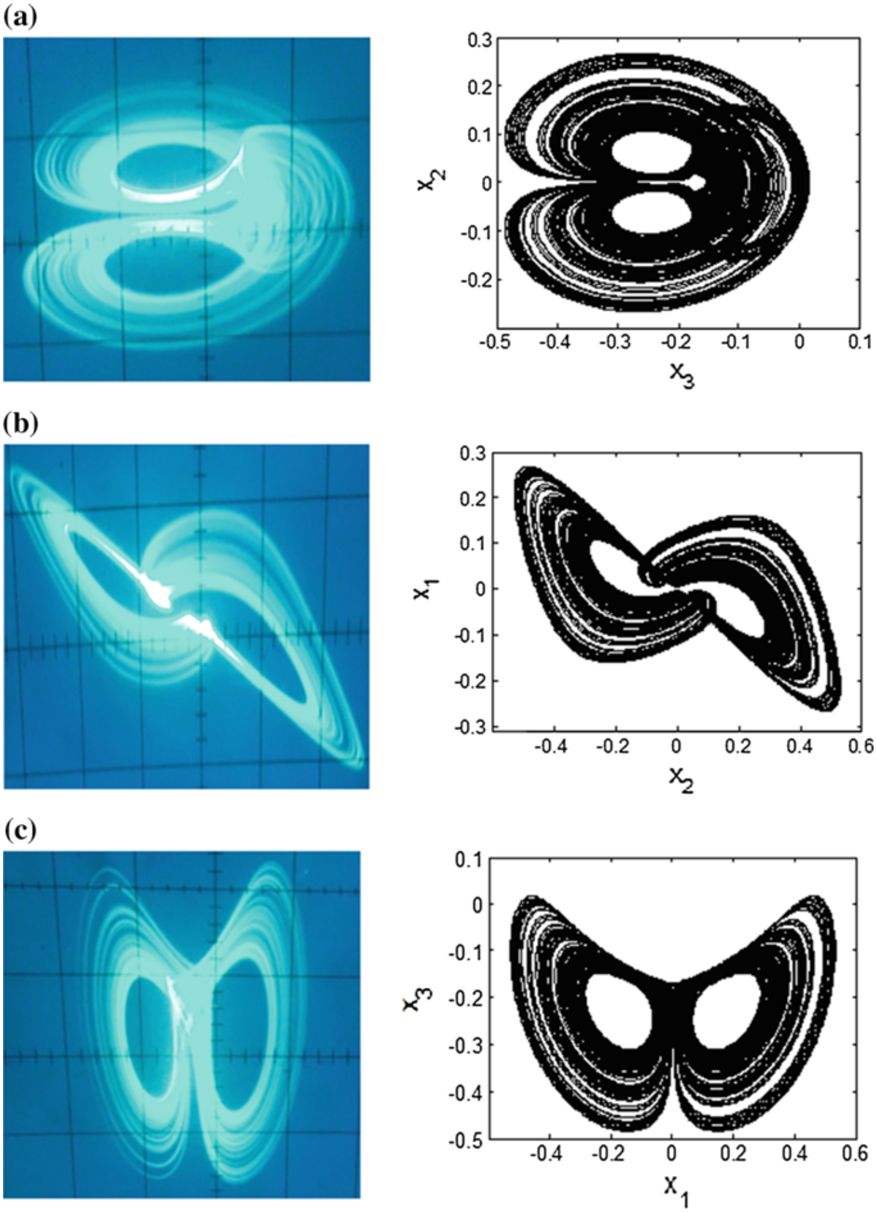


Fig. 10 Experimental phase portraits (*left*) obtained from the circuit using a dual trace oscilloscope in XY mode; the corresponding numerical phase portraits are shown on the *right*, obtained by a direct integration of the system (1) with $(a, b, c) = (0.6, 5, 0.145)$: **a** projection onto the (X_3, X_2)-plane, **b** projection onto the (X_1, X_2)-plane, and **c** projection onto (X_1, X_3)-plane. The scales are $X = 1V/div$ and $Y = 0.5V/div$ for all pictures

With a time unit of 10^4, the parameters of system (1) are expressed in terms of the values of capacitors and resistors as follows (provided that the critical relationships $10^4 R_2C_1 = 1$, $10^4 R_6C_2 = 1$, $5.10^5 R_8C_2 = 1$, $5.10^5 R_9C_3 = 1$ are satisfied):

$$a = \frac{10^{-4}}{R_1C_1}, b = \frac{10^{-4}}{R_3C_1}, c = \frac{10^{-4}}{R_{10}C_3}. \tag{4}$$

We briefly recall that the time scaling process offers to analog devices (operational amplifiers and analog multipliers) the possibility of operating under their bandwidth. Furthermore, time scaling offers the possibility of simulating the behavior of the system at any given frequency by expressing the real time variable τ versus the analog computation time variable $t(t = 10^{-n}\tau)$, allowing the simulation frequency to be less than the real frequency by a factor of order 10^{+n}. In the latter expression, the positive integer depends on the values of resistors and capacitors used in the analog simulator. A photograph of the experimental analog simulator in operation is shown in Fig. 9, while a comparison between numerical and experimental phase portraits is provided in Fig. 10.

From the graphs in Fig. 10, it clearly appears that the dynamics of the Newton–Leipnik system is well reproduced by the analog simulator.

5 Concluding Remarks

This chapter has focused on a methodological analysis of the Newton–Leipnik system considered as a prototypal dynamical system with multiple attractors. Regions of multiple attractors in the parameter space were depicted using bifurcation diagrams based on appropriate techniques (e.g., forward and backward bifurcation diagrams). Furthermore, basins of attraction of various competing attractors were computed, showing nontrivial basin boundaries, thus suggesting possible jumps between different coexisting attractors in experiment. Moreover, one piece of interesting information that can be gained from basins of attraction is the chance of the appearance of attractors in a real system. A suitable electrical circuit (i.e., analog simulator) was designed that was shown to reproduce the Newton–Leipnik attractor. Combined with numerical techniques, the proposed analog computer may be particularly useful for exploring the parameter space in view of tracking further regions of multiple attractors in the Newton–Leipnik model. We stress that the approach followed in this chapter may be exploited advantageously in the investigation of other nonlinear dynamical systems exhibiting multiple attractors.

References

1. Masoller, C.: Coexistence of attractors in a laser diode with optical feedback from a large external cavity. Phys. Rev. A **50**, 2569–2578 (1994)
2. Cushing, J.M., Henson, S.M.: Blackburn: multiple mixed attractors in a competition model. J. Biolog. Dyn. **1**, 347–362 (2007)
3. Upadhyay, R.K.: Multiple attractors and crisis route to chaos in a model of food-chain. Chaos, Solitons Fractals **16**, 737–747 (2003)
4. Massoudi, A., Mahjani, M.G., Jafarian, M.: Multiple attractors in Koper-Gaspard model of electrochemical. J Electroanal. Chem. **647**, 74–86 (2010)
5. Li, C., Sprott, J.C: Coexisting hidden attractors in a 4-D simplified Lorenz system. Int. J. Bifurc. Chaos **24**, 1450034 (2014)
6. Leipnik, R.B., Newton, T.B.: Double strange attractors in rigid body motion with linear feedback control. Phys. Lett. A **86**, 63–87 (1981)
7. Vaithianathan, V., Veijun, J.: Coexistence of four different attractors in a fundamental power system model. IEEE Trans. Cir. Syst. **I46**, 405–409 (1999)
8. Kengne, J.: Coexistence of chaos with hyperchaos, period-3 doubling bifurcation, and transient chaos in the hyperchaotic oscillator with gyrators. Int. J. Bifurc. Chaos **25**(4), 1550052 (2015)
9. Pivka, L., Wu, C.W., Huang, A.: Chua's oscillator: a compendium of chaotic phenomena. J.Frankl. Inst. **331B**(6), 705–741 (1994)
10. Kuznetsov, A.P., Kuznetsov, S.P., Mosekilde, E., Stankevich, N.V.: Co-existing hidden attractors in a radio-physical oscillator. J. Phys. A Math. Theor. **48**, 125101 (2015)
11. Kengne, J., Njitacke, Z.T., Fotin, H.B.: Dynamical analysis of a simple autonomous jerk system with multiple attractors. Nonlinear Dyn. **48**, 751–765 (2016)
12. Li, C., Hu, W., Sprott, J.C., Wang, X.: Multistability in symmetric chaotic systems. Eur. Phys. J. Spec. Top. **224**, 1493–1506 (2015)
13. Letellier, C., Gilmore, R.: Symmetry groups for 3D dynamical systems. J. Phys. A. Math. Theor. **40**, 5597–5620 (2007)
14. Rosalie, M., Letellier, C.: Systematic template extraction from chaotic attractors: I. Genus-one attractors with inversion symmetry. J. Phys. A Math. Theor. **46**, 375101 (2013)
15. Rosalie, M., Letellier, C.: Systematic template extraction from chaotic attractors: II. Genus-one attractors with unimodal folding mechanisms. J. Phys. A Math. Theor. **48**, 235100 (2015)
16. Xuedi, W., Lixin, T.: Bifurcation analysis and linear control of the Newton-Leipnik system. Chaos, Solitons Fractals **27**, 31–38 (2006)
17. Lofaro, T.: A model of the dynamics of the Newton-Leipnik attractor. Int. J. Bifurc. Chaos **7**(12), 2723–2733 (1997)
18. Wang, X., Gao, Y.: The inverse optimal control of chaotic system with multiple attractors. Mod. Phys. Lett. B **21**, 1199–2007 (2007)
19. Hendrik, R.: Controlling chaotic systems with multiple strange attractors. Phys. Lett. A **300**, 182–188 (2002)
20. Strogatz S.H.: Nonlinear Dynamics and Chaos. Addison-Wesley, Reading
21. Nayfeh, A.H., Balachandran, B.: Applied Nonlinear Dynamics: Analytical, Computational and Experimental Methods. Wiley, New York
22. Leonov, G.A., Kuznetsov, N.V.: Hidden attractors in dynamical systems. From hidden oscillations in Hilbert-Kolmogorov, Aizerman, and Kalman problems to hidden chaotic attractor in Chua circuits. Int. J. Bifurc. Chaos **23**, 1793–6551 (2013)
23. Leonov, G.A., Kuznetsov, N.V., Mokaev, T.N.: Homoclinic orbits, and self-excited and hidden attractors in a Lorenz-like system describing convective fluid motion. Eur. Phys. J. Spec. Top. **224**, 1421–1458 (2015)
24. Wolf, A., Swift, J.B., Swinney, H.L., Wastano, J.A.: Determining Lyapunov exponents from time series. Physica D **16**, 285–317 (1985)
25. Pisarchik, A.N., Feudel, U.: Control of multistability. Phys. Rep. **540**(4), 167–218 (2014)
26. Hens, C., Dana, S.K., Feudel, U.: Extreme multistability: attractors manipulation and robustness. Chaos **25**, 053112 (2015)

27. Chedjou, J.C., Fotsin, H.B., Woafo P., Domngang, S.: Analog simulation of the dynamics of a van der Pol oscillator coupled to a Duffing oscillator. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **48**, 748–756 (2001)
28. Zhao, R., Song, Y.: Circuit realization of Newton-Leipnik chaotic system via EWB. Chinese control and decision conference, Yantai, Shandong, pp. 5111–5114 (2008)

Multivaluedness Aspects in Self-Organization, Complexity and Computations Investigations by Strong Anticipation

Alexander Makarenko

Abstract Since the introduction of strong anticipation by D. Dubois, numerous investigations of concrete systems have been proposed. In this chapter, new examples of discrete dynamical systems with anticipation are considered. The mathematical formulation of problems, possible analytical formulas for solutions, and numerical examples of possible solutions are proposed. One of the most interesting properties in such systems is the possible multivaluedness of the solutions. This can be considered from the point of view of dynamical chaos and complex behavior. We present examples of periodic and complex solutions, properties of attractors, and possible applications in self-organization. The main peculiarity is the strong anticipation property. General new possibilities include the possible multivaluedness of the dynamics of automata. Possible interpretations of such behavior of cellular automata are discussed. Further prospects for development of automata theory and hypercomputation are proposed.

Keywords Nature-inspired · Strong anticipation · Multivalued solutions · Chaos · Self-organization · Hypercomputation

1 Introduction. Short Outlook of Singlevaluedness and Multivaluedness in Nature and in Models

The recent science of self-organization is now important to the study of nature, living systems, and investigations into social organization. Here we outline as examples only some achievements: dissipative structures theories and synergetic (I. Prigogine, G. Nicolis, H. Haken), sociodynamics (W. Weidlich, G. Haag, D. Helbing), cellular automata (B. Chopard, M. Droz, S. Wolfram, L. Chua), morphogenesis theory (A. Turing), artificial neuronets (J. Hopfield), and many others.

A. Makarenko (✉)
Institute for Applied System Analysis at National Tech.
University of Ukraine (KPI), Kyiv, Ukraine
e-mail: makalex@i.com.ua, makalex51@gmail.com

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_3

There exist many models of such objects and their solutions including parabolic equations of heat and mass transfer, kinetic equations, equations with memory and space nonlocality, ordinary differential equations, cellular automata, and discrete equations. Such models have a variety of solutions: stationary and periodic solutions, deterministic chaos, solitons and autowaves, "chimera" solutions, blowup (collapses), synchronization, fragmentation, "ideal" turbulence by A. Sharkovski and colleagues (see, for example, P. Shuster, A. Pikovski, Yu. Maystrenko, E. Mosekilde, A. Sharkovski, A. Samarski and S. Kurdiumov, P. Sloot, A. Hoekstra). In most cases, since I. Newton, one of the main requirements is the *single-valuedness* of the solutions in suitable spaces of the solutions (frequently in very complicated spaces, for example in Sobolev spaces).

At the same time, our understanding of computational processes in nature and artificial devices has improved. Examples are automata theory (M. Sipser, B. Cooper) and cellular automata theory (J. von Neumann, S. Wolfram, L. Chua, A. Illiachinski). Applications are distributed from the quantum level (G. t 'Hooft, A. Zellinger, G. Grossing, H.-T. Elze, K. Zuse, S. Wolfram). We remark that previously, most such objects of computational theory and applications were single-valued (classical automata theory, computer architecture, etc.).

However, in parallel to the problems with single-valued solutions, many phenomena were found in nature and engineering that have models with *multivalued solutions* (that is, the existence of possibly many solution values at a given moment of time). We remark also that mathematical tools for considering multivalued solutions have been developed. As examples of tools we recall variational inequalities, differential inclusions, and differential equations with discontinuous coefficients and nonlinear sources. A number of models with intrinsically multivalued solutions have recently been proposed. Examples are multivalued fields (H. Kleinert), mechanical systems with discontinuities (M. Zak, A. Ioffe), turbulence (J. Leray), multivalued solitons (V. Vachnenko), control theory and differential games (A. Chikriy, R. T. Rockafellar), nonclassical mechanical systems with friction and collisions (J. Moreau, C. Glocker), systems with multivalued Hamiltonians in field theory (M. Henneaux, C. Teitelboim, J. Zanelli), multivalued functionals in field theory (S. Novikov), nonuniqueness phenomena (ghost) in field theory (L. Faddeev, A. Grib), Hamiltonian inclusions system (R. T. Rockafellar), and of course inverse problems.

But recently it has been found that the investigation of systems with anticipation is also very interesting. The term "anticipation" was first attached to systems with intrinsic models for predicting the evolution of the systems and their environment (see [4, 5, 9–12, 15–18] and many others). But now since the work of Daniel M. Dubois (Belgium), the notion of "strong anticipation" has been introduced and investigated. In the case of strong anticipation, a system doesn't have models for predictions of future states of the system, but take into account the presence of virtual future states in evolution. D. Dubois also introduced the notion of incursive and hyperincursive systems (with possible multivaluedness of solutions). Also, D. Dubois was the first to describe the elementary single element with hyperincursion. One of the most important examples of a model with anticipation follows from modeling large social systems. Some examples of investigations of systems with strong anticipation were described in [9–12, 15–18]. We remark that one of the newest and most interesting properties in such systems is the possible multivaluedness of the solutions. Because of this, it is worth considering the properties of multivalued solutions from the point of view of synchronization investigations.

Thus multivaluedness as a phenomenon has many manifestations, at least in mathematical models of different natural and artificial objects. So further investigations of possible multivaluedness in mathematical problems are of interest. Moreover, the manifestation of multivaluedness and its interpretation can be important for further understanding of self-organization, computation theory, living system properties including consciousness, observability and measurements, complexity, and others. In fact, presumably all systems and problems that have been considered previously on the basis of models with single-valued solutions have a counterpart with multivalued solutions. Especially interesting may be interpretations related to nature and social behavior.

Systems with anticipation (with advanced effects) constitute one of the classes of systems with possible multivaluedness. We note that there now exist many examples of systems with anticipation: electromagnetic theory (R. Feynman, D. Dubois), nonlocal field theory (N. Nielsen), economic problems (L. Gardini), control problems, evolution on lattices, neuron models, social systems (A. Makarenko, L. Leydersdorff and D. Dubois), neuronets and consciousness (A. Makarenko), and many others. But till now, systems with anticipation (especially with strong anticipation) haven't been considered from the point of view of relevancy to self-organization theory and computation theory. So the first goal of this paper is to call attention to such problems. For this, we describe some examples of multivalued behavior in systems with strong anticipation. We also offer a discussion on possible directions for investigation and some achievements.

The structure of this paper is as follows. In Sect. 2, we give a brief description of strong anticipation. Section 3 is devoted to a short review of different systems with strong anticipation—cellular automata, neural networks, discrete equations, partial differential equations—and to the illustration of their multivalued solutions. A discussion of multivalued solution properties and possible research problems is presented in Sect. 4. Sections 5–7 deal with some possible consequences of strong anticipation in computation theory, consciousness theory, and uncertainty problems. For lack of sufficient space, we offer only a brief description of examples and ideas. Thus the main goal of this paper is to present a general view of multivaluedness in models with strong anticipation, some interpretations, and indications of prospects for further investigation.

2 Strong Anticipation Property

The idea of strong anticipation was introduced the early 1990 s in the works of D. Dubois; see [4, 5]:

Definition of an incursive discrete strong anticipatory system: an incursive discrete system is a system which computes its current state at time t, as a function of its states at past times $\ldots, t-3, t-2, t-1$, present time t, and even its states at future times $t+1, t+2, t+3, \ldots$

$$x(t+1) = A(\ldots, x(t-2), x(t-1), x(t), x(t+1), x(t+2), \ldots, p), \tag{1}$$

where the variable x at future times $t+1, t+2, t+3, \ldots$ is computed in using the equation itself.

Definition of an incursive discrete weak anticipatory system: an incursive discrete system is a system which computes its current state at time t, as a function of its states at past times $\ldots, t-3, t-2, t-1$, present time t, and even its predicted states at future times $t+1, t+2, t+3, \ldots$

$$x(t+1) = A(\ldots, x(t-2), x(t-1), x(t), x^*(t+1), x^*(t+2), \ldots, p), \tag{2}$$

where the variable x^* at future times $t+1, t+2, t+3, \ldots$ are computed in using the predictive model of the system [4].

Many results on systems with strong anticipation have been described in papers by D. Dubois, his coauthors, and followers (see [2]). Such investigations originated from formulations specific for computer science. But the great variety of such systems opens new possibilities for investigations. One of them is self-organization, or synergetics. So in this paper we will try to point out such possibilities especially with respect to multivaluedness of solutions.

Thus as a further research problem in the field of self-organization, one should consider systems with strong anticipation. We remark that the simplest yet rather general counterparts of Eqs. 1 and 2 are the equations for the continuous-time case.

3 Examples of Investigated Problems with Anticipation

In pursuit of the general goal of this paper, in this section we give several examples of strong anticipatory systems to illustrate the emergence of new peculiarities of their solutions. We give only a few illustrations because of lack of space.

In this section we briefly describe the possibilities of the strong anticipation property as manifested in some problems.

3.1 Two-Step Discrete-Time Anticipatory Models

The proposed model has two features (see details at [15]). The first is its two-step nature (that is, the values in two advanced steps are used in formulas). The second is that the function $f(x)$ in the model has a piecewise-linear character and looks like the transition function of neurons in neuronets [12]. We note that the piecewise character of the nonlinearity usually allows one to simplify the mathematical investigations.

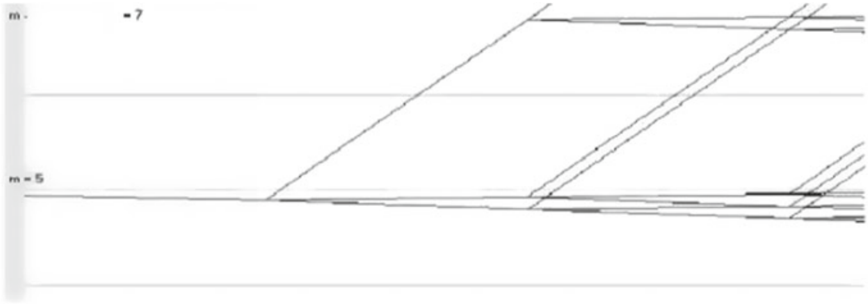


Fig. 1 Graph of branching of a solution. On the x-axis we display the discrete time steps, and on the y-axis we display the p_t values. The parameters values are $\alpha = 2$, $m = 0$. Limitations for visualization of the values of variables are as follows: maximum $m + 7$ and minimum $m - 5$, without representing the branches, which are thrown out to infinity

We write down the proposed model in discrete time steps as follows (here p_{t+1} is the value of the variable p at moment $(t + 1)$ of time, $t = 0, 1, 2, \ldots$)

$$
\begin{aligned}
p_{t+1} &= \alpha \cdot m + (1 - \alpha) p_t - \alpha \cdot f(p_{t+2} - p_{t+1}), \\
p_{t+2} &= \alpha \cdot m + (1 - \alpha) p_{t+1} - \alpha \cdot f(p_{t+2} - p_{t+1}).
\end{aligned}
$$

The function $f(x)$ depends on the parameter α and has the following expression:

$$
\begin{aligned}
f(x) &= 0 \qquad x \leq 0, \\
f(x) &= \alpha \cdot x \quad x \in \left(0, \frac{1}{\alpha}\right], \\
f(x) &= 1 \qquad x > \frac{1}{\alpha}.
\end{aligned}
$$

Software has been developed so that it is possible to visualize some branches of the states of the map, but only those that are not thrown out to infinity (see [15]). This facilitates understanding of the processes that take place in the region of ambiguity. In the case of the equations given above, we obtain a threefold solution (two branches tend to infinity and one branch remains in a restricted region of space). In Fig. 1 we display the branching of solution at discrete time steps. We display the discrete time steps on the x-axis and the value of p on the y-axis. In other calculations we have seen the increasing periods of cycles and increase in the number of different branches in the course of time, that is, we have seen a tendency to phenomena that we can call “chaos” (see Fig. 2).

In Fig. 2 we can see the similar behavior of solutions on different branches.

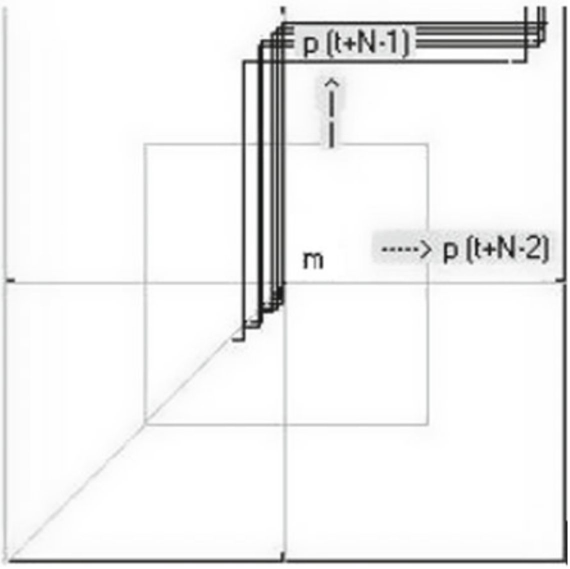


Fig. 2 Map in state space. The sequence of (p_t, p_{t+1}) values in the plane is represented for the solution, which corresponds to the parameters and initial conditions from Fig. 1

3.2 Logistic Equation with Strong Anticipation [9]

Another well-known example of discrete time equations with complex behavior is the classical logistic (or Verhulst) equation (see, for example, [18]). It is of interest to study the counterparts to this equation with strong anticipation taken into consideration.

The discrete dynamics equation with strong anticipation, which is a modification of the well-known logistic equation, was considered. The proposed equation has the form

$$x_{n+1} = \lambda x_n(1 - x_n) - \alpha x_{n+1}^2, \tag{3}$$

where $\alpha \neq 0$ (for $\alpha = 0$ we have the classical logistic map) is an anticipatory factor. Multivalued behavior and fractals were found. Examples of multivalued periodic solutions and chaos were considered. Also, multivalued fractal properties and bifurcations were investigated.

3.3 Cellular Automata with Anticipation

Another very general class of models is that of cellular automata. The basic issues of the classical cellular automata approach are a regular subdivision of space on equal (regular) cells, discrete time for considering the evolution of cell states, and special (local) rules for the dynamics of cell states (see [7, 16, 19]). Recently we have investigated such models with anticipation, namely the "Game of Life" with anticipation, the movement of pedestrian crowds, and other general problems [16].

First of all, let us recall a brief definition of cellular automata (CA). A cellular automaton is a discrete dynamical system that represents a set of identical cells and connections between them. The cells create the lattice of the cellular automaton. Lattices can be of different types, differing both in dimension and the form of the cells [16].

The local rule for cell k on the lattice Z_d is the transformation T_k that transforms the state $s_k(t)$ in S (here S denotes the space of all states of the cell) of the cell with index k at moment t to the state $s_k(t+1)$ in S of the same cell at moment $(t+1)$. Namely,

$$s_i(t+1) = G_i(\{s_i(t)\}, R), \tag{4}$$

where N_k is some neighborhood of the cell k on the lattice Z_d; $\{s_k(t)\}$ is the set of the cell's states within N_k; and the result of the transformation T_k depends only on the states of elements within the neighborhood N_k (locality); R is a set of parameters for rules that define the transformations. The collection of states of the cells at a given time is called a "configuration" (we denote it by $C(t)$). The collection of local transformations T_k defines the global transformation G on the configuration space C, where C is the space of all cell states: $C(t+1) = G(C(t))$.

The initial data configuration $C(0)$ is defined at initial moment $t = 0$. The set of transformations $\{T_k\}$ or transformation G defines the cellular automaton on the lattice Z_d with the cell's state space S.

In the case of strong anticipation, Eq. 4 in the additive case can have the form

$$s_i(t+1) = (1-\alpha)G_i(\{s_i(t)\}, R) + \alpha G_i(\{s_i(t+1)\}, R). \tag{5}$$

The factor α corresponds to accounting of strong anticipation. The case $\alpha = 0$ corresponds to the absence of anticipation.

3.4 The "Game of Life" as an Example of a Cellular Automaton

One of the basic examples of a CA is the well-known "Game of Life" by J. Conway. At time t, let some subset of the cells in the array be alive. In such a case, $S = \{0, 1\}$, where 1 corresponds to the living state and 0 to the dead state in some interpretations. The living cells at time $t+1$ are determined by those at time t according to the following evolutionary rules. Note that accounting of strong anticipation follows to the CA equations of the form

$$s_i(t+1) = G_i(\{s_i(t)\}, \ldots, \{s_i(t+g(i))\}, R). \tag{6}$$

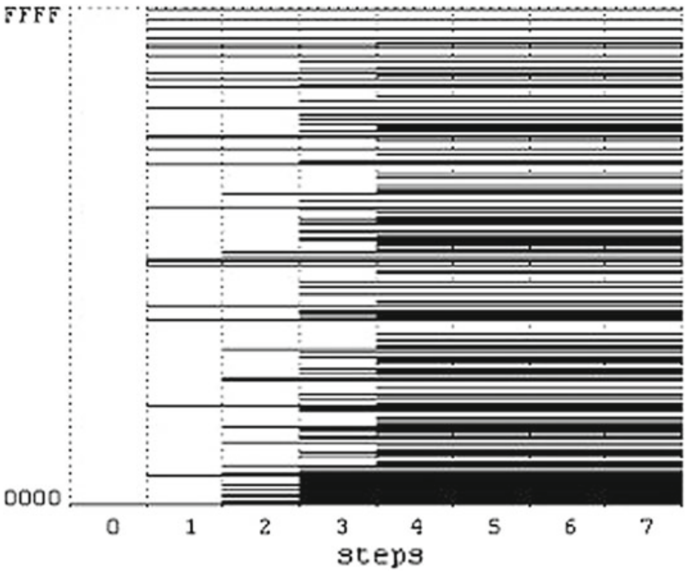


Fig. 3 A large number of configurations coexisting in a system (additive next state function, $\alpha = 0.5$, initial state is 0000). Discrete time moments are represented on the x-axis. The configuration's index of its collection of 16 cells is represented on the y-axis in a special alphabet

The main peculiarities of such an equation is that such equations can have multivalued solutions (of course within some range of parameters).

Figure 3 presents an example of a well-developed solution with multivaluedness within the model "LifeA." On the x-axis we display the values of discrete time steps. On the y-axis we show the configurations that exist at each given moment of time. The configurations are marked by special indices from 0000 to FFFF in a special alphabet. Figure 3 corresponds to a cellular automaton with 16 cells and periodic boundary conditions [16].

Such behavior is interesting and promising for investigations of self-organization. We remark that the main peculiarities in such systems are possible multivaluedness and spatial inhomogeneity of the solutions' behavior.

3.5 Neural Networks with Anticipation

Another class of very important object with anticipation and discrete time dynamics is neural networks with strong anticipation. We have already investigated some such models [12]. Here we propose only a very brief description of calculations for understanding their possible behavior. One of the simplest variants of such models with strong anticipation has the form (counterpart for Hopfield's neural networks)

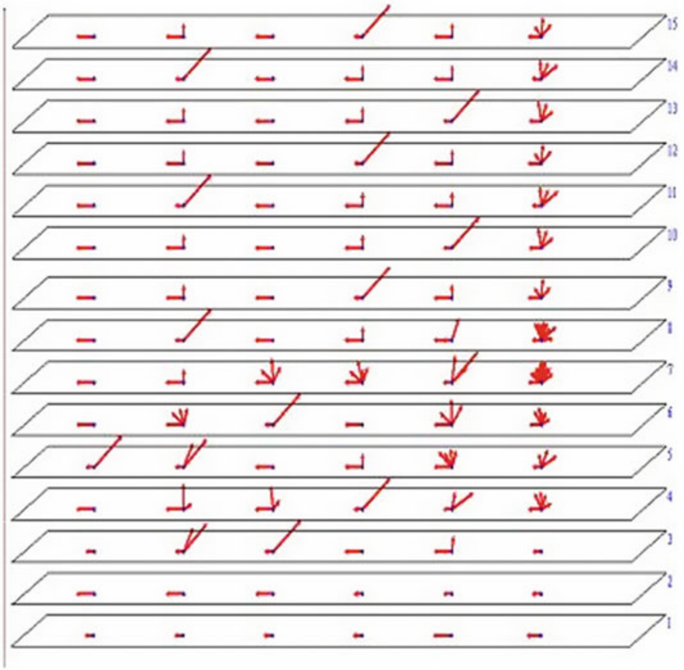


Fig. 4 Behavior of multivalued solutions. The numbers at the *right* side of the figure correspond to discrete time moments. The network has six elements. The inhomogeneous solution in space is seen. The length of an *arrow* corresponds to the value of the neuron's state. Many *arrows* originating from a given point in the figure correspond to multiple values of a cell's state. The length of the *arrow* corresponds to the value of the solution on a given branch at a given moment in time

$$x_j(n+1) = f\left((1-\alpha)\sum_{i=1}^{N} w_{ji}x_i(n) + \alpha\sum_{i=1}^{N} w_{ji}x_i(n+1)\right). \tag{7}$$

Here $x_j(n)$ is the value of the jth element at the moment n, w_{ji} is the weight matrix, and α is the anticipation parameter. The case $\alpha = 0$ corresponds to the absence of anticipation (with classical neuronet). In Fig. 4 we propose the example of possible behavior of the model of Eq. 7 at different time moments.

Two properties of artificial neural network solutions are represented in Fig. 4: multivaluedness and inhomogeneities in the elements' behavior. Both are rather new and of interest for the development of synchronization investigations both in theory and in applications.

3.6 Boundary Value Problems with Anticipatory Boundary Conditions

Distributed systems with continuous space and time variables constitute a special field of investigation in self-organization theory (since the works of I. Prigogine, H. Haken, A. Kolmogorov, A. Turing). The investigation of strong anticipation in such systems is only in its initial stages. So in this subsection and in Sect. 4 we offer a description of some research problems and interpretations of solutions in the multivalued case for distributed systems.

New examples of distributed systems with strong anticipation have been considered, namely systems of hyperbolic differential equations with special boundary conditions that include strong anticipation [11]. The mathematical formulation of problems and analytical formulas for solutions and interpretations of derived distributed solutions are proposed.

We consider some hyperbolic systems with anticipation. Such problems in the classical case originated in considering oscillations of an elastic string in a bounded domain, oscillations of voltage, and flows in transmission lines (see examples in [18], E.Yu. Romanenko and A. Sharkovski (1999) and references therein).

The basic classical problems have the following form. We consider a system of two first-order hyperbolic equations in a bounded space domain:

$$u_t + u_x = 0, \tag{8}$$

$$v_t - v_x = 0, \tag{9}$$

$$(x, t) \in \Omega = [0, 1] \times (-\infty, +\infty).$$

Initial conditions should be added to make the problem solution unique. For simplicity, in this paper we add the following conditions:

$$v(0, t) = \phi_1(x), \quad x \in [0, 1], \tag{10}$$

$$u(0, t) = \phi_2(x), \quad x \in [0, 1]. \tag{11}$$

Also, the boundary conditions should be considered for $x = 0$ and $x = 1$ and all $t \geq 0$ for completeness. Interesting results were obtained with the following boundary conditions in the works of A. Scharkovski and colleagues [18]:

$$u(0, t) = v(0, t), \quad t \in (-\infty, +\infty), \tag{12}$$

$$v(1, t) = f(u(1, t)), \quad t \in (-\infty, +\infty). \tag{13}$$

To discuss the possible effects of strong anticipation, we restrict attention in this chapter to the following form of boundary condition at the right point of the segment $x = 1$:

$$v(1, t + 1) = f(u(1, t + 1), u(0, t + 2)). \tag{14}$$

Eq. 14 replaces the condition of Eq. 13. Thus we need to study the equation

$$u(0, t+2) = f(u(0, t), u(0, t+2)), \tag{15}$$

similarly to how it was done for the problem of Eqs. 8–13.

Equation 15 can be a very complex object. Due to the possible multivaluedness of the solution, a strict mathematical formulation of results will be forthcoming.

We propose a multivalued counterpart for the complex behavior of solutions and their "ideal" turbulence from [18]. The main new property is the possibility of multivalued fields (that is, the existence of solutions of partial differential equations with many values at each space point at a given moment in time). This can be interesting for field equations, for example the Dirac equations.

Since [18], it has been known that a classical system without anticipation can have interesting and complex behavior: periodic, relaxation oscillations, oscillations with decreasing characteristic length, and chaotic regimes ("ideal" turbulence). But for examples with strong anticipation we can observe different behaviors on different branches, multivalued chaotic solutions, etc.

4 Possible Properties of Solutions and Further Research Problems in Physics

4.1 Multivaluedness and Self-Organization

Some generalizations for cases with strong anticipating can be proposed using the results described in Sect. 3. At first, the solutions in such cases can be multivalued, as we described above. So the fractal dimension of solutions can be greater than that in the single-valued case. Moreover, the behavior of the solutions can be much more complex. Different regimes on the different branches of solutions can exist (some examples of such different branches can be found in [15]). So each branch can have different stochastic properties, different types of relaxation behavior, etc. But probably the most interesting is the common behavior of a mixture of the branches in the multivalued case. Especially interesting is the problem of limiting behavior as $t \to \infty$: limit solutions, attractors of such solutions, probability properties of limiting objects. But new problems also arise: observability of solutions, selection of the single-valued trajectories of the system, complexity of its limiting objects and such complexity measures, searching adequate mathematical spaces for considering problems and solutions. In particular, the problems of adequate definition of periodic behavior (and also of dynamical chaos), Lemeray's staircase, Poincaré's bifurcation diagram, ergodicity, and mixing are of interest. Also, many other problems from self-organization theory [3] can be reconsidered with strong anticipation.

4.2 General Distributed Media with Anticipation

Many more models can be proposed for the case of continuous media with some kind of anticipation. The simplest variant is that in which only the nonlinear source $f(u)$ in the equations of such media (where $u(x, t)$ is the field value) has the anticipatory property, that is,

$$f_1 = f(u(x,t), u(x,t+\tau)) \tag{16}$$

in simplest case or

$$f_2 = \int\limits_{-\infty}^{+\infty} f(u(x,\tau))K(x,t-\tau)d\tau \tag{17}$$

in a more complex case. An example of such a model is a parabolic equation of diffusion with such a nonlinear source. As an example of such a system, we may propose models with continuous variables of neural activation fields (counterparts for known models of S. Amari or H. Wilson and J. Cowan). We remark that much more complex and well-developed models of such phenomena are the counterparts of well-known strict models of media with space nonlocality and memory in theoretical physics.

4.3 Oscillators and Chimeras and in Media with Memory, Nonlocality and Anticipation. New Research Possibilities

Many interesting new solutions have been found for distributed media: dynamical chaos, oscillations, autowaves, synchronization, and quite recently “chimera” states. Till now, mostly the classical parabolic equations of hydrodynamics and diffusion have been used (for example, the Navier–Stokes equations).

But now it is recognized that more accurate equations with memory and space nonlocality effects should be used (see, for example, [3]). Here we describe some possibilities for posing new research problems. First of all, we consider models with memory (relaxation in hydrodynamics). In such a case, one of the classes of models is infinite systems of ordinary differential equations of second order in time. Such systems recall systems of coupled oscillators. So the problems of energy transitions between different scales receive new solutions (transfer from large to small scales). Secondly, the new possibilities provide nonlocality. This follows from the possible origin of new “chimera” states in hydrodynamics. And finally, anticipation follows from the possibilities of multivalued “chimera” states in distributed systems.

4.4 Chimera States in Systems with Strong Anticipation

Recall that chimera states are highly inhomogeneous transitive solutions in which different types of behavior coexist in space [1]. The chimera states are the solutions in chains of elements or in distributed media that have different behaviors in different domains of the space. For example, such systems can have coexisting domains with chaotic and "smooth" behavior in different regions of space. One discrete system with possible chimeras is as follows [1]:

$$z_i^{t+1} = f(z_i^t) + \frac{\sigma}{2P} \sum_{j=i-P}^{i+P} (f(z_j^t) - f(z_i^t)). \tag{18}$$

Here z_i^t is the value of the ith element at discrete time t, σ is a parameter, and f is a nonlinear function.

The next example involves chains of oscillators:

$$\frac{d\psi_k(t)}{dt} = \omega - \frac{1}{2R} \sum_{j=k-R}^{k+R} \sin(\psi_k(t) - \psi_j(t) + \gamma), \tag{19}$$

where $\psi_k(t)$ is the value of the kth element at the time moment t.

The corresponding example to Eq. 18 with chimera state in case of strong anticipation is

$$\begin{aligned} z_i^{t+1} = (1-\alpha) \left(f(z_i^t) + \frac{\sigma}{2P} \sum_{j=i-P}^{i+P} (f(z_j^t) - f(z_i^t)) \right) + \\ \alpha \left(f(z_i^{t+1}) + \frac{\sigma}{2P} \sum_{j=i-P}^{i+P} .(f(z_j^{t+1}) - f(z_i^{t+1})) \right), \end{aligned} \tag{20}$$

where α is a coefficient of anticipation.

The corresponding example to Eq. 19 with a chimera state is

$$\begin{aligned} \frac{d\psi_k(t)}{dt} = \omega - \frac{1-\alpha}{2R} \sum_{j=k-R}^{k+R} \sin(\psi_k(t) - \psi_j(t) + \gamma) + \\ \frac{\alpha}{2R} \sum_{j=k-R}^{k+R} \sin(\psi_k(t+\tau) - \psi_j(t+\tau) + \gamma). \end{aligned} \tag{21}$$

It is accepted that the main source of chimera states is nonlocality in equations (see, for example, Eqs. 20, 21). We remark that the nonlocality described in [3] essentially expands the cases with possible origins of chimeras. But the possibilities of chimeras in systems with anticipation are absolutely new. For example, we should note the possibilities of ‘multivalued chimeras’ coexisting with different chimeras on different branches of the multivalued solutions and the coexistence of chimeras and “smooth” behavior in different branches of the solution.

4.5 Possible Research of Physical Systems

So far in this section we have described the possible properties of mathematical objects. Currently, the quantity of experiments on the properties of expressions of strict anticipation is very small (though they are made, especially in neuroscience and psychology). Therefore, we describe a hypothetical possibility concerning the physical realization of effects arising if we take the opportunity of physical realization of systems with the strong anticipation property (there are several works in favor of such opportunities, beginning with those of H. Tetrode, R. Feynman, D. Dubois et al.).

Sources of the laws of probability. In the case of physical realizability of strong anticipation, anticipation can be a model not only of probability laws, but also of their physical mechanism.

The disruption of monitoring and instability stochastics in experiments. In experimental work with complex systems with stochastic processes, sometimes failure monitoring is observed, i.e., the system suddenly jumps from one statistical law to another (e.g., according to Victor Ivanenko, the author of systems with amplifiers). We can assume that such a phenomenon could correspond to jumps from one branch to another of a multivalued solution.

The appearance of an ensemble of systems in statistical physics, particularly in the Boltzmann scheme. In statistical physics, the Boltzmann distribution came from calculation with ensembles (configurations). However, in this scheme, bands were purely speculative designs. Assuming strong anticipation, one can assume that the configuration corresponding to the branches in the multivalued structure is substantially parallel.

Problems of self-organization and synergy. Classic problems of self-organization can also obtain new interpretations primarily because of possible ambiguity. The choice of appropriate single-valued functionals among many possible is one of the problems. In addition, an interesting problem is to assess the degree of disequilibrium in the future, making polysemy and branching.

Generalized quantum mechanics, where the probabilities are set by anticipation. The usual quantum mechanics is a special case. In the classical scheme of quantum mechanics, there is an equation for the probability density that is complemented by the hypothesis of reduction (collapse) of the wave function of the measurements (the reduction of the system to a certain state). However, a question still remains about the physical sense, the laws of probability, and the sources of quantum mechanics. Assuming the existence of a strong physical anticipation, one can consider the laws of probability generating a mechanism with strong anticipation.

Everett's environment. There exists an interpretation of quantum mechanics by Everett that assumes the existence of a plurality of parallel branching worlds. However, the mechanism of occurrence of such a pattern has not yet been suggested. It makes sense to consider the possibility of generation of Everett's picture of the world through the mechanism of a type of strong anticipation. Generally, it is important to consider an environment with strong anticipation for which Everett's pattern is a special case.

Cellular models of quantum fields. It is known (G. t 'Hooft, H.-T. Elze, et al.) that the equations of classical quantum mechanics can be derived from models of cellular space on a quantum scale (Planck scale) under the limiting transition to zero cell length. Consideration of strong anticipation in the original cellular models (recall the anticipation approach to Eqs. 20, 21 in Sect. 4.4) can lead to differential equations for generalized quantum mechanics and just quantum gravity with anticipation. In such a case, the solutions of the processes of measurement correspond to choosing (or building) single-valued solutions by a measurement procedure.

5 Possible Ambiguity in the Theory of Computing Machines and Computer Architecture

One of the most promising new methodologies for modeling is the so-called cellular automata approach. According to this method, models are built from simple elements with local interaction with neighbors. The elements are distributed according to a rule based on simple geometry (lattices). Some details of the approach to a description of crowds and pedestrians will be described in the next section. But here we recall the Game of Life by J. Conway as the simplest example of a cellular automaton (CA). Recent investigations of cellular automata in physics and applied mathematics have shown that in spite of the simple description of the elements and rules for element dynamics, they can represent any phenomena in nature, including- self-organization, chaos, and complex behavior of an entire system. An outstanding example is modeling of hydrodynamic flows. For example, one of the CA applications is that simple (and frequently obvious) rules that correspond to a single pedestrian's behavior can lead to reliable modeling of a crowd as a complex object.

We describe new possibilities of CA in theory and applications including anticipation (advanced equations). We propose examples of CA with anticipation and some applications.

We propose a list of some properties and manifestations of anticipatory properties in different systems and processes: global sustainable development as strongly anticipative processes; regional sustainable development as weak anticipatory processes; origin of scenarios of evolution of complex social systems as the consequences of anticipation manifestation; self-referencing, reflexivity, and mentality aspects in anticipation agents; medical manifestation of anticipation including schizophrenia; new consciousness models that are based on the anticipatory effects in the brain and artificial intelligence; quantum-mechanical, microphysics, gravitation, and anticipatory analogies in the behavior of large complex systems.

Also, examples of anticipatory aspects in automata theory and computers have been considered. A new scheme for probability investigation is proposed for systems with anticipation. Neural network models, cellular automata for crowd movement, sports, communication, and social networks are proposed.

5.1 Possible Development of the Theory of Cellular Automata

In order to discuss the possible consequences of the new multivalued solutions in cellular automata, we first of all will analyze the evolution of computation theory and applications. After the initial period of automata theory development, the next stage came in connection with the study of the Game of Life by J. Conway. There were found many types of solutions of cellular automata, their relationship with the general theory of automata, the connection with formal language theory and formal grammars, etc. The next stage began with the involvement of methods of physics, primarily statistical methods in cellular automata studies. The beginning of the application of physics methods in cellular automata in the 1980s is associated with the names of S. Wolfram, U. Frisch, T. Toffoli, B. Chopard, et al. for the simulation of hydrodynamics. Attractors were studied in parallel for local and global maps, languages, and machines related to cellular automata, the possibility of using CA algorithms and programs, etc. Both deterministic and probabilistic cellular automata and sequential automata were constructed with a complicated structure (nonlocal, second-order (hierarchical), heterogeneous, with memory).

The next essentially new stage can be roughly related to the development of the theory of quantum computing, including quantum cellular automata. The beginning of this stage is usually connected with R. Feynman's articles (1982, 1986) for quantum computation. The fundamentally important accomplishments are the formulation of the concept of a quantum machine, including CA, quantum Turing machines, quantum computation, and logic applied to other quantum CA promising the formulation of local aspects of CA in terms of local algebras of operators in the spirit of the results of G. Haag.

Currently, all of the above areas continue to evolve (see the works by S. Mura, J. Crutchfield, T. Toffoli, B. Chopard, and many others). Cellular automata become more and more popular [7, 19]. Moreover, in view of hypotheses about the cellular structure of space-time at the micro level (the Planck length, D. Finkelshteyn's work (1973), and many others), it is sometimes declared that the universe is a giant cellular automaton with appropriate computational processes; see the book by S. Wolfram [19] (2002).

Note that the development of cellular automata occurred approximately as follows. There were certain concepts (such as classical or quantum computing) that then were embedded into the concept of cellular automata. Of course, the proposed cellular automata with anticipation [8] partially admit such an immersion into existing classical concepts. We call this the direct way of theory development. However, the solutions of the new CA models open new paths for the development of a theory that can be called the "inverse" or "reconstruction of concepts."

5.2 Problem of the Reconstruction of Concepts

Again, the development of cellular automata, both classical and quantum, follows a "direct" mode of development. That is, from original concepts, the CA were proposed, and then their properties and applications were studied. The study of CA has led to equivalent representations (languages, automata, dynamical systems). All of this is discussed in the framework of a predetermined structure of space-time. We can see that our research into CA with anticipation (and subsequent generalizations) has led to new and interesting results.

But there is a new source of problems, concepts, generalizations, and interpretations. We call this way "reverse" in research or "reconstruction" concepts. It comes from the fact that the primary source, the basic element of the CA unit cell, is given with its laws of the evolution of states. We assume that the existence of the enveloping concepts (metasystems) and others with known properties were not originally supposed to, and they can be built, removed, and identified only as properties of solutions of a CA system. For example, for conventional CA such concepts are: machine, language, dynamical system, equivalence, etc. The introduction of anticipation in CA leads to ambiguity in the solutions (we call these *cellular automata with multivaluedness*). To specify multiple meanings, we will add the letter M to the corresponding object. Then the concepts of the theory of automata with multivaluedness (AM) should be introduced. Turing machines (TM) that allow ambiguity can be marked as (TMM), algorithms with multivaluedness as (AlgM), formal languages with multivaluedness as (LM). Then it is necessary to revise the Church–Turing thesis on computability by means of automata (TChM) to the Turing test for artificial intelligence (AIMtest).

In physics, this leads to the possibility of considering multivalued statistical mechanics (SMM), multivalued cosmology (KM), multivalued quantum mechanics (CMM), and the multiple structures of matter, etc. Note that Everett's interpretation of quantum mechanics can be useful for such considerations. Conversely, multivalued CA can help in understanding Everett's picture of the universe. It seems that new opportunities for understanding the nature of probability can be proposed. More precisely, it can lead to new research on multiple-valued solutions in CA and evolutionary objects: multivalued chaos, cycles, bifurcations, Markov chains, attractors.

These are relatively obvious examples of "reconstruction" concepts. But no limits for generalizations exist. In fact, the structure of the "reverse" direction of research in CA can lead to generalized elements. Therefore, generalizations, first, can concern the state spaces of CA (or configurations). Also, integers and real numbers can change other options (nonstandard analysis, p-adic analysis; ∞-structure, logical calculus, topological objects (e.g., homology), distributions). Then state operators, secondary operators (operads), etc., can be proposed. To all this diversity, more general rules—generalized operators, multivalued, discrete, or continuous delay (and anticipation)—can be considered. There can be different definitions of neighborhoods (including implicit, fuzzy, time-dependent, and state). Further generalizations can be viewed using the CA as CA cells with elements of artificial intelligence. Thus, elements of the CA can be self-reflective elements with infinite recurrence level. Each proposed architecture can induce construction of new structures, etc. We note only that the obvious way along this path suggests the language of categories and functors (including facilities for the study of the limit), which is becoming commonplace in theoretical computer science and quantum field theory. We also point out that many of the issues discussed can also be applied to the model equations in the form of another type (e.g., differential).

Some features of cellular automata with strong anticipation are described in [8]. We stress here only some interesting features. Firstly, we mention the application in the theory of automata and cellular automata. Cellular automata with strong anticipation can implement nonclassical logic and can be non-Turing machines, which are important for hypercomputation. In particular, the possible worlds approach of S. Kripke and J. Hintikka as well as the ideas of N. Belnap can be implemented. Consideration of the implementation of computers in strong anticipative elements is promising. Also it is possible to consider issues in consciousness and effects on cognitive science [13].

6 Interpretation of Mental Reactions and the Problem of Consciousness

Recently, studies of processes in the brain have attracted more and more attention. This list of achievements is impressive: a model of a single neuron, the theory of neural networks, neuroinformatics, cognitive science, brain–computer interface, and many others.

However, we are still far from a final solution (and even from the existence of a common paradigm) because of the object's complexity. One of the key problems is the explanation of the relationship of such entities as brain, mind, and consciousness.

In view of these uncertainties, it seems appropriate to consider different factors and phenomena, including those that have not yet been fully recognized. Naturally, there is a large number of factors. However, one of the most promising is the manifestation of the foresight properties (or more precisely of anticipation in the sense indicated below).

In this regard, this chapter describes some of the already established facts about anticipation in living systems, the properties of neural networks with anticipation, and especially the possible consequences for the consideration of the problem of consciousness.

6.1 Property of Anticipation (Foresight) in Theory and Nature

There are many ways to describe anticipation. Perhaps for intuitive understanding, the closest explanation is “Anticipation is the presentation of the result of a process that occurs before its real achievements and serving as a means of feedback in the construction activities." This concept can be met, though without being formalized, many times in the context of philosophy, economics, psychology, and medicine. More recently (probably in the last thirty or forty years), anticipation emerged as a topic of experimental research and new theoretical concepts. In the neurosciences, the most famous investigations were made by B. Libet and numerous subsequent studies (a review can be found, for example, in [6] and many others).

On the theoretical level, in the field of biology and modeling, systems and models have explicitly described anticipation (R. Rosen). Significant development and formalization of the anticipation concept was introduced by D. Dubois.

One of the accepted concepts in classical neuroscience is to consider brain processes within the paradigm of neural network architecture for activity of the brain. Here we describe some possible consequences of strong anticipation in such models of brain activity.

6.2 The Possible Role of Strong Anticipation in Thinking

Section 3.5 briefly illustrates the results of the study of neural network models based on strong anticipation. The main new feature of this behavior compared to other models is a multivalued solution in the sense of simultaneous multiple virtual item states. Naturally, the simulation results indicate the need for further search analogues and manifestations of such effects in actual experiments. However, even the current results indicate a possible utility of the concept of powerful anticipation. So here we present some of the possible consequences of the adoption of the existence of effects such as strong anticipation.

Firstly let us discuss the problem of consciousness. Currently, there are many concepts of consciousness, such as psychological, biological, philosophical, and physical (unfortunately, it is impossible to give references in such a short paper). We describe how the phenomenon of consciousness might look like in the case of strong anticipating accounting. The physical elements of the brain on different levels (from neurons to microtubules and Hameroff cells) can assume many virtual states. The examples how it may looks can be found in [16] in cellular automata case – the Game of Life in view of the strong anticipation. Then the elementary act of consciousness is the realization of one of the plurality of virtual states. Incidentally, it is strongly reminiscent of usual quantum mechanics complete with a measurement process (especially in Everett’s interpretation). Acceptance of this interpretation can be useful in real applications. This concept can be applied, for example, to the problem of schizophrenia. Then schizophrenia can be associated with a disorder of the mechanism of choice of the "ordinary" person among the plurality of virtual personalities. Also, a virtual set of states can be correlated with uncertainty in mental processes. Also, B. Libet's experiments on anticipation can correspond to a slow unconscious evaluation of virtual values of states, which is then accompanied by a conscious reaction and choice of a single state from among many possibilities.

7 Uncertainty and Probability

The history of the spread of probabilistic concepts has more than three hundred years of development, consisting of a stream of outstanding discoveries, and mathematical and physical interpretations of the results. However, the main problem of the sources, meaning, nature, and adequacy of the laws of probability does not yet have a final solution. This is evident from the steady stream of publications on these issues. Although nowadays there have appeared many surveys and monographs, the latest studies describe new problems and approaches to major issues. Following the first modern research and review, it is possible to allocate such directions conditionally:

- classical approach with equal outcomes (from the eighteenth century);
- frequency approach (primarily associated with the name of R. von Mises);
- the theory of possibilities (propensity theory);
- axioms of probability theory associated with the name of A. Kolmogorov;
- logical approach;
- probabilistic aspects of decision theory and game approach, including upper and lower probability;
- quantum probability;
- nonprobabilistic uncertainty (V.I. Ivanenko);
- probability and foundations of physics;
- applied statistics and others.

Generally speaking, these areas can be arbitrarily structured in four blocks.

1. The axiomatic construction of the foundations of the theory.
2. Computation of probability characteristics of distributions etc., particularly in prior centuries (for example, Markov chains).
3. Identification of new problems and systems in which it is possible to use probabilistic schemes for the calculation of uncertainties.
4. The physical interpretation of probability structures (for example, in statistical physics).

Recently, in blocks 3 and 4, many new ideas (such as p-adic probability) have introduced the concept of parastatistics problems of probability, decision-making approaches in quantum mechanics, and multiple nondeterministic chaos. As a rule, these new applications of probability schemes appear in connection with the consideration of new mathematical problems describing various real processes.

In this chapter we have briefly noted new possibilities of research on the basis of uncertainty properties of systems with strong anticipation. The ideas described here are mostly relevant to block 3 above, that the study of models in which the manifestation of the properties that simulate probability is possible. The basic design centers on the possible emergence of ambiguities in the evolution of solutions in problems with strong anticipation and the origin in this uncertainty of the multivaluedness of solutions and the emergence of many possible paths for given values of solutions [14]. This can lead to the analogue of classical counting "frequencies of realizations of states" in probable events. A general scheme of such a structure and specific examples, particularly cellular automata and artificial neural networks, were given. We also discussed the hypothesis of a possible physical realization of such systems and their relationship with the occurrence of the actual laws of probability.

8 Conclusion

In this chapter we have considered the problem of complex behavior for a new class of objects, namely for chains and networks with strong anticipation. The possible multivaluedness of solutions leads to new interesting properties in the framework of existing concepts. But new features can also appear (for example nonheterogeneous multivalued synchronization) that are very promising for further investigation. We described only the first results of investigations and only some new possible forms of research problems. But the obvious mathematical novelty of the proposed problems and the possible great importance to applications (for example, for social systems, computation and signal processing theory, consciousness investigations) leads to the need for further investigation.

Thus, this chapter described examples of the introduction of anticipation in terms of cellular automata, neural networks, and differential and discrete equations. One of the important results is the possibility of the emergence of multivalued solutions. An immediate consequence is the possibility of posing problems of multivalued machines, languages, chaos, etc. Moreover, the results on multivalued solutions (including multivalued solutions by strong anticipation) described in this chapter can lead to hypotheses for discussion that can have a multivalued nature but for which we, as observers, can usually see only a single-valued picture of the world.

Acknowledgements The author would like to express his appreciation to D. Krusinsky, Yu. Yatsuk, V. Biliuga, A. Stashenko, and S. Lazarenko for computer realizations of models and to V. Statkevich for assistance in preparing the text of this chapter.

References

1. Abrams, D., Strogatz, S.: Chimera states for coupled oscillators. Phys. Rev. Lett. **93**, 174102–174105 (2004)
2. Dubois, D.M. (ed.): Int. J. Comput. Anticip. Syst. **26** (2014). www.sia.hec.ulg.ac.be/CASYSIJCAS-VOLUME-26-TOC.pdf
3. Danilenko, V., Danevich, T., Makarenko, A., Skurativskyi, S., Vladimirov, V.: Self-organization in Nonlocal Non-equilibrium Media. Kyiv, Subbotin S.I. Institute of Geophysics, NAS of Ukraine (2011)
4. Dubois, D.: Generation of fractals from incursive automata, digital diffusion and wave equation systems. BioSystems **43**, 97–114 (1997)
5. Dubois, D.: Incursive and hyperincursive systems, fractal machine and anticipatory logic. In: Dubois, D. (ed.) AIP Conference Proceedings, vol. 573, pp. 437–451 (2001)
6. Dubois, D.: Natural and Artificial intelligence, language, consciousness, emotions and anticipation. In: Dubois, D. (ed.) AIP Conference Proceedings, vol. 1303, pp. 236–245 (2010)
7. Ilachinski, A.: Cellular Automata. A Discrete Universe. World Scientific Publishing, Singapore (2001)
8. Krushinskiy, D., Makarenko A.: Cellular automata with anticipation. examples and presumable applications. In: Dubois, D. (ed.) AIP Conference Proceedings, vol. 1303, pp. 246–254 (2010)
9. Lazarenko, S., Makarenko, A.: Investigation of complex multivalued solutions in discrete dynamical systems with anticipation. In: Abstracts Book of 10th International Conference CASYS 2011, Liege, Belgium, August 2011, p. 1 (2011)
10. Makarenko, A.: Anticipating in modeling of large social systems - neuronets with internal structure and multivaluedness. Int. J. Comput. Anticip. Syst. **13**, 77–92 (2002)
11. Makarenko, A.: Some distributed systems with anticipation. Int. J. Comput. Anticip. Syst. **24**, 21–32 (2010)
12. Makarenko, A.: Neural networks with anticipation and some problems of complexity theory. In: Proceedings of International Conference on Complex Systems: Synergy of Control, Communications and Computing – COSY 2011, Ohrid, Macedonia, pp. 257–262 (2011)
13. Makarenko, A.: Possible effects of anticipation role in neurophysiology and thinking (short abstracts). In: Proceedings of the XVI International Conference On Neurocybernetics (ICNC–12), 24-28 September 2012, Rostov-on-Don, p. 3 (in Russian)
14. Makarenko, A.: On some model construction with ambigiuty and simulation of probability laws. 1. Elementary examples and posing of problems. Analysis, models and control **1**, Book of papers. ESC “IASA” NTUU “KPI”, Kyiv, pp. 54–72 (2013) (in Russian)
15. Makarenko, A., Stashenko, A.: Some two- steps discrete-time anticipatory models with ‘boiling’ multivaluedness. In: Dubois, D. (ed.) AIP Conference Proceedings, vol. 839, pp. 265–272 (2006)
16. Makarenko, A., Goldengorin, B., Krushinsky, D.: Game ‘life’ with anticipatory property. In: Umeo, H. et al (eds.) Proceedings of the International Conference ACRI’08. LNCS, vol. 5191, pp. 77–82 (2008)
17. Rosen R.: Anticipatory Systems. 2nd edn. Springer, Berlin (2012)
18. Scharkovski, A., Maystrenko, Yu., Romanenko, E.: Difference Equations and their Applications. Kluwer, Dordrecht (1993)
19. Wolfram, S.: New Kind of Science. Wolfram Media Inc., USA (2002)

Part II
Nonlinear Dynamics—Selected Applications

Nonlinear Modeling of Continuous-Wave Spin Detection Using Oscillator-Based ESR-on-a-Chip Sensors

Jens Anders

Abstract In this chapter, an advanced nonlinear energy-based modeling of LC tank oscillators used as sensors for ensembles of electron or nuclear spins is presented. Recently, this oscillator-based sensing principle has been gaining significant attention in the electron spin resonance community for biomedical and material science applications. Since the sensing principle relies on the coupling between a harmonic oscillator (the spin ensemble) and an intrinsically nonlinear electrical oscillator, it presents an excellent example of the high relevance of nonlinear dynamical systems modeling for practical sensing applications. In order to provide a self-contained overview, after a short general motivation that highlights the relevance of the topic, the chapter begins with a description of the experimental setup of the oscillator-based spin-detection approach, which is somewhat different from that for conventional resonator-based detection. In this description, it is shown how continuous-wave spin-detection experiments can be carried out using LC tank oscillators by monitoring the oscillation frequency when sweeping the static magnetic field B_0. At this point, it is also explained how standard field modulation using a modulation field B_m parallel to B_0 can be used to increase the signal-to-noise ratio by means of phase-sensitive detection using a conventional lock-in amplifier. Then the interaction between the nonlinear electrical oscillator and the spin ensemble is modeled using the solution of the Bloch equation in the steady state, which models the dynamics of the spin ensemble, and the magnetic energy associated with the inductor of the LC tank oscillator. In this way, under steady-state conditions as they occur in continuous-wave ESR and NMR experiments, expressions for spin-related changes in inductance and resistance can be derived, which are in turn related to changes in the oscillation frequency and amplitude of the LC tank oscillator. To quantify the resulting change in oscillation frequency and also derive expressions for the expected noise floor, which eventually determines the achievable limit of detection, the chapter then provides a detailed discussion of the nonlinear modeling of LC tank oscillators in the presence of noise. The resulting model of the LC tank oscillator is subsequently used to find analytical

J. Anders (✉)
University of Ulm, Insitute of Microelectronics, Albert-Einstein-Allee 43,
89081 Ulm, Germany
e-mail: jens.anders@uni-ulm.de

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_4

expressions for the limit of detection of frequency-sensitive oscillator-based spin detectors. Finally, experimental results from a prototype realization of an oscillator-based CMOS ESR-on-a-chip detector are used to validate the accuracy of the derived signal and noise models.

1 Motivation

Over the past ten years, miniaturized inductive spin detectors have gained significant attention in the research community [1, 2, 6, 7, 9–11, 15, 16, 18, 27, 29, 30, 34, 35, 38–40, 42]. Here in the nuclear magnetic resonance (NMR) community, most research in this area focuses either on miniaturizing the classical resonator-based approach using MEMS-like techniques [9, 10, 15, 27, 34, 35] or on using methods based on integrated circuit (IC) technology [2, 6, 7, 18, 38] or a combination of the two [9, 38]. In contrast, in the electron spin resonance (ESR) community, due to the more stringent requirements on the detector's operating frequency, in addition to miniaturized resonator-based detectors [11, 29, 30, 39, 40], since 2008, cf. [42], also a novel approach that uses an LC tank oscillator to sense the spin ensemble is used. This ESR-on-a-chip oscillator-based detection approach has recently gained significant interest in the ESR community [1, 16, 21, 42], because in a given CMOS (complementary metal oxide semiconductor) technology, oscillators can be realized at higher operating frequencies than the amplifiers required for resonator-based detection. Since the achievable signal-to-noise ratio (SNR) in inductively detected spin resonance experiments improves almost quadratically with said operating frequency, the oscillator-based approach has the potential to significantly improve the achievable sensitivity of inductively detected spin resonance experiments.

2 Introduction: Oscillator-Based Spin Detection

Oscillator-based frequency-sensitive spin sensing was first introduced in the context of electron spin resonance (ESR) experiments in [42] and further improved in [1] and [16]. The required experimental setup is shown conceptually in Fig. 1. This setup is somewhat different from conventional resonator-based ESR experiments; cf., e.g., [32]: In the conventional resonator-based approach, a resonator containing the spin ensemble is irradiated through a circulator with a fixed-frequency RF-signal at a frequency ω_{RF}. The resonator converts the incident power into a fixed-frequency RF-magnetic field—the so-called B_1-field—at the sample location. This fixed-frequency B_1-field can then trigger the transitions between the different energy levels associated with different spin states if its frequency matches the resonant condition $\omega_{\mathrm{RF}} = -\gamma B_0$, where γ is the gyromagnetic ratio of the particle under

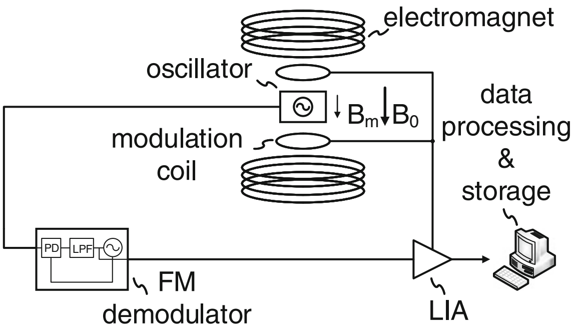


Fig. 1 Illustration of the experimental setup using a fixed-frequency oscillator for continuous-wave spin-detection experiments

investigation,[1] and B_0 is the applied static magnetic field. In the resonator-based detection scheme, a spectrum is then recorded by sweeping the B_0-field in and out of resonance and measuring the change in the reflected power caused by the change in sample magnetization due to the spin resonance effect; cf. [32]. In contrast, in the frequency-sensitive oscillator-based approach of Fig. 1, the detector is realized as an LC-tank oscillator that both produces the B_1-field necessary to excite the spin ensemble and detects the resulting spin-resonance-induced change in sample magnetization as a change in its oscillation frequency.

More specifically, to perform an oscillator-based spin-detection experiment according to Fig. 1, the nominal oscillation frequency $\omega_{\mathrm{osc},0}$, i.e., the oscillation frequency in the absence of the spin resonance effect, is chosen to lie in the desired frequency range, corresponding to the center static magnetic field strength of the utilized magnet. The spin resonance effect is then detected by observing the frequency change of the oscillator as a result of a corresponding change in the spin-related sample magnetization as the applied B_0-field is swept through the resonance condition $B_{0,\mathrm{res}} = -\omega_{\mathrm{osc},0}/\gamma$. This change in oscillation frequency can occur because in contrast to conventional resonator-based spin-detection experiments, the LC-tank oscillator, which produces the B_1-field, is bidirectionally coupled to the spin ensemble; i.e., not only does the B_1-field generated by the oscillator perturb the spin ensemble, but the spin system can also influence the oscillator. An example spectrum of such a continuous-wave spin-detection experiment using an oscillator as detector is shown in Fig. 2a. According to the figure, for B_0-fields far away from the resonance condition, the oscillator oscillates with its nominal oscillation frequency $\omega_{\mathrm{osc},0}$. As the B_0-field is swept through the resonance condition $B_{0,\mathrm{res}} = -\omega_{\mathrm{osc},0}/\gamma$, a resonant change in the oscillation frequency, which displays a dispersive line shape, can be observed. Since typical B_0-field sweeps according to Fig. 2a are performed at low sweep rates, the resulting so-called direct detection spectra are plagued by low-frequency noise and drifts in the setup. As a solution, standard continuous-wave spin-detection experiments employ so-called field modulation,[2] which is illustrated

[1] As an example, for an isolated hydrogen nucleus and an isolated electron, γ takes on values of $\gamma_{^1\mathrm{H}} = 2\pi \cdot 42.576\,\mathrm{MHz}$ and $\gamma_{\mathrm{e}^-} = -2\pi \cdot 28.024\,\mathrm{GHz}$, respectively.

[2] This is also the case for resonator-based detection.

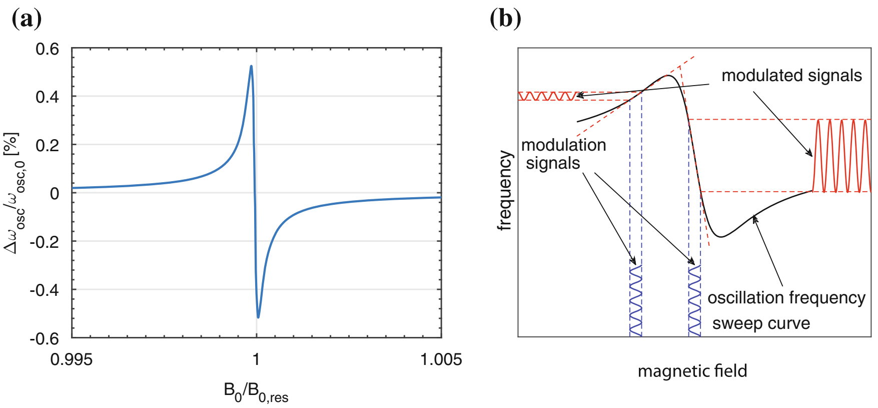


Fig. 2 **a** Illustration of the direct detection spectrum when a B_0-field sweep is used to perform an oscillator-based spin-detection experiment. **b** Illustration of field modulation for SNR enhancement in continuous-wave spin-detection experiments

in Fig. 2b. According to the figure, at every B_0-field point a modulating magnetic field B_m, which is pointing in the same direction as the B_0-field, is used to transfer the resulting frequency shift to the frequency of the modulating field f_m. If f_m is chosen sufficiently large, i.e., outside the spectral region corrupted by low-frequency noise and drifts, a significant improvement in the achievable SNR can be achieved. Here it should be noted that because the spin resonance information is encoded in the frequency of the oscillation, an FM-demodulation, e.g., using a PLL-based FM-demodulator, needs to be performed before a standard lock-in amplifier can be used to extract the spin resonance information at the modulation frequency f_m.

After this qualitative description of spin resonance experiments using LC tank oscillators, the remainder of the chapter will be devoted to quantifying both the expected signal, i.e., the spin-induced change in oscillation frequency, and the expected noise floor, which is defined by the frequency noise of the LC tank oscillator. More specifically, in order to evaluate the performance of frequency-sensitive oscillator-based spin detectors compared to other inductive spin-detection schemes, we will derive analytical expressions for the achievable spin sensitivity of the detector; cf. [12].

3 Magnetic-Energy-Based Modeling of the Spin-Induced Inductance Change in Continuous-Wave Oscillator-Based Spin-Detection Experiments

The interaction of an ensemble of noncoupled spins represented by its corresponding magnetization $\mathbf{M} = (M_x, M_y, M_z)$ with an applied magnetic field $\mathbf{B}$ can be described using the famous Bloch equation, cf., e.g., [41], which in the laboratory frame of reference reads

$$\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} = \gamma \cdot \mathbf{M} \times \mathbf{B} - \frac{M_\mathrm{x} \cdot \mathbf{e}_\mathrm{x} + M_\mathrm{y} \cdot \mathbf{e}_\mathrm{y}}{T_2} - \frac{M_\mathrm{z} - M_0}{T_1} \cdot \mathbf{e}_\mathrm{z}, \tag{1}$$

where γ is the gyromagnetic ratio, $\mathbf{e}_\mathrm{x}$, $\mathbf{e}_\mathrm{y}$, and $\mathbf{e}_\mathrm{z}$ are the unit vectors in the x-, y- and z-directions of the laboratory reference frame, M_0 is the steady-state sample magnetization, and T_1 and T_2 are the longitudinal and transversal relaxation times, respectively.

To model spin-detection experiments using the Bloch equation, it turns out to be convenient to transform the Bloch equation from the laboratory into a rotating frame of reference, which rotates with an angular frequency ω_F around the z-axis of the laboratory frame. Since for a continuous-wave spin-detection experiment, the applied magnetic field in the most general case can be written according to $\mathbf{B} = 2\,B_{1\mathrm{x}} \cdot \cos(\omega_{\mathrm{B}1} t) \cdot \mathbf{e}_\mathrm{x} + 2\,B_{1\mathrm{y}} \cdot \cos(\omega_{\mathrm{B}1} t) \cdot \mathbf{e}_\mathrm{y} + B_0 \cdot \mathbf{e}_\mathrm{z}$, the most convenient choice for the angular velocity of the rotating frame is $\omega_\mathrm{F} = \omega_{\mathrm{B}1}$. With this choice of ω_F and ignoring frequency components at $2\,\omega_{\mathrm{B}1}$, which are far away from resonance and have therefore a minimal effect on the magnetization, the Bloch equation in the rotating frame becomes

$$\frac{\mathrm{d}\mathbf{M}'}{\mathrm{d}t} = \gamma \cdot \mathbf{M}' \times \mathbf{B}_\mathrm{eff} - \frac{M_{\mathrm{x}'} \cdot \mathbf{e}_{\mathrm{x}'} + M_{\mathrm{y}'} \cdot \mathbf{e}_{\mathrm{y}'}}{T_2} - \frac{M_{\mathrm{z}'} - M_0}{T_1} \cdot \mathbf{e}_{\mathrm{z}'}, \tag{2}$$

where $\mathbf{M}' = (M_{\mathrm{x}'}, M_{\mathrm{y}'}, M_{\mathrm{z}'})$, $\mathbf{e}_{\mathrm{x}'}$, $\mathbf{e}_{\mathrm{y}'}$, and $\mathbf{e}_{\mathrm{z}'}$ are the unit vectors along the x'-, y'-, and z'-axes of the rotating frame of reference[3] and

$$\mathbf{B}_\mathrm{eff} = \left(B_0 + \frac{\omega_{\mathrm{B}1}}{\gamma}\right) \cdot \mathbf{e}_{\mathrm{z}'} + B_{1\mathrm{x}} \cdot \mathbf{e}_{\mathrm{x}'} + B_{1\mathrm{y}} \cdot \mathbf{e}_{\mathrm{y}'}. \tag{3}$$

Therefore, the individual components of Eq. (2), can be written according to

$$\frac{\mathrm{d}M_{\mathrm{x}'}}{\mathrm{d}t} = \Delta\omega \cdot M_{\mathrm{y}'} - \gamma\, B_{1\mathrm{y}}\, M_{\mathrm{z}'} - \frac{M_{\mathrm{x}'}}{T_2} \tag{4a}$$

$$\frac{\mathrm{d}M_{\mathrm{y}'}}{\mathrm{d}t} = \gamma\, B_{1\mathrm{x}}\, M_{\mathrm{z}'} - \Delta\omega \cdot M_{\mathrm{x}'} - \frac{M_{\mathrm{y}'}}{T_2} \tag{4b}$$

$$\frac{\mathrm{d}M_{\mathrm{z}'}}{\mathrm{d}t} = -\gamma\, B_{1\mathrm{x}}\, M_{\mathrm{y}'} + \gamma\, B_{1\mathrm{y}}\, M_{\mathrm{x}'} - \frac{M_{\mathrm{z}'} - M_0}{T_2}, \tag{4c}$$

where $\Delta\omega = \omega_{\mathrm{B}1} - \omega_L$ and $\omega_L = -\gamma\, B_0$ is the so-called Larmor frequency. From Eq. (4), the resonant nature of the excitation through the RF-magnetic field B_1 becomes obvious: for a choice of $\omega_{\mathrm{B}1} = \omega_L = -\gamma B_0$, the effective magnetic field in the rotating frame of reference, $\mathbf{B}_\mathrm{eff}$, has no residual component in the z'-direction,

[3] Note that $\mathbf{e}_\mathrm{z} = \mathbf{e}_{\mathrm{z}'}$.

and the terms $\Delta\omega$ in Eq. (4) evaluate to zero. As a consequence, the magnetization rotates about the axis of the applied B_1-field. For a choice of $\omega_{\mathrm{B1}} \neq \omega_L$, $\mathbf{B}_{\mathrm{eff}}$ has a residual component in the z'-direction that alters the axis of rotation and the "efficiency" of the resulting rotation. Since the rotation of the net magnetization is essentially "efficient"[4] only for small offsets $\Delta\omega$ between ω_{B1} and ω_L, the spin evolution described by the Bloch equation is clearly a resonant phenomenon. At this point it should also be emphasized that due to the resonant nature of the Bloch equation, of the two circularly polarized components contained in the linearly polarized B_1-field only one is in (or near) resonance with the spin ensemble at a given B_0-field according to $\omega_{\mathrm{B1}} = \omega_L = -\gamma B_0$. The nonresonant component, which in a frame of reference rotating at $\omega_F = \omega_{\mathrm{B1}} = \omega_L$ is spectrally located at $2\omega_{\mathrm{B1}}$, has only a marginal effect on the evolution of the spin magnetization; cf. Bloch-Siegert shift [37]. Which of the two circularly polarized fields is in resonance is—for a given B_0-field—determined by the sign of the gyromagnetic ratio γ.

The solution of Eq. (4) in the steady state, i.e., for $\mathrm{d}\mathbf{M}'/\mathrm{d}t = 0$, can be written as

$$M_{\mathrm{x}',\mathrm{ss}} = \frac{T_2\left(\Delta\omega\, T_2\, \gamma\, B_{1x} - \gamma\, B_{1y}\right)}{1 + \Delta\omega^2\, T_2^2 + T_1\, T_2\, \left((\gamma\, B_{1x})^2 + \left(\gamma\, B_{1y}\right)^2\right)} \cdot M_0 \tag{5a}$$

$$M_{\mathrm{y}',\mathrm{ss}} = \frac{T_2\left(\gamma\, B_{1x} + \Delta\omega\, T_2\, \gamma\, B_{1y}\right)}{1 + \Delta\omega^2\, T_2^2 + T_1\, T_2\, \left((\gamma\, B_{1x})^2 + \left(\gamma\, B_{1y}\right)^2\right)} \cdot M_0 \tag{5b}$$

$$M_{\mathrm{z}',\mathrm{ss}} = \frac{1 + \Delta\omega^2\, T_2^2}{1 + \Delta\omega^2\, T_2^2 + T_1\, T_2\, \left((\gamma\, B_{1x})^2 + \left(\gamma\, B_{1y}\right)^2\right)} \cdot M_0. \tag{5c}$$

The corresponding solution in the laboratory frame of reference can be calculated from the results of Eq. (5) according to

$$M_{\mathrm{x},\mathrm{ss}} = M_{\mathrm{x}',\mathrm{ss}} \cos\left(\omega_{\mathrm{B1}} t\right) - M_{\mathrm{y}',\mathrm{ss}} \sin\left(\omega_{\mathrm{B1}} t\right) \tag{6a}$$

$$M_{\mathrm{y},\mathrm{ss}} = M_{\mathrm{x}',\mathrm{ss}} \sin\left(\omega_{\mathrm{B1}} t\right) + M_{\mathrm{y}',\mathrm{ss}} \cos\left(\omega_{\mathrm{B1}} t\right) \tag{6b}$$

$$M_{\mathrm{z},\mathrm{ss}} = M_{\mathrm{z}',\mathrm{ss}}. \tag{6c}$$

The transversal components of the magnetization in the laboratory frame can also be written using a phasor notation, which will become convenient later, where the time-domain signals are related to the phasors according to

$$M(t) = \hat{M} \cdot \cos\left(\omega t + \varphi_M\right) = \mathrm{Re}\left\{\bar{M} \cdot \exp\left[\omega t\right]\right\} = \mathrm{Re}\left\{\hat{M} \cdot \exp\left[j\varphi_M\right] \cdot \exp\left[\omega t\right]\right\}, \tag{7}$$

where $\bar{M} = \hat{M} \cdot \exp\left[j\varphi_M\right]$ is the complex phasor with amplitude $\hat{M}$ and phase φ_M.

[4] Here efficient refers to the fact that small—compared to the static B_0-field—perpendicular fields B_1 produce a significant rotation of the magnetization $\mathbf{M}$ away from its equilibrium orientation along the z-axis.

Introducing this phasor notation into Eq. (6), the transversal magnetization in the laboratory frame can be written as

$$\bar{M}_{\mathrm{x,ss}} = M_{\mathrm{x',ss}} + j\, M_{\mathrm{y',ss}} \tag{8a}$$
$$\bar{M}_{\mathrm{y,ss}} = -j\, M_{\mathrm{x',ss}} + M_{\mathrm{y'}}. \tag{8b}$$

In the following, we will relate the steady-state magnetization in a continuous-wave experiment given by Eqs. (5) and (6) to a change in impedance of a coil filled with said magnetization. To this end, we can first start with the general relation between the **H**-field and the **B**-field of an inductor filled with a linear material with magnetization **M**, which can be written as

$$\mathbf{B} = \mu_0 \cdot (\mathbf{H} + \mathbf{M}), \tag{9}$$

where μ_0 is the free-space permeability. Moreover, the magnetic fields **B** and **H** can be associated with an energy density per unit volume w_m according to

$$w_\mathrm{m} = \frac{1}{2} \cdot \mathbf{B} \cdot \mathbf{H} = \frac{1}{2} \cdot \mu_0 \cdot (\mathbf{H} + \mathbf{M}) \cdot \mathbf{H} = \frac{1}{2} \cdot \mu_0 \cdot \left(\mathbf{H}^2 + \mathbf{M} \cdot \mathbf{H}\right). \tag{10}$$

Assuming that the **H**-field is produced by a current $i(t)$ running through an inductor with inductance L, the energy density of Eq. (10) can be related to an electrical power according to

$$P = \frac{\partial W_\mathrm{m}}{\partial t} = v(t) \cdot i(t) = \left[\frac{\mathrm{d}\,[\mathrm{L} \cdot \mathrm{i(t)}]}{\mathrm{d}t} + R_\mathrm{spin} \cdot i(t)\right] \cdot i(t), \tag{11}$$

where $W_\mathrm{m} = \int w_\mathrm{m} \mathrm{d}V$ is the total magnetic energy associated with the inductor, $v(t)$ is the voltage drop across the inductor, and R_spin models a potential increase in the loss associated with the inductor due to out-of-phase components of the magnetization **M** and the **H**-field.

Further assuming a sinusoidal waveform at a frequency ω_B1 for the current $i(t)$ according to $i(t) = \hat{I} \cdot \cos(\omega_\mathrm{B1} t)$, as it occurs in continuous-wave spin-detection experiments, and noting that of the product term $i(t)^2$ only the DC component needs to be considered,[5] Eq. (11) can be rewritten according to

$$P = \frac{\partial W_\mathrm{m}}{\partial t} = \frac{j}{2} \cdot \omega_\mathrm{B1} \cdot L \cdot \hat{I}^2 + \frac{1}{2} \cdot R_\mathrm{spin} \cdot \hat{I}^2. \tag{12}$$

A second expression for the power associated with the magnetic energy of the inductor can be derived by inserting the expression given by Eq. (10) for w_m into $P = \partial \int w_\mathrm{m} \mathrm{d}V / \partial t$, yielding

[5] The spectral component of $i(t)$ at $2\omega_\mathrm{B1}$ produces an H-field component at the same frequency that is far away from resonance and therefore does not need to be considered.

$$\begin{aligned} P &= \frac{1}{2} \cdot \int_V \left[\mu_0 \cdot \frac{\mathrm{d}}{\mathrm{d}t} \left[\mathbf{H}\left(t, \mathbf{r}\right) \right]^2 + \mu_0 \cdot \frac{\mathrm{d}}{\mathrm{d}t} \left[\mathbf{M}\left(t, \mathbf{r}\right) \cdot \mathbf{H}\left(t, \mathbf{r}\right) \right] \right] \mathrm{d}V \\ &= \mu_0 \cdot \int_V \left[\mathbf{H} \cdot \frac{\mathrm{d}\mathbf{H}}{\mathrm{d}t} + \frac{\mathbf{M}}{2} \cdot \frac{\mathrm{d}\mathbf{H}}{\mathrm{d}t} + \frac{\mathbf{H}}{2} \cdot \frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} \right] \mathrm{d}V, \end{aligned} \tag{13}$$

where in the last line, to avoid notational clutter, the arguments of all vector-valued functions have been dropped.

For the special case of sinusoidal time-dependencies of the **H**- and **M**-fields as they occur in continuous-wave spin-detection experiments, Eq. (13) can also be rewritten in phasor notation, resulting in

$$\begin{aligned} P &= j \cdot \omega_{\mathrm{B1}} \cdot \mu_0 \cdot \int_V \left[\frac{\bar{\mathbf{H}} \cdot \bar{\mathbf{H}}}{2} + \frac{\bar{\mathbf{M}} \cdot \bar{\mathbf{H}}}{4} + \frac{\bar{\mathbf{H}} \cdot \bar{\mathbf{M}}}{4} \right] \mathrm{d}V \\ &= \frac{j}{2} \cdot \omega_{\mathrm{B1}} \cdot \mu_0 \cdot \int_V \bar{\mathbf{H}}^2 \mathrm{d}V + \frac{j}{2} \cdot \omega_{\mathrm{B1}} \cdot \mu_0 \cdot \int_{V_\mathrm{s}} \bar{\mathbf{H}} \cdot \bar{\mathbf{M}}\, \mathrm{d}V, \end{aligned} \tag{14}$$

where $\bar{\mathbf{H}}$ and $\bar{\mathbf{M}}$ are the phasor representations of the applied magnetic field **H** and the sample magnetization **M**, and V_s is the sample volume, where $\mathbf{M} \neq \mathbf{0}$. At this point it should be mentioned that a factor of $1/2$ had to be introduced into Eq. (14) to take care of the fact that the phasors used in this chapter are amplitude phasors and not rms-phasors.

Replacing the generic phasors $\bar{\mathbf{H}}$ and $\bar{\mathbf{M}}$ by those occurring in a real continuous-wave spin-detection experiment, i.e., $\bar{\mathbf{H}} = H_{1\mathrm{x}}\mathbf{e}_\mathrm{x} + H_{1\mathrm{y}}\mathbf{e}_\mathrm{y}$[6] and $\bar{\mathbf{M}} = \bar{M}_{\mathrm{x,ss}}\, \mathbf{e}_\mathrm{x} + \bar{M}_{\mathrm{y,ss}}\, \mathbf{e}_\mathrm{y}$, Eq. (13) can be rewritten according to

$$P = \frac{j \cdot \omega_{\mathrm{B1}}}{2} \cdot \mu_0 \cdot \int_V \bar{\mathbf{H}}^2 \mathrm{d}V + \frac{\omega_{\mathrm{B1}}}{2} \cdot \int_{V_\mathrm{s}} \begin{pmatrix} H_{1\mathrm{x}} \\ H_{1\mathrm{y}} \end{pmatrix} \cdot \begin{pmatrix} j\, M_{\mathrm{x',ss}} - M_{\mathrm{y',ss}} \\ M_{\mathrm{x',ss}} + j\, M_{\mathrm{y',ss}} \end{pmatrix} \mathrm{d}V. \tag{15}$$

Further assuming that the H-field is produced by a current running through the inductor and introducing the unitary H-field of the inductor according to $\bar{\mathbf{H}}_\mathrm{u} = \bar{\mathbf{H}}/\hat{I} = H_{1\mathrm{xu}} \cdot \mathbf{e}_\mathrm{x} + H_{1\mathrm{yu}} \cdot \mathbf{e}_\mathrm{y}$, where $\hat{I}$ is the amplitude of the (sinusoidal) current running through the inductor, Eq. (15) can be rewritten according to

$$\begin{aligned} P = {} & \frac{j \cdot \omega_{\mathrm{B1}}}{2} \cdot \mu_0 \cdot \int_V \bar{\mathbf{H}}_\mathrm{u}^2 \mathrm{d}V \cdot \hat{I}^2 \\ & + \frac{\omega_{\mathrm{B1}}}{2} \cdot \int_{V_\mathrm{s}} \begin{pmatrix} H_{1\mathrm{xu}} \\ H_{1\mathrm{yu}} \end{pmatrix} \cdot \frac{1}{\hat{I}} \cdot \begin{pmatrix} j\, M_{\mathrm{x',ss}} - M_{\mathrm{y',ss}} \\ M_{\mathrm{x',ss}} + j\, M_{\mathrm{y',ss}} \end{pmatrix} \mathrm{d}V \cdot \hat{I}^2. \end{aligned} \tag{16}$$

[6] Note that in the phasor notation, only the component of **H** rotating at $\omega_{\mathrm{B1}} = \omega_L$, i.e., in resonance with the spin ensemble, has to be considered, cf. the effective B-field in the rotating frame of Eq. (3), and that therefore, the phasor amplitudes in the x- and y-directions are $H_{1\mathrm{x}}$ and $H_{1\mathrm{y}}$ and not $2\, H_{1\mathrm{x}}$ and $2\, H_{1\mathrm{y}}$.

Next, one can proceed by comparing Eqs. (12) and (16) and associating all terms of the right-hand side of Eq. (16) with inductances and resistances according to

$$L_0 = \mu_0 \int_V \bar{\mathbf{H}}_{\mathrm{u}}^2 \mathrm{d}V \tag{17a}$$

$$L_{\mathrm{spin}} = \int_{V_{\mathrm{s}}} \left[H_{1\mathrm{xu}} \cdot \frac{M_{\mathrm{x'},\mathrm{ss}}}{\hat{I}} + H_{1\mathrm{yu}} \cdot \frac{M_{\mathrm{y'},\mathrm{ss}}}{\hat{I}} \right] \mathrm{d}V \tag{17b}$$

$$R_{\mathrm{spin}} = -\omega_{\mathrm{B1}} \cdot \int_{V_{\mathrm{s}}} \left[H_{1\mathrm{xu}} \cdot \frac{M_{\mathrm{y'},\mathrm{ss}}}{\hat{I}} - H_{1\mathrm{yu}} \cdot \frac{M_{\mathrm{x'},\mathrm{ss}}}{\hat{I}} \right] \mathrm{d}V. \tag{17c}$$

Finally, substituting the results of Eq. (5) into Eq. (17) and using the approximation $\mathbf{H}_{\mathrm{u}} = (1/\mu_0) \cdot (B_{1\mathrm{xu}} \cdot \mathbf{e}_{\mathrm{x}} + B_{1\mathrm{yu}} \cdot \mathbf{e}_{\mathrm{y}})$,[7] one obtains

$$L_0 = \mu_0 \int_V \bar{\mathbf{H}}_{\mathrm{u}}^2 \mathrm{d}V = \frac{1}{\mu_0} \int_V \bar{\mathbf{B}}_{\mathrm{u}}^2 \mathrm{d}V \tag{18a}$$

$$L_{\mathrm{spin}} = \gamma T_2^2 \Delta\omega \cdot \int_{V_{\mathrm{s}}} \frac{M_0 \cdot \left(B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2\right)}{1 + \Delta\omega^2 T_2^2 + T_1 T_2 \left((\gamma B_{1x})^2 + \left(\gamma B_{1y}\right)^2\right)} \mathrm{d}V \tag{18b}$$

$$R_{\mathrm{spin}} = -\gamma T_2 \omega_{\mathrm{B1}} \cdot \int_{V_{\mathrm{s}}} \frac{M_0 \cdot \left(B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2\right)}{1 + \Delta\omega^2 T_2^2 + T_1 T_2 \left((\gamma B_{1x})^2 + \left(\gamma B_{1y}\right)^2\right)} \mathrm{d}V, \tag{18c}$$

where V_{s} is the sample volume, $\Delta\omega = \omega_{\mathrm{B1}} - \omega_L$ and $\omega_L = -\gamma B_0$ is the Larmor frequency, γ is the gyromagnetic ratio, M_0 is the steady-state sample magnetization, T_1 and T_2 are the longitudinal and transversal relaxation times, respectively, ω_{B1} is the frequency of the B_1-field produced by the inductor, and $B_{1\mathrm{xu}}$ and $B_{1\mathrm{yu}}$ are the unitary magnetic fields of the inductor in the x- and y-directions, respectively.

The results of Eq. (18) can be interpreted as follows: L_0 is the conventional inductance of the inductor when filled with air/vacuum,[8] and L_{spin} and R_{spin} are the changes in inductance and resistance of the (nonideal, i.e., lossy) inductor when filled with a sample that undergoes a change in its magnetization in response to a spin resonance experiment.

[7] According to this simplifying assumption, the B_1-field that interacts with the spin ensemble according to the Bloch equation can be calculated from the current running in the coil that produces the H_1-field by assuming a relative permeability of $\mu_{\mathrm{r}} = 1$, i.e., ignoring the small effect of the nonzero sample susceptibility. This assumption essentially linearizes the intrinsically nonlinear interaction between the B_1-field and the spin ensemble.

[8] Please note that it was assumed that the coil used for the spin-detection experiment has only B_{u}-field components in the x- and y-directions because those are the ones interacting with the spin ensemble. Should the coil also produce a field component in the z-direction, this component would need to be considered in computing L_0 but would not change the expression for L_{spin}.

Before further discussing the general case described by Eq. (18), it is instructive to first look at the nonsaturated case, which occurs for $1 \gg T_1\, T_2\, \left((\gamma\, B_{1x})^2 + \left(\gamma\, B_{1y}\right)^2 \right)$. Under these conditions, it is useful to introduce a complex susceptibility $\chi = \chi' - j\,\chi''$, which relates the complex H_1-field phasor produced by the inductor to the spin-related complex transversal magnetization phasor $\bar{M}_{\mathrm{x,ss}} + j\, \bar{M}_{\mathrm{y,ss}}$ according to

$$\bar{M}_{\mathrm{x,ss}} + j\, \bar{M}_{\mathrm{y,ss}} = \left(\chi' - j\,\chi''\right) \cdot \left(2\, H_{1\mathrm{x}} + j\, 2\, H_{1\mathrm{y}}\right) = \frac{2}{\mu_0} \cdot \left(\chi' - j\,\chi''\right) \cdot \left(B_{1\mathrm{x}} + j\, B_{1\mathrm{y}}\right), \tag{19}$$

where $\bar{M}_{\mathrm{x,ss}}$ and $\bar{M}_{\mathrm{y,ss}}$ are defined according to Eq. (6). Then, χ' and χ'' can be expressed as

$$\chi' = \frac{\Delta\omega\, T_2^2}{1 + \Delta\omega^2\, T_2^2} \cdot \gamma\, \mu_0\, M_0 \tag{20a}$$

$$\chi'' = -\frac{T_2}{1 + \Delta\omega^2\, T_2^2} \cdot \gamma\, \mu_0\, M_0, \tag{20b}$$

where $\Delta\omega = \omega_{\mathrm{B1}} - \omega_L$, T_2 is the transversal sample relaxation time, γ is the gyromagnetic ratio, μ_0 is the free-space permeability, and M_0 is the steady-state sample magnetization.

Here, it should be pointed out that the expressions for the linear susceptibility given in Eq. (20) differ by a factor of two from those defined in previous works; cf., e.g., [42]. This discrepancy arises from the fact that in the treatment above, the susceptibility relates the complex H_1-field, which accounts for both the H_1-fields in the x- and y-directions, to the complex steady-state magnetizations in those directions. In previous treatments, cf., e.g., [42], the susceptibility only related the H_1-field in the x-direction to the complex magnetization in the same direction, resulting in a complex susceptibility that is a factor of two smaller. In an oscillator-based spin detector, the spin-induced change in frequency will turn out to be approximately proportional to the spin-related inductance L_{spin} defined in Eq. (17b), cf. Sect. 5.2, which is proportional to the total magnetic energy associated with the spin ensemble. Therefore, since this total magnetic energy is in turn a function of the spin-related magnetizations in both the x- and y-directions, both components should be considered in the definition of the complex susceptibility to obtain a correct expression for the total spin-related inductance and the resulting total spin-induced frequency shift.

Keeping this in mind and assuming a nonsaturated, homogeneous[9] sample, the equations for L_{spin} and R_{spin} given in Eq. (18) can be further simplified by introducing the complex susceptibility of Eq. (20) according to

[9] Namely, $M_0 \neq M_0(\mathbf{r})$, that is, the sample magnetization is constant over the sample volume.

$$\begin{aligned} L_{\mathrm{spin}} &= \frac{\chi'}{\mu_0} \cdot \int_{V_{\mathrm{s}}} B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2 \mathrm{d}V \\ &= \chi' \cdot \frac{\int_{V_{\mathrm{s}}} B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2 \mathrm{d}V}{\int_{V} B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2 \mathrm{d}V} \cdot \frac{1}{\mu_0} \cdot \int_{V} B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2 \mathrm{d}V \\ &\approx \chi' \cdot \eta \cdot L_0 \end{aligned} \tag{21a}$$

$$\begin{aligned} R_{\mathrm{spin}} &= -\frac{\chi'' \omega_{\mathrm{B1}}}{\mu_0} \cdot \int_{V_{\mathrm{s}}} B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2 \, \mathrm{d}V \\ &\approx -\chi'' \, \eta \, \omega_{\mathrm{B1}} \cdot L_0, \end{aligned} \tag{21b}$$

where $\eta \approx V_{\mathrm{s}}/V_{\mathrm{coil}}$ is the so-called filling factor, V_{s} being the sample volume and V_{coil} the sensitive coil volume. The filling factor η indicates how much of the sensitive volume of the inductor is filled with the active material.

Since it will turn out to be important for the derivation of the limit of detection (LOD) of frequency-sensitive oscillator-based spin detectors in Sect. 5.2, we will next consider the more general case in which saturation effects cannot be ignored.[10] Under these conditions, it is impossible to define a linear susceptibility independent of $B_{1\mathrm{x}}$ and $B_{1\mathrm{y}}$ such as the one defined in Eq. (20). Instead, to capture the nonlinear relation between B_1 and M_{ss}, one can define a B_1-dependent susceptibility according to

$$\chi' = \frac{\Delta\omega \, T_2^2}{1 + \Delta\omega^2 \, T_2^2 + T_1 \, T_2 \, \left((\gamma \, B_{1x})^2 + \left(\gamma \, B_{1y}\right)^2 \right)} \cdot \gamma \, \mu_0 \, M_0 \tag{22a}$$

$$\chi'' = -\frac{T_2}{1 + \Delta\omega^2 \, T_2^2 + T_1 \, T_2 \, \left((\gamma \, B_{1x})^2 + \left(\gamma \, B_{1y}\right)^2 \right)} \cdot \gamma \, \mu_0 \, M_0, \tag{22b}$$

where $\Delta\omega = \omega_{\mathrm{B1}} - \omega_L$, T_1 and T_2 are the longitudinal and the transversal sample relaxation times, γ is the gyromagnetic ratio, μ_0 is the free-space permeability, M_0 is the steady-state sample magnetization, and $B_{1\mathrm{x,y}}$ are the x- and y-components of the B_1-field.

4 Phase and Frequency Noise in LC Tank Oscillators

Large parts of the following discussions and derivations are extracted from the extensive treatment of phase and frequency noise in LC tank oscillators in [8, Chap. 5] and [5].

[10] That is, for $1 \leq T_1 \, T_2 \, \left((\gamma \, B_{1x})^2 + \left(\gamma \, B_{1y}\right)^2 \right)$.

4.1 Preliminary Considerations

In the following discussion, we will develop analytical expressions for the oscillation frequency and amplitude as well as the frequency noise of an LC tank oscillator. From the modeling process, we will see that in order to describe the frequency uncertainty of an oscillator, it is not necessary to deal with actual phase noise, and that the most natural description for the frequency uncertainty of an oscillator is its frequency fluctuation with units of Hz^2/Hz or (angular) frequency fluctuation with units of (rad/s)^2/Hz. The frequency noise process has the advantage that in contrast to phase noise, it can be modeled by a (cyclo-) stationary process whose (averaged) variance is constant over time. While the frequency noise process is sufficient to find the limit of detection for the frequency-sensitive oscillator-based spin detector discussed in this chapter, standard measurement instruments such as signal source analyzers display their results as phase noise. Phase noise naturally has units of rad^2/Hz. Therefore, to be able to relate the measured results from signal source analyzers to the expressions for the frequency noise of LC tank oscillators derived in this chapter, a transformation between the two processes will be provided in Sect. 6.

4.2 Nonlinear Oscillator Modeling

An LC tank oscillator can be modeled mathematically as a second-order system of nonlinear ordinary differential equations according to $\dot{x} = f(x)$. One possible approach to studying its properties is the Andronov–Hopf bifurcation theory; cf. [17]. The Andronov–Hopf bifurcation theory allows one to determine the existence of a limit cycle as is required for an electrical oscillator by embedding the oscillator at hand into a family of oscillators $\dot{x} = f(x, \mu)$ with a family of equilibrium points $x_0(\mu)$, such that $f(x_0(\mu), \mu) = 0$, and analyzing the eigenvalues of the Jacobian $J_f(x, \mu)$ associated with f evaluated at $x = x_0$. The stability of the limit cycle can then be assessed from the sign of the first Lyapunov coefficient; cf. [17].

Instead of following this approach, due to the limited space available, here we will not go into the details of the Hopf bifurcation theory but ensure the existence of a stable limit cycle by resorting to the theory of weakly nonlinear oscillators, which is discussed, e.g., in [24], and rewriting our system as a perturbed harmonic oscillator.[11] More details about general oscillator modeling can be found in [17] and [24].

[11] The modeling equations resulting from the stochastic averaging method used in this chapter have the form of a weakly nonlinear oscillator.

4.3 Nonlinear Oscillator Modeling in the Presence of Noise

Phase noise is probably one of the most-discussed topics in circuit design, and there is a variety of models that have been published in the open literature ranging from very simple linear time-invariant models as proposed by Leeson [26] to more complicated linear time-varying models as proposed by Hajimiri [19] and nonlinear models with varying degrees of complexity [13, 20, 23, 25]. As was already discussed by Lax in [25], one fundamental problem with all of these models arises from the fact that an oscillator is a nonlinear system far away from thermal equilibrium, and for such systems even the most sophisticated methods proposed up to now can be seen only as heuristics, because they all rely on modeling based on stochastic differential equations (SDEs). An excellent review article covering this topic in more detail has been published by Mathis et al. [28]. In that article, the authors explain how the Langevin approach of introducing additive noise sources into the deterministic system description fails for nonlinear dynamical systems because it leads to physical inconsistencies. Here, the major problem that arises in going from a linear system, where the Langevin approach leads to physically consistent results, to a nonlinear system is the coupling between the moments of the stochastic process, which is described by the nonlinear system SDE. This results in a situation in which the stochastically averaged SDE is in general not identical to the deterministic system to which the noise sources have been added. Consequently, the self-consistency of the Langevin approach, which is given for linear systems, where the stochastically averaged SDE indeed corresponds to the deterministic systems to which the noise sources have been added, is no longer given for nonlinear systems.

4.4 A Model for the Phase Noise in LC Tank Oscillators Based on a Special Case of Bogoliubov's Asymptotic Method

In this section, we will proceed by tackling the problem of oscillator frequency noise using the method outlined in the book by Stratonovich [36]. This approach is based on the idea of introducing equivalent noise sources for the different noisy components in the oscillator and solving the resulting SDE using the so-called stochastic averaging method. Here, one should keep in mind that according to the previous section, this Langevin-type approach has to be seen as a heuristic. Nevertheless, as we shall discover when we compare the model with measured data, the heuristic approach can be empirically justified.

Bearing this intrinsic limitation in mind, one can proceed by investigating the model of a noisy LC tank oscillator, whose circuit diagram and equivalent model are shown in Fig. 3a and b. Here, it should be noted that the noise analysis based on Stratonovich's method for this circuit topology was first presented in [5]. The

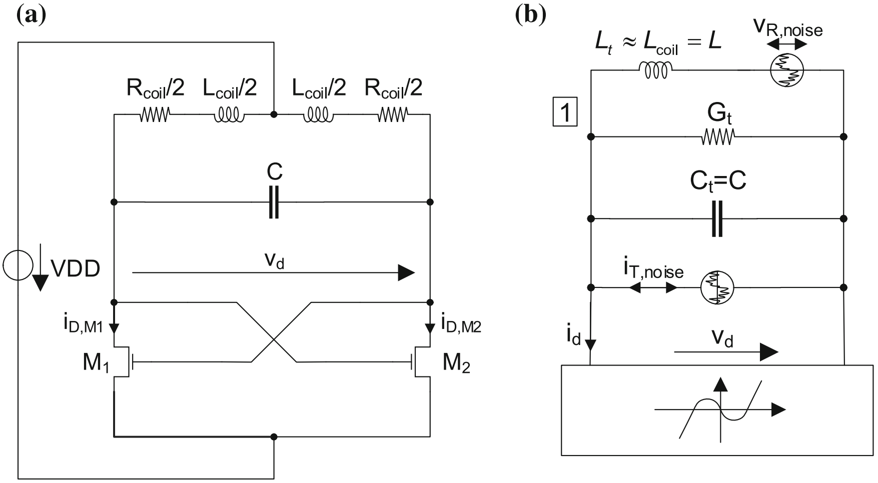


Fig. 3 **a** Schematic of a simple LC tank oscillator and **b** equivalent circuit including additive noise sources, in which the cross-coupled pair is modeled as a static nonlinearity and the LC tank is replaced by an equivalent parallel model

corresponding analyses for a current-starved LC tank oscillator with its active devices in strong and weak inversion, respectively, can also be found in the open literature in [3, 4].

From Fig. 3b it can be seen that in order to introduce noise into the model, two noise sources—one modeling the thermal noise of the coil resistance and a second one modeling the noise of the active cross-coupled transistor pair—have been added compared to the noise-free case. At this point, it is important to note that the choice of state variables, which is of minor importance in the deterministic case, in which a different choice of state variables would not alter the estimates of the oscillation amplitude and frequency, is of prime importance for obtaining even qualitatively accurate results for the frequency noise spectra of an LC tank oscillator. More specifically, depending on the state variables used, a single noise source shown in Fig. 3b can lead to a different slope in the calculated frequency noise spectrum. Although this ambiguity might sound confusing or even seem to be an indication of an inadequate model, it merely reflects the fact that modeling noise in nonlinear systems far away from equilibrium cannot be seen and treated as a straightforward extension of the equilibrium case. To solve this modeling problem, we used measured frequency noise spectra obtained using prototype realizations of the proposed oscillator-based spin detector to calibrate the model. The physical reasoning behind this calibration approach is that the ultimate detection limit of a frequency-sensitive spin detector is defined by the white part of the frequency noise spectrum. At the same time, it is reasonable to assume that this detection limit should result from the intrinsic white voltage and current noise sources present in the oscillation circuit. Therefore, a reasonable choice of state variables is guided by the goal of obtaining flat frequency noise spectra from both the white voltage noise source $v_{\mathrm{noise,R}}$ and the white part of the current noise source $i_{\mathrm{noise,T}}$. The application of this calibration approach leads to the rather unusual situation that in order to model the noise originating from the coil resistance, one has to use the differential tank voltage $x_R = v_d$ as state variable, while in order to obtain the right contribution to the overall noise stemming from the nonlinear cross-coupled transistor pair, one has to use the flux-type state variable $x_T = \omega_{\mathrm{LC}} \int v_d \mathrm{d}t$ instead.

With this choice of state variables, one can proceed by first calculating the resulting amplitude and frequency noise originating in the thermal noise produced by the resistance of the tank inductor, $v_{\mathrm{noise,R}}$. To this end, one has to set $i_{\mathrm{noise,T}}$ to zero and then write Kirchhoff's current law for node $\boxed{1}$ in Fig. 3b and Kirchhoff's voltage law for the loop $v_{\mathrm{d}} - v_{\mathrm{noise,R}} = L \cdot \mathrm{d}i_{\mathrm{L}}/\mathrm{d}t$. Then, introducing the I-V characteristic of the cross-coupled transistor pair, $i_{\mathrm{d}} = -(G_{\mathrm{m0}}/2)\, v_{\mathrm{d}} + (G_{\mathrm{m0}}/2)/(4\, V_{\mathrm{DD}}^2) v_{\mathrm{d}}^3$, and setting $x_R \triangleq v_{\mathrm{d}}$, one obtains the following second-order differential equation describing the oscillator circuit:

$$\ddot{x}_R + \omega_{\mathrm{LC}}^2\, x_R = \varepsilon \left[\frac{1}{C} \left(4\, V_{\mathrm{DD}}^2 \left(1 - \frac{1}{\alpha_{\mathrm{od}}} \right) - 3\, x_R^2 \right) \dot{x}_R + \frac{\omega_{\mathrm{LC}}^2}{\varepsilon} \cdot v_{\mathrm{noise,R}} \right], \qquad (23)$$

where $\alpha_{\mathrm{od}} = (G_{\mathrm{m0}}/2)/G_t$, $\varepsilon = \alpha_{\mathrm{od}}/(4\, V_{\mathrm{DD}}^2)$ and $\omega_{\mathrm{LC}} = 1/\sqrt{LC}$. This equation can be further simplified by introducing a first-order estimate of the oscillation amplitude according to $A_0 = 4/\sqrt{3}\,\sqrt{1 - 1/\alpha_{\mathrm{od}}} \cdot V_{\mathrm{DD}}$, cf. [5], and the normalized noise process $\tilde{v}_{\mathrm{noise,R}} \triangleq v_{\mathrm{noise,R}}/\varepsilon$, resulting in

$$\ddot{x}_R + \omega_{\mathrm{LC}}^2\, x_R = \varepsilon \left[\frac{3}{C} \left(\frac{A_0^2}{4} - x_R^2 \right) \dot{x}_R + \omega_{\mathrm{LC}}^2 \cdot \tilde{v}_{\mathrm{noise,R}} \right] = \varepsilon\, f_R(x_R, \dot{x}_R, t), \qquad (24)$$

where the normalized noise process $\tilde{v}_{\mathrm{noise,R}}$ has been introduced in order to ensure that for $\varepsilon = 0$, the system formally corresponds to the harmonic oscillator. As discussed in [36], there are other possible choices of introducing normalized noise processes that in most cases all lead to the same result. For example, A_0 can be calculated by applying the method of averaging, cf. [24], to the deterministic system, i.e., the system in which both noise sources in Fig. 3b are set to zero.

Similarly, one can derive the state equation for $v_{\mathrm{noise,R}} = 0$ and a nonzero $i_{\mathrm{noise,T}}$ according to

$$\ddot{x}_T + \omega_{\mathrm{LC}}^2\, x_T = \varepsilon \left[\frac{1}{C} \left(\frac{3}{4} A_0^2 - \left(\frac{\dot{x}_T}{\omega_{\mathrm{LC}}} \right)^2 \right) \dot{x}_T - \frac{\omega_{LC}}{C}\, \tilde{i}_{\mathrm{noise,T}} \right] = \varepsilon\, f_T(x_T, \dot{x}_T, t), \qquad (25)$$

where $\tilde{i}_{\mathrm{noise,T}} \triangleq i_{\mathrm{noise,T}}/\varepsilon$.

At this point, it is not obvious at all how in a nonlinear system the approach of treating the two noise sources independently—i.e., essentially assuming the superposition principle to apply—and, moreover, using two different sets of state variables, can lead to a consistent result. The solution to this problem will be to apply the so-called linearization method, cf. [36], where the two nonlinear SDEs given by Eqs. (24) and (25) are linearized around the trajectories of a corresponding deterministic system, the so-called smoothly averaged system. Fortunately, this approach will lead to equations for both the amplitude and phase noise in the two different systems under consideration, which are linear in $\tilde{v}_{\text{noise,R}}$ and $\tilde{i}_{\text{noise,T}}$, respectively, and whose deterministic parts are identical, providing an a posteriori justification of the utilized superposition approach.

Due to the limited space available in this chapter, it is impossible to discuss in detail here the special case of Bogoliubov's method, which was used to derive the SDEs governing the behavior of the amplitude and phase fluctuations of the LC tank oscillator. Instead, we refer the interested reader to the book by Stratonovich [36] for a complete discussion of the required transformation steps.

For the purpose of this chapter, it should suffice to say that the core of this method is a transformation of the noisy state equations describing the system behavior in the presence of noise into polar coordinates, resulting in the **equations in standard form**; cf. [36]. Then the equations in standard form are further transformed to obtain two systems: a first one describing the fast fluctuations of the circuit and a second system describing the slowly varying components, the so-called smoothly averaged system.

Applying this method to the system at hand and noting that the deterministic parts of the two systems describing the amplitude and phase fluctuations, δA and $\delta\varphi$, in response to the voltage as well as the current noise source are identical, one obtains the following linear system of equations describing the amplitude and phase fluctuations of the LC tank oscillator in the presence of a lossy coil and a nonideal, noisy cross-coupled transistor pair:

$$\begin{aligned}\delta\dot{A} &= -\frac{3\,G_{\text{m0}}}{64\,C\,V_{\text{DD}}^2}\left(A_0^2 - 3\,A_{\text{sm}}^2\right)\delta A \\ &\quad - \left[\omega_{\text{LC}}\,v_{\text{noise,R}} - \frac{1}{C}\,i_{\text{noise,T}}\right]\sin\left(\omega_{\text{LC}}\,t + \varphi_{\text{sm}}\right) \qquad \text{(26a)}\\ \delta\dot{\varphi} &= \frac{9\,G_{\text{m0}}^2\,A_{\text{sm}}^2}{4096\,A_0\,C^2\,V_{\text{DD}}^4\,\omega_{\text{LC}}}\left(3\,A_0^2 - 5\,A_{\text{sm}}^2\right)\delta A \\ &\quad - \left[\frac{\omega_{\text{LC}}}{A_0}v_{\text{noise,R}} + \frac{1}{A_0\,C}i_{\text{noise,T}}\right]\cos\left(\omega_{\text{LC}}\,t + \varphi_{\text{sm}}\right), \qquad \text{(26b)}\end{aligned}$$

where the transformed noise processes $\tilde{v}_{\text{noise,R}}$ and $\tilde{i}_{\text{noise,T}}$ have been again replaced by the original noise processes $v_{\text{noise,R}} = \varepsilon\,\tilde{v}_{\text{noise,R}}$ and $i_{\text{noise,T}} = \varepsilon\,\tilde{i}_{\text{noise,T}}$. The smoothly averaged trajectories can be calculated as the solution of the smoothly averaged system

$$\dot{A}_{\text{sm}} = \frac{3\,A_{\text{sm}}\,A_0^2}{8\,C}\left[1 - \left(\frac{A_{\text{sm}}}{A_0}\right)^2\right] \tag{27a}$$

$$\dot{\varphi}_{\text{sm}} = -\frac{9\varepsilon}{256\,\omega_{\text{LC}}\,C^2}\left(27A_{\text{sm}}^4 - 14A_{\text{sm}}^2\,A_0^2 + 2A_0^4\right). \tag{27b}$$

The steady-state solution of Eq. (27a) for $\dot{A}_{\text{sm}} = 0$ directly provides the desired smoothly averaged amplitude, A_{sm}, according to

$$A_{\text{sm}} = A_0 = \frac{4}{\sqrt{3}}\sqrt{1 - \frac{1}{\alpha_{\text{od}}}} \cdot V_{\text{DD}}. \tag{28}$$

Then, substituting this finding into Eq. (27b), one obtains for the smoothly averaged phase

$$\varphi_{\text{sm}} = -\frac{(\alpha_{\text{od}} - 1)^2}{16 \cdot Q_{\text{coil}}^2}\,t + \varphi_0, \tag{29}$$

where φ_0 is an arbitrary integration constant, which can be modeled as a random variable uniformly distributed in the interval $[0,\ 2\,\pi]$.

The smoothly averaged amplitude and phase of Eqs. (28) and (29) can then be inserted into Eq. (26) to arrive at the SDEs describing the fluctuational deviations of the oscillation amplitude and phase δA and $\delta\varphi$:

$$\delta\dot{A} = -\frac{3}{32}\frac{A_0^2\,G_{\text{m0}}}{C\,V_{\text{DD}}^2}\delta A - \left[\omega_{\text{LC}}\,v_{\text{noise,R}} - \frac{i_{\text{noise,T}}}{C}\right]\sin\left(\omega_{\text{osc}}\,t + \varphi_0\right) \tag{30a}$$

$$\delta\dot{\varphi} = -\frac{9}{2048}\frac{A_0^3\,G_{\text{m0}}^2\,\delta A}{C^2\,V_{\text{DD}}^4\,\omega_{\text{LC}}} - \left[\frac{\omega_{\text{LC}}}{A_0}v_{\text{noise,R}} + \frac{i_{\text{noise,T}}}{A_0\,C}\right]\cos\left(\omega_{\text{osc}}\,t + \varphi_0\right), \tag{30b}$$

where

$$\omega_{\text{osc}} = \omega_{\text{LC}}\left(1 - \frac{(\alpha_{\text{od}} - 1)^2}{16 \cdot Q_{\text{coil}}^2}\right). \tag{31}$$

Before proceeding, it is instructive to take a closer look at the structure of Eq. (30): Eq. (30) presents a system of (weakly) coupled nonautonomous SDEs with a linear homogeneous part and that is also linear in both driving processes $v_{\text{noise,R}}$ and $i_{\text{noise,T}}$. Therefore, since both $v_{\text{noise,R}}$ and $i_{\text{noise,T}}$ can be modeled as Gaussian random processes, δA and $\delta\varphi$ will inherit this property, and the second-order moments will be sufficient to characterize them. Moreover, since both driving processes have zero mean, both δA and $\delta\varphi$ will also inherit this property, and it is sufficient to calculate the autocorrelation functions of δA and $\delta\varphi$ in order to completely characterize them.

The situation is even further simplified by the relatively weak coupling between Eqs. (30a) and (30b) and the fact that for frequency-sensitive spin resonance experiments one is interested only in the frequency noise produced by the LC tank oscillator. Then, noting that the amplitude noise of oscillators is orders of magnitude smaller than their phase noise and that the scaling coefficient of δA in Eq. (30b) is also small in magnitude, the error introduced in ignoring the influence of the amplitude noise δA on the phase noise $\delta\varphi$ is small. Finally, one can use the relation between the instantaneous phase φ_i and instantaneous frequency ω_i of a signal given by $\omega_i = \dot{\varphi}_i$ to derive a simplified equation for the frequency noise of an LC tank oscillator from Eq. (30) according to

$$\delta\omega \approx -\frac{\omega_{\mathrm{LC}}}{A_0} v_{\mathrm{noise,R}} \cos\left(\omega_{\mathrm{osc}}\, t + \varphi_0\right) - \frac{1}{A_0\, C} i_{\mathrm{noise,T}} \cos\left(\omega_{\mathrm{osc}}\, t + \varphi_0\right). \tag{32}$$

4.5 Frequency Noise Due to the Coil Resistance

Using the linearity of Eq. (32) with respect to $v_{\mathrm{noise,R}}$ and $i_{\mathrm{noise,T}}$, the solution of Eq. (32) can be derived using the superposition principle, which allows one to independently derive the fraction of the (angular) frequency noise process $\delta\omega$ that is caused by the thermal noise associated with the coil resistance $v_{\mathrm{noise,R}}$ and that produced by $i_{\mathrm{noise,T}}$. The total frequency noise is then found by adding the two partial noise power spectral densities (PSDs) to obtain the total noise PSD. To this end, one can start by modeling the process $v_{\mathrm{noise,R}}$ as a scaled white noise process $\xi(t)$ with a (double-sided) PSD of $2kTR_{\mathrm{coil}}$, where k is Boltzmann's constant and T is absolute temperature. Then, inserting this model of $v_{\mathrm{noise,R}}$ into Eq. (32) and setting $i_{\mathrm{noise,T}} = 0$, one obtains the following (partial) autocorrelation $R_{\delta\omega,\delta\omega,R}(\tau)$ of the noise process $\delta\omega$ produced by $v_{\mathrm{noise,R}}$ alone:

$$\begin{aligned} R_{\delta\omega,\delta\omega,R}(\tau) &= 2kTR_{\mathrm{coil}} \left(\frac{\omega_{\mathrm{LC}}}{A_0}\right)^2 \left\langle \cos\left(\omega_{\mathrm{osc}} t + \varphi_0\right) \cos\left(\omega_{\mathrm{osc}}\left(t+\tau\right) + \varphi_0\right)\right\rangle \\ &\quad \left\langle \xi(t)\xi(t+\tau)\right\rangle \\ &= kTR_{\mathrm{coil}} \left(\frac{\omega_{\mathrm{LC}}}{A_0}\right)^2 \cos(\omega_{\mathrm{osc}}\tau)\delta(\tau) \\ &= kTR_{\mathrm{coil}} \left(\frac{\omega_{\mathrm{LC}}}{A_0}\right)^2 \delta(\tau), \end{aligned} \tag{33}$$

where $\delta(\tau)$ is the Dirac delta function and the index R denotes the fact that this is the contribution of the coil resistance R_{coil} to the overall autocorrelation function. In deriving this result, the assumption that φ_0 is a random variable uniformly distributed in the interval $[0, 2\,\pi]$, which is in addition statistically independent of the random process $\xi(t)$ at all times t was used. Since $\delta\omega$ is a zero-mean process and $R_{\delta\omega,\delta\omega,R}(\tau)$ depends only on the time shift τ and not on the absolute time epoch t, $\delta\omega$ is clearly a wide-sense stationary (WSS) process, cf. [31], whose PSD can be computed using the Wiener–Khinchin theorem, cf. [31], according to

$$\boxed{S_{\delta\omega,\delta\omega,R}(\omega) = kT R_{\text{coil}} \left(\frac{\omega_{\text{LC}}}{A_0}\right)^2}\,, \tag{34}$$

where again, the index R indicates that $S_{\delta\omega,\delta\omega,R}(\omega)$ is the contribution to the overall PSD from the coil resistance alone. From Eq. (34) one sees that the coil resistance indeed establishes a constant noise floor for the frequency noise, as it was required from the model.

4.6 Frequency Noise Due to the Cross-Coupled Transistor Pair

To complete the frequency noise analysis of the LC tank oscillator, it remains to compute the contribution from the cross-coupled transistor pair. Although the analysis is in principle very similar to that of the previous section, it is complicated by the fact that the current noise produced by the cross-coupled transistor pair is bias-dependent and has a PSD that displays both a flat region, corresponding to a scaled white noise process, and a so-called 1/f-noise region around DC where the PSD falls off with a slope of 10 dB per decade. To account for these facts, one can proceed by first calculating the contribution of the white noise part of the transistor noise and subsequently including the effect of the 1/f-part of the transistor noise spectrum.

In order to calculate the effect of the white noise part of the transistor noise spectrum on the frequency noise PSD, one first has to relate the parallel current noise source of Fig. 3b to the physical noise sources given by the drain noise of the two transistors in the cross-coupled pair.

A very suitable model for the noise produced by a single MOSFET can be found in [14, Chap. 6]. According to this model, the white noise part of the (double-sided) PSD of the drain current noise of a single transistor in saturation is given by

$$S_{i_{\text{nD}},i_{\text{nD}}}(\omega) = 2kT\,\gamma_{\text{nD}}\,G_{\text{m}}, \tag{35}$$

where γ_{nD} is the gate excess noise factor, which takes on values of $\approx 2/3$ and ≈ 1 in weak inversion and strong inversion, respectively, and G_{m} is the (gate) transconductance of the MOSFET; cf. [14]. Then, noting that the differential tank current i_{d} shown in Fig. 3b can be expressed in terms of the drain currents of the two transistors according to $i_{\text{d}} = (i_{\text{D,M1}} - i_{\text{D,M2}})/2$, one obtains for the noise current of the cross-coupled transistor pair

$$i_{\text{nd}} = \frac{i_{\text{nD,M1}} - i_{\text{nD,M2}}}{2}, \tag{36}$$

where $i_{\text{nD,M1}}$ and $i_{\text{nD,M2}}$ are the drain noise currents of transistors M1 and M2, respectively. Since the two noise sources $i_{\text{nD,M1}}$ and $i_{\text{nD,M2}}$ are associated with physically different devices, they are statistically independent and the autocorrelation $R_{i_{\text{nd}},i_{\text{nd}}}(t,\tau)$ of the noise process i_{nd} can be written as

$$R_{i_{\text{nd}},i_{\text{nd}}}(t,\tau)=\frac{1}{4}\left[R_{i_{\text{nD,M1}},i_{\text{nD,M1}}}(t,\tau)+R_{i_{\text{nD,M2}},i_{\text{nD,M2}}}(t,\tau)\right]. \tag{37}$$

At this point, it is important to note that the gate transconductances of M1 and M2, $G_{\text{m1,2}}$, are functions of the oscillation voltage v_d. This in turn introduces a bias dependence in the noise current PSD of each MOSFET in the cross-coupled pair according to Eq. (35).

Fortunately, due to the periodic nature of v_d, the resulting frequency noise process is cyclostationary. Therefore, in order to still be able to calculate the corresponding spectrum by means of the Wiener–Khinchin theorem, a WSS process can be derived from the original cyclostationary process by averaging over one period of the oscillation signal.

To derive a closed-form expression for the frequency noise produced by the noise in the cross-coupled pair, one can assume that both transistors are in strong inversion and saturation for the entire oscillation period. Under these conditions, their gate transconductance can be written as

$$G_{\text{m0}}=\beta\cdot(V_{\text{P}}-V_{\text{S}})=\beta\cdot V_{\text{P}}=\beta\cdot\frac{V_{\text{G}}-V_{\text{T0}}}{n}, \tag{38}$$

where $V_{\text{S}}=0$ is the source potential of the transistor, V_{P} is the pinch-off voltage, V_{G} is the gate potential, V_{T0} is the threshold voltage, $\beta=\mu_n C_{\text{ox}}W/L$ is the transfer parameter of the MOSFET, and n is the slope factor; cf. [14].

Due to the periodic oscillation of v_d with an amplitude of A_0, the time-dependence of the gate voltages of the transistors M1 and M2 can be written as $V_{\text{G1,2}}=V_{\text{DD}}\pm A_0/2\cos(\omega_{\text{osc}}t)$. Then, the resulting time-averaged autocorrelation of the transistor drain currents can be written as

$$\begin{aligned}\bar{R}_{i_{\text{nD,M1,2}},i_{\text{nD,M1,2}}}(\tau)&=2kT\gamma_{\text{nD}}\,\beta\,\overline{V_{\text{P1,2}}}\,\delta(\tau)\cos(\omega_{\text{osc}}\tau)\\&=2kT\,\gamma_{\text{nD}}\,\beta\,\overline{V_{\text{P1,2}}}\,\delta(\tau),\end{aligned} \tag{39}$$

with

$$\overline{V_{\text{P1,2}}}\triangleq\frac{1}{T_{\text{osc}}}\int_0^{T_{\text{osc}}}V_{\text{P1,2}}(t)\text{d}t=\frac{V_{\text{DD}}-V_{\text{T0}}}{n}, \tag{40}$$

where $T_{\text{osc}}=2\pi/\omega_{\text{osc}}$ is the oscillation period.

Substituting the result of Eq. (40) into Eq. (39) and using the resulting expression to compute the time-averaged autocorrelation of i_{nd} by averaging Eq. (37) over T_{osc}, one obtains

$$\bar{R}_{i_{\text{nd}},i_{\text{nd}}}(\tau)=kT\,\gamma_{\text{nD}}\,G_{\text{m0}}\delta(\tau), \tag{41}$$

where $G_{\mathrm{m0}} = \beta\,(V_{\mathrm{DD}} - V_{\mathrm{T0}})/n$ is the gate transconductance of a single MOSFET in the cross-coupled pair for $v_d = 0$.

Finally, the result of Eq. (41) can be used to compute the time-averaged autocorrelation of the frequency noise originating in the cross-coupled transistor pair $\bar{R}_{\delta\omega,\delta\omega,T}(\tau)$ from Eq. (32) according to

$$\bar{R}_{\delta\omega,\delta\omega,T}(\tau) = \frac{1}{2} kT \frac{\gamma_{\mathrm{nD}}\, G_{\mathrm{m0}}}{A_0^2\, C^2}\,\delta(\tau), \tag{42}$$

where it was again assumed that φ_0 is uniformly distributed in the interval $[0, 2\,\pi]$.

Equation (42) can be brought into a format similar to that of Eq. (34) by noting that $G_{\mathrm{m0}}/2 = G_t \cdot \alpha_{\mathrm{od}}$ and $G_t \approx 1/(Q_{\mathrm{coil}}^2 R_{\mathrm{coil}}) \approx R_{\mathrm{coil}} C/L$. Then, substituting these approximations into Eq. (42), the PSD of the frequency noise due to the white noise in the cross-coupled transistor can be written as

$$\boxed{S_{\delta\omega,\delta\omega,T}(\omega) = kT\,\alpha_{\mathrm{od}}\gamma_{\mathrm{nD}} \left(\frac{\omega_{\mathrm{LC}}}{A_0}\right)^2 R_{\mathrm{coil}},} \tag{43}$$

where $\alpha_{\mathrm{od}} \triangleq (G_{\mathrm{m0}}/2)/G_t$ is the so-called overdrive parameter, which in order to ensure startup of the oscillator has to be chosen larger than one.

4.7 Total Frequency Noise of the LC Tank Oscillator

Combining the results of Eqs. (34) and (43) yields that the total PSD of the frequency noise can be obtained as the sum of the partial contributions from the coil resistance and the active transistor pair according to

$$\boxed{S_{\delta\omega,\delta\omega,\mathrm{tot,white}}(\omega) = kT \left(\frac{\omega_{\mathrm{LC}}}{A_0}\right)^2 (1 + \alpha_{\mathrm{od}}\,\gamma_{\mathrm{nD}})\, R_{\mathrm{coil}},} \tag{44}$$

where again, $\alpha_{\mathrm{od}} \triangleq (G_{\mathrm{m0}}/2)/G_t$ is the overdrive parameter introduced previously.

The frequency noise resulting from the 1/f-noise in the cross-coupled pair can be calculated using a procedure very similar to the one used for the white noise part, the main difference being the nonzero correlation time of the 1/f-noise process. Due to the limited space available in this chapter, a detailed derivation will be omitted here, and the interested reader is referred to [8]. From [8], the final result for the PSD of the frequency noise process originating from the 1/f-noise in the transistors of the cross-coupled pair is given by

$$\boxed{S_{\delta\omega,\delta\omega,T,1/f}(\omega) = \frac{\pi}{32}\frac{\alpha_{\mathrm{od}}\,\mu_n\,\alpha_{\mathrm{H}}\,q\,\omega_{\mathrm{LC}}^2}{L_{\mathrm{T}}^2\,n}R_{\mathrm{coil}}\left[3+8\left(\frac{V_{\mathrm{DD}}-V_{\mathrm{T0}}}{A_0}\right)^2\right]\frac{1}{|\omega|},} \tag{45}$$

where $\alpha_{\mathrm{od}} = (G_{\mathrm{m0}}/2)/G_t$, $\omega_{\mathrm{LC}} = 1/\sqrt{L\,C}$, L and C being the tank inductance and capacitance, respectively, μ_n is the mobility of the carriers in the transistors of the cross-coupled pair, α_{H} is the Hooge parameter, a unitless parameter ranging from about 10^{-4}–10^{-6}, q is the elementary charge, k is Boltzmann's constant, T is absolute temperature, n is the slope factor, L_{T} is the length of the transistors in the cross-coupled pair, and C_{ox} is their oxide capacitance; cf. [14].

5 Signal Calculation and Limit of Detection of Frequency-Sensitive Oscillator-Based Spin Detectors

5.1 Numerical Model for the Signal Calculation

In order to derive the LOD of the frequency-sensitive oscillator-based detection method described in this chapter, it is first necessary to derive an expression for the spin-induced change in oscillation frequency based on the modeling of the spin ensemble in Sect. 3 and the modeling of the oscillation frequency and amplitude provided in Sect. 4.2. To this end, we will first proceed by deriving an accurate nonlinear signal model suitable for numerical simulations. As a starting point, one can note that both the coil inductance and the coil resistance are functions of the resonant spin magnetization during a spin resonance experiment according to Eq. (18). Since furthermore, Eqs. (31) and (28) suggest that both the frequency and the amplitude of oscillation are functions of said inductance and resistance, it becomes obvious that also these two will be affected by the spin resonance effect. In this section, we will focus on the change of the oscillation frequency for which experimental results have already been published in the open literature; cf., e.g., [1, 42]. To find the desired change in oscillation frequency, one can use Eq. (31) as a starting point and replace the tank inductance L in the equation by the total inductance including the spin-induced change in inductance L_{spin} according to $L = L_0 + L_{\mathrm{spin}} = L_{\mathrm{tot}}$, where L_0 and L_{spin} are defined in Eq. (18). In this way, Eq. (31) can be rewritten according to

$$\omega_{\mathrm{osc,spin}} \approx \frac{1}{\sqrt{\left(L_0+L_{\mathrm{spin}}\right)C}}\cdot\left(1-\frac{\left(\alpha_{\mathrm{od,spin}}-1\right)^2}{16\cdot Q_{\mathrm{coil,spin}}^2}\right), \tag{46}$$

where the additional subscript "spin" denotes that the quantity is affected by the spin resonance, $Q_{\mathrm{coil,spin}} = \omega_{\mathrm{osc,spin}}\cdot L_{\mathrm{tot}}/R_{\mathrm{coil,spin}}$, and $\alpha_{\mathrm{od,spin}} = G_{\mathrm{m0}}/(2\cdot G_{\mathrm{t,spin}})$, where $G_{\mathrm{t,spin}} \approx (R_{\mathrm{coil,tot}}\cdot Q_{\mathrm{coi,spin}}^2)^{-1} \approx R_{\mathrm{coil,tot}}\cdot\omega_{\mathrm{LC}}\cdot L_0$ with $\omega_{\mathrm{LC}} = 1/\sqrt{L_0\cdot C}$

and $R_{\text{coil,tot}} = R_{\text{coil}} + R_{\text{spin}}$, R_{coil} being the coil resistance in the absence of spin resonance, and L_{spin} and R_{spin} are given by

$$L_{\text{spin}} = \gamma\, T_2^2 \Delta\omega \cdot \int_{V_s} \frac{M_0 \cdot \left(B_{1\text{xu}}^2 + B_{1\text{yu}}^2\right)}{1 + \Delta\omega^2\, T_2^2 + T_1\, T_2\, \left((\gamma\, B_{1x})^2 + \left(\gamma\, B_{1y}\right)^2\right)} \mathrm{d}V \tag{47a}$$

$$R_{\text{spin}} = -\gamma\, T_2\, \omega_{\text{B1}} \cdot \int_{V_s} \frac{M_0 \cdot \left(B_{1\text{xu}}^2 + B_{1\text{yu}}^2\right)}{1 + \Delta\omega^2\, T_2^2 + T_1\, T_2\, \left((\gamma\, B_{1x})^2 + \left(\gamma\, B_{1y}\right)^2\right)} \,\mathrm{d}V. \tag{47b}$$

Since L_{spin} and R_{spin} are functions of the (thereby itself spin-dependent) B_1-field[12] produced by the coil, we next have to relate said B_1-field to the oscillation amplitude of Eq. (28). This can be done by noting that for sufficiently large coil quality factors, the amplitude of the coil current can be approximated according to

$$\hat{I}_{\text{coil,spin}} \approx \frac{\hat{v}_{\text{d,spin}}}{\omega_{\text{osc,spin}} \cdot L_{\text{tot}}} \approx \frac{A_{0,\text{spin}}}{\omega_{\text{osc,spin}} \cdot L_{\text{tot}}}, \tag{48}$$

where $\hat{v}_{\text{d,spin}}$ is the amplitude of the differential tank voltage in the presence of the spin resonance and

$$A_{0,\text{spin}} = \frac{4}{\sqrt{3}} \sqrt{1 - \frac{1}{\alpha_{\text{od,spin}}}} \cdot V_{\text{DD}}. \tag{49}$$

Here, the spin-dependence of the amplitude arises from the dependence of α_{od} on the equivalent tank conductance G_t, which is a function of both Q_{coil} and R_{spin}, as expressed above.

Next, one can relate the current running through the tank inductance to the B_1-field assuming a simple loop coil whose axis is pointing in the x-direction[13] according to

$$\hat{B}_{1\text{x,spin}} \approx \frac{1}{2} \cdot \mu_0 \cdot \frac{\hat{I}_{\text{coil,spin}}}{d_{\text{coil}}} \approx \frac{1}{2} \cdot \mu_0 \cdot \frac{A_{0,\text{spin}}}{\omega_{\text{osc,spin}} \cdot L_{\text{tot}} \cdot d_{\text{coil}}}, \tag{50}$$

where the factor of $1/2$ accounts for the fact that only half of the current running through the coil produces a circularly polarized field rotating in the appropriate direction to excite the spin ensemble.

[12] The spin-dependence of the B_1-fields is not denoted by an additional subscript "spin" in Eq. (47) to avoid notational clutter.

[13] Here, if a better model accuracy is required, one could, e.g., use a Biot–Savart solver to relate the coil current to a position-dependent vector-valued B_1-field $\mathbf{B}_1(\mathbf{r})$.

In principle, the spin-induced frequency change can now be computed by inserting the expression for $B_{1\text{x,spin}}$ of Eq. (50) into Eq. (47) to obtain expressions for L_{spin} and R_{spin}, which can then be used to calculate $\alpha_{\text{od,spin}}$ and $Q_{\text{coil,spin}}$ before one can finally calculate the total spin-induced frequency shift from Eq. (46). However, the situation is further complicated by the fact that both L_{spin} and R_{spin} are functions of $\Delta\omega = \omega_{\text{B1}} - \omega_L = \omega_{\text{B1}} + \gamma B_0$. Since ω_{B1} is the frequency of the B_1-field produced by the coil, which is in the case of oscillator-based detection equal to the oscillation frequency, i.e., $\omega_{\text{B1}} = \omega_{\text{osc,spin}}$, Eq. (46) is in fact a fixed-point equation. To make this point even more obvious, one can write

$$B_{1,\text{spin}} = f_1\left(\omega_{\text{osc,spin}}, B_{1,\text{spin}}\right) \tag{51}$$

$$\omega_{\text{osc,spin}} = f_2\left(\omega_{\text{osc,spin}}, B_{1,\text{spin}}\right), \tag{52}$$

which in vector form reads

$$\mathbf{x} \triangleq \begin{pmatrix} B_{1,\text{spin}} \\ \omega_{\text{osc,spin}} \end{pmatrix} = \begin{pmatrix} f_1(\omega_{\text{osc,spin}}, B_{1,\text{spin}}) \\ f_2(\omega_{\text{osc,spin}}, B_{1,\text{spin}}) \end{pmatrix} = \mathbf{f}(\mathbf{x}). \tag{53}$$

This vector fixed-point equation can be solved numerically using the intrinsic values, i.e., the values in the absence of the shift due to spin resonance, as initial conditions, according to

$$\mathbf{x}_0 = \begin{pmatrix} B_{1,0} \\ \omega_{\text{osc},0} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \cdot \mu_0 \cdot \frac{A_0}{\omega_{\text{LC}}\, L_0\, d_{\text{coil}}} \\ \omega_{\text{LC}} \left(1 - \frac{(\alpha_{\text{od}}-1)^2}{16\, Q_{\text{coil}}^2}\right) \end{pmatrix}, \tag{54}$$

where $\omega_{\text{LC}} = 1/\sqrt{L_0\, C}$ and A_0 is the oscillator amplitude given by Eq. (28).

5.2 Limit of Detection

While the signal model derived in the previous section is very useful for numerical signal calculations, it needs to be significantly simplified in order to derive closed-form expressions for the achievable LOD for frequency-sensitive oscillator-based spin detectors, which is the topic of this section. To achieve this goal, we can again start from Eq. (31) but this time ignore the relatively weak—compared to the dependence of L_{tot}—spin-dependence of both α_{od} and Q_{coil}, yielding

$$\omega_{\text{osc,spin}} \approx \frac{1}{\sqrt{\left(L_0 + L_{\text{spin}}\right) C}} \cdot \left(1 - \frac{(\alpha_{\text{od}} - 1)^2}{16 \cdot Q_{\text{coil}}^2}\right). \tag{55}$$

Next, we can observe that for typical spin resonance experiments, the spin-induced change in the inductance L_{spin} is much smaller than L_0. This has two major consequences for the modeling of the spin-induced frequency change: First, it implies that the spin-induced frequency change is most often relatively small, and therefore, the fixed-point equation defined in Eq. (53) can be approximately solved by a fixed-point iteration stopped after the first step. This essentially allows one to compute the spin-induced inductance from

$$L_{\mathrm{spin}} = \gamma\, T_2^2 \cdot \int_{V_{\mathrm{s}}} \frac{M_0 \cdot \left(\omega_{\mathrm{osc},0} - \omega_{\mathrm{L}}\right) \cdot \left(B_{1\mathrm{xu}}^2 + B_{1\mathrm{yu}}^2\right)}{1 + \left(\omega_{\mathrm{osc},0} - \omega_{\mathrm{L}}\right)^2\, T_2^2 + T_1\, T_2\, \left(\left(\gamma\, B_{1x}\right)^2 + \left(\gamma\, B_{1y}\right)^2\right)} \mathrm{d}V, \tag{56}$$

where $\omega_{\mathrm{osc},0} = \omega_{\mathrm{osc,spin}}\left(L_{\mathrm{spin}} = 0\right)$ is the oscillation frequency in the absence of spin resonance.

Second, the resulting explicit equation for the oscillation frequency can further be simplified by developing the right-hand side of Eq. (55) into a first-order Taylor series in L_{spin} about the point L_0, resulting in

$$\omega_{\mathrm{osc,spin}} \approx \omega_{\mathrm{osc},0} \cdot \left(1 - \frac{1}{2} \cdot \sqrt{\frac{L_0 + L_{\mathrm{spin}}}{L_0}} \cdot \frac{L_{\mathrm{spin}}}{L_0}\right) \approx \omega_{\mathrm{osc},0} \cdot \left(1 - \frac{1}{2} \cdot \frac{L_{\mathrm{spin}}}{L_0}\right). \tag{57}$$

Therefore, the spin-induced change in oscillation frequency is to first order given by

$$\Delta\omega_{\mathrm{osc,spin}} = \omega_{\mathrm{osc,spin}} - \omega_{\mathrm{osc},0} \approx -\frac{1}{2} \cdot \omega_{\mathrm{osc},0} \cdot \frac{L_{\mathrm{spin}}}{L_0}. \tag{58}$$

Before proceeding further, it is instructive to note that by inserting the expression of Eq. (17b) for L_{spin}, the spin-induced frequency change can be written as

$$\Delta\omega_{\mathrm{osc,spin}} \approx -\frac{1}{2} \cdot \frac{\omega_{\mathrm{osc},0}}{\mu_0 \cdot \hat{I} \cdot L_0} \cdot \int\limits_{V_{\mathrm{s}}} \bar{\mathbf{B}}_{\mathrm{u}} \cdot \bar{\mathbf{M}}_{\mathrm{ss}} \mathrm{d}V, \tag{59}$$

where $\bar{\mathbf{B}}_{\mathrm{u}} = B_{1\mathrm{xu}}\mathbf{e}_{\mathrm{x}} + B_{1\mathrm{yu}}\mathbf{e}_{\mathrm{y}}$ and $\bar{\mathbf{M}}_{\mathrm{ss}} = \bar{M}_{\mathrm{x,ss}}\mathbf{e}_{\mathrm{x}} + \bar{M}_{\mathrm{y,ss}}\mathbf{e}_{\mathrm{y}}$ with $\bar{M}_{\mathrm{x,ss}}$ and $\bar{M}_{\mathrm{y,ss}}$ defined in Eq. (8). Interestingly, this simplified detection equation is conceptually very similar to the detection equation of resonator-based spin detectors derived from the reciprocity principle, cf. [22].

Toward a simple closed-form expression for the LOD of oscillator-based spin detectors, Eq. (58) can be further simplified using the approximate expression of L_{spin} given by Eq. (21), yielding

$$\Delta\omega_{\mathrm{osc,spin}} \approx -\frac{1}{2} \cdot \omega_{\mathrm{osc},0} \cdot \eta \cdot \chi'. \tag{60}$$

The signal in a spin resonance experiment according to Fig. 1 can be defined as the peak-to-peak variation of the direct-detection dispersion spectrum. The noise floor, on the other hand, is given by the total white noise floor due to the coil resistance and the cross-coupled transistor pair given by Eq. (44). Then the spot SNR, i.e., the

SNR in a bandwidth of $\Delta f = 1\,\text{Hz}$, of the experiment can be defined according to

$$\text{SNR}(B_1) = \frac{\Delta\omega_{\text{osc,spin,pp}}}{\sqrt{2S_{\delta\omega,\delta\omega,\text{tot,white}}(\omega)}}. \tag{61}$$

To find the maximum SNR as a function of the applied B_1-field, we can use Eq. (50) as a starting point and ignore the dependence of the oscillation amplitude and frequency on the spin resonance,[14] yielding

$$I_{\text{coil}} \approx \frac{A_0}{\omega_{\text{osc},0}\, L_0}. \tag{62}$$

Furthermore, for a simple loop coil, the current I_{coil} can be related to the B_1-field according to $I_{\text{coil}} = 2\, B_1\, d_{\text{coil}}/\mu_0$. Also, assuming a uniform B_{u}-field, the inductance L_0 can be approximated by $L_0 \approx V_{\text{coil}} \cdot B_{\text{u}}^2/\mu_0$, where V_{coil} is the sensitive volume of the coil and $B_{\text{u}} \approx \mu_0/d_{\text{coil}}$. Then, the oscillation amplitude can be rewritten as a function of the B_1-field according to

$$A_0 \approx 2 \cdot B_1 \cdot \omega_{\text{osc},0} \cdot \frac{B_{\text{u}} \cdot V_{\text{coil}}}{\mu_0}. \tag{63}$$

Inserting this finding into Eq. (44) and further noting that the peak-to-peak signal $\Delta\omega_{\text{osc,spin,pp}}$ can be computed by replacing χ' in Eq. (60) by Eq. (22), the SNR of Eq. (61) can be rewritten according to

$$\text{SNR}(B_1) = \frac{B_1\, B_{\text{u}}\, V_{\text{s}} \sqrt{1 + T_1 T_2(\gamma B_1^2) + (T_2\omega_L^2)}}{\sqrt{2}\mu_0 \cdot \sqrt{(1+\alpha_{\text{od}}\gamma_{\text{nD}})kT\, R_{\text{coil}}} \cdot \sqrt{1+\gamma^2 B_1^2 T_1^2 T_2^2}} \cdot \chi_0\, \omega_{\text{osc},0}, \tag{64}$$

where $\chi_0 = M_0/H_0 = M_0/(B_0/\mu_0)$ and we used the fact that $\eta \approx V_{\text{s}}/V_{\text{coil}}$. Then, noting that for all practical values of B_1, T_2 and ω_L, we have that $1 + T_1 T_2(\gamma B_1^2) \ll (T_2\omega_L^2)$, the expression for the SNR can be further simplified according to:

$$\text{SNR}(B_1) = \frac{B_1\, B_{\text{u}}\, V_{\text{s}}\, T_2}{\sqrt{2}\mu_0 \cdot \sqrt{(1+\alpha_{\text{od}}\,\gamma_{\text{nD}})\, kT\, R_{\text{coil}}} \cdot \sqrt{1+\gamma^2 B_1^2 T_1^2 T_2^2}} \cdot \chi_0 \omega_{\text{osc},0}^2, \tag{65}$$

where we have used the assumption $|\omega_L| = \omega_{\text{osc},0}$. At this point it is instructive to note that for spin-half particles χ_0 can be calculated from the spin concentration of the sample N_{spin} according to $\chi_0 = \mu_0 N_{\text{spin}} \gamma^2 \hbar^2/(4kT)$, where $\hbar$ is the reduced Planck constant, cf., e.g., [42]. From Eq. (64) it is clear, that the SNR is a strong function of the B_1-field. In fact, the maximum theoretically achievable SNR is obtained for $B_1 \mapsto \infty$ and given by:

[14] This again corresponds to stopping the fixed-point iteration after the first step.

$$\mathrm{SNR}_{\max} = \frac{B_\mathrm{u}\, V_\mathrm{s}}{\sqrt{2}\mu_0 \cdot \sqrt{(1+\alpha_\mathrm{od}\,\gamma_\mathrm{nD})\; kT\; R_\mathrm{coil}}} \cdot \sqrt{\frac{T_2}{T_1}} \cdot \omega_{\mathrm{osc},0} \cdot \chi_0 \cdot B_0, \tag{66}$$

where we have used the assumption that $|\gamma B_0| = \omega_{\mathrm{osc},0}$. Therefore, for an optimum sample with $T_2 = 2T_1$, the achievable SNR becomes:

$$\mathrm{SNR}_{\max,\mathrm{opt}} = \frac{B_\mathrm{u}\, V_\mathrm{s}}{\mu_0 \cdot \sqrt{(1+\alpha_\mathrm{od}\,\gamma_\mathrm{nD})\; kT\; R_\mathrm{coil}}} \cdot \omega_{\mathrm{osc},0} \cdot \chi_0 \cdot B_0, \tag{67}$$

From Eq. (67), the achievable spin sensitivity N_min of frequency-sensitive oscillator-based spin detectors can be defined according to

$$N_\mathrm{min} \triangleq \frac{3 \cdot N_\mathrm{spin} \cdot V_\mathrm{s}}{\mathrm{SNR}_{\max,\mathrm{opt}}}, \tag{68}$$

where $N_\mathrm{spin} \cdot V_\mathrm{s}$ is the total number of spins in the sample leading to the SNR of $\mathrm{SNR}_{\max,\mathrm{opt}}$, and the factor of 3 has been introduced to be compatible with standard ESR literature. Then, inserting the finding of Eq. (67) into Eq. (68), one finally obtains

$$\boxed{N_\mathrm{min} = \frac{12\, kT\, \sqrt{(1+\alpha_\mathrm{od}\gamma_\mathrm{nD})\; kT\; R_\mathrm{coil}}}{\gamma^3\, \hbar^2\, B_\mathrm{u}\, B_0^2}}. \tag{69}$$

This result is up to factor of two identical to the result presented in [5] and up to a factor of $2\sqrt{1+\alpha_\mathrm{od}\gamma_\mathrm{nD}}$ identical to that presented in [42]. These differences exist because in previous works, cf., e.g., [42], the spin-induced frequency shift was calculated from an expression for the total inductance according to $L_\mathrm{tot} = L_0 \cdot (1 + \eta\chi')$, and the expression used for χ' did not take into account the entire magnetic energy associated with the spin-induced magnetization but only that associated with the x-component of the steady-state magnetization, explaining the difference of a factor of two. The factor of $\sqrt{1+2\,\alpha_\mathrm{od}\,\gamma_\mathrm{nD}}$ arises due to the noise produced by the cross-coupled differential pair, which was not considered in [42].

At this point it is instructive to note that assuming $\gamma_\mathrm{nD} \approx 1$ and for a reasonable choice of $\alpha_\mathrm{od} = 3$, Eq. (69) can be rewritten according to

$$N_\mathrm{min} = \frac{24\, kT\, \sqrt{kT\; R_\mathrm{coil}}}{\gamma^3\, \hbar^2\, B_\mathrm{u}\, B_0^2}, \tag{70}$$

which is identical to the theoretical spin sensitivity of resonator-based spin detectors derived in [12].

6 Comparison with Measured Data

Before concluding this chapter, it is instructive to compare both the signal and the noise models developed above against measured data from the prototype sensor realization presented in [1]. To this end, we will first use the numerical model for the

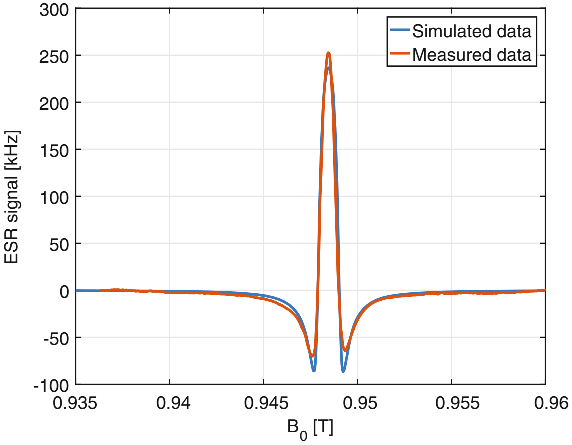


Fig. 4 Comparison of simulation results obtained using the numerical model of Sect. 5.1 and measured data obtained using the detector presented in [1]. The parameters used for the simulation are spin density $N_{\text{spin}} = 2 \times 10^{27}$ spins/m^3, filling factor $\eta = 3.1 \times 10^{-4}$, $T_1 = T_2 = 62$ ns, temperature $T = 290$ K, tank capacitance $C = 200$ fF, tank inductance $L_0 = 170$ pH, coil resistance $R_{\text{coil}} = 0.8\,\Omega$, transconductance of a single transistor $G_{\text{m0}} = 2$ mS, corresponding to a supply voltage of $V_{\text{DD}} = 0.6$ V, modulation amplitude $B_{\text{m}} = 400\,\mu$T. The measured sample has a size of approximately $(7\,\mu\text{m})^3$

signal calculation given by Eq. (53) and the initial values of Eq. (54) to solve the fixed-point iteration required to calculate the resulting change in oscillation frequency. Once the direct detection spectrum is calculated in this way, it is a straightforward task to include the field-modulation as it is applied in the measurement. To perform the comparison between model and experiment, the detector presented in [1] was used together with a small sample of DPPH (2,2-diphenyl-1-picrylhydrazyl)—a standard ESR sample—with an approximate size of $(7\,\mu\text{m})^3$. The corresponding measured and simulated spectra are shown in Fig. 4. The corresponding simulation parameters are listed in the figure caption. According to the figure, the agreement between the model and the measured data is very good. The residual error in the peak height can most likely be attributed to the imperfect estimation of the sample size using a light microscope. The residual error in the line shape can most likely be attributed to the uncertainty in the modeling of the actual amplitude of the utilized modulation field B_{m} and the actual B_1-field produced by the oscillator.

Finally, we will compare the frequency noise model developed in Sect. 4.3 with simulation results from a commercial circuit simulator (Cadence's SpectreRF) and measured data from the prototype oscillator-based spin detector presented in [1]. Here, in order to be able to compare the frequency noise predicted by the model against the phase noise spectra simulated in SpectreRF and those measured using a signal source analyzer, one can apply the following transformation:

$$S_{\delta\varphi,\delta\varphi}(\omega) \approx \frac{1}{\omega^2}\, S_{\delta\omega,\delta\omega}(\omega) = \frac{1}{f^2}\, S_{\delta f,\delta f}(2\pi f), \tag{71}$$

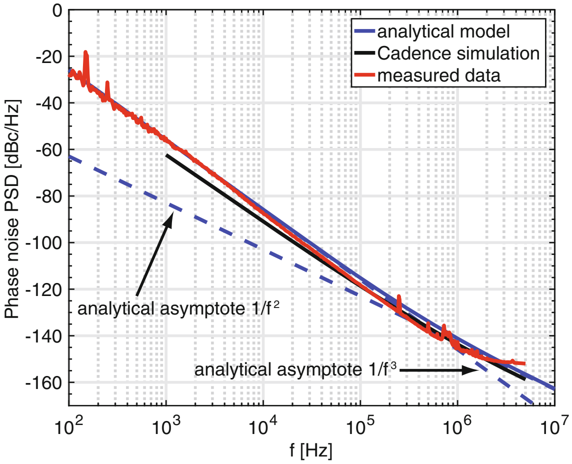


Fig. 5 Comparison of the phase noise predicted by the model of Sect. 4.3 with simulations in Cadence's SpectreRF and with measured data from the detector prototype presented in [1]. The experimental conditions and the data used for the model-based calculation are oscillator supply voltage VDD = 1.5 V, $f_{\mathrm{osc},0} = 27\,\mathrm{GHz}$, $L_0 = 170\,\mathrm{pH}$, $R_{\mathrm{coil}} = 0.8\Omega$, $\alpha_{\mathrm{od}} = 5$, $\alpha_{\mathrm{H}} = 10^{-6}$, $L_{\mathrm{T}} = 120\,\mathrm{nm}$, $V_{\mathrm{T0}} = 0.4\,\mathrm{V}$, and $\gamma_{\mathrm{nD}} = 2/3$

which relates the PSD of the frequency noise process $\delta\omega(t)$ to that of the phase noise process $\delta\varphi(t)$. The corresponding simulated and measured spectra are displayed in Fig. 5. The experimental conditions and the data used for the model-based calculation are listed in the figure caption. According to the figure, the model prediction is in good agreement with the measured data and therefore is validated for its use in the limit-of-detection calculation of Sect. 5.2.

7 Conclusion and Outlook

Today, the field of oscillator-based spin-detection is receiving significant attention in the research community, because it allows for the realization of inexpensive, highly sensitive miniaturized detectors in integrated circuit technology [1, 16, 21, 42]. Moreover, thanks to the possibility of realizing oscillators in CMOS IC technology at very high frequencies, cf., e.g., [33], which bears the potential of significantly improving the achievable spin sensitivity compared to the first presented prototypes of oscillator-based spin detectors [1, 16, 21, 42], a further increase in research activities in this field can be expected. This increasing interest clearly mandates the availability of a good model of this intrinsically nonlinear type of sensor, which accurately predicts the expected performance. The modeling presented in this chapter reflects these needs by extending the state-of-the-art [5, 42] by incorporating both inhomogeneous distributions of the oscillator's B_1-field and nonlinear effects in the oscillator into the model. The resulting improved accuracy is essential for a future quantitative modeling as is, e.g., required for quantitative spin-detection experiments.

References

1. Anders, J., Angerhofer, A., Boero, G.: K-band single-chip electron spin resonance detector. J. Magn. Reson. **217**, 19–26 (2012)
2. Anders, J., Chiaramonte, G., SanGiorgio, P., Boero, G.: A single-chip array of nmr receivers. J. Magn. Reson. **201**(2), 239–249 (2009)
3. Anders, J., Ortmanns, M.: Frequency noise of CMOS LC tank oscillators operating in weak inversion. In: 2013 European Conference on Circuit Theory and Design (ECCTD), pp. 1–4 (2013)
4. Anders, J., Ortmanns, M., Boero, G.: Frequency noise in current-starved cmos lc tank oscillators. In: Nonlinear Dynamics of Electronic Systems, Proceedings of NDES 2012, pp. 1–4 (2012)
5. Anders, J., Ortmanns, M., Boero, G.: Noise in frequency-sensitive esr detectors. In: MATHMOD 2012, vol. 7, pp. 451–456 (2012)
6. Anders, J., SanGiorgio, P., Boero, G.: A fully integrated IQ-receiver for NMR microscopy. J. Magn. Reson. **209**(1), 1–7 (2011)
7. Anders, J., SanGiorgio, P., Deligianni, X., Santini, F., Scheffler, K., Boero, G.: Integrated active tracking detector for MRI-guided interventions. Magn. Reson. Med. **67**(1), 290–296 (2012)
8. Anders, J.: Fully-integrated CMOS Probes for Magnetic Resonance Applications. Ph.D thesis, EPFL, Lausanne (2011)
9. Badilita, V., Kratt, K., Baxan, N., Anders, J., Elverfeldt, D., Boero, G., Hennig, J., Korvink, J.G., Wallrabe, U.: 3d solenoidal microcoil arrays with CMOS integrated amplifiers for parallel MR imaging and spectroscopy. In: 2011 IEEE 24th International Conference on Micro Electro Mechanical Systems (Mems), pp. 809–812 (2011)
10. Badilita, V., Kratt, K., Baxan, N., Mohmmadzadeh, M., Burger, T., Weber, H., von Elverfeldt, D., Hennig, J., Korvink, J.G., Wallrabe, U.: On-chip three dimensional microcoils for MRI at the microscale. Lab on a Chip **10**(11), 1387–1390 (2010)
11. Blank, A., Dikarov, E., Shklyar, R., Twig, Y.: Induction-detection electron spin resonance with sensitivity of 1000 spins: en route to scalable quantum computations. Phys. Lett. A **377**(31–33), 1937–1942 (2013)
12. Boero, G., Bouterfas, M., Massin, C., Vincent, F., Besse, P.A., Popovic, R.S., Schweiger, A.: Electron-spin resonance probe based on a 100 mu m planar microcoil. Rev. Sci. Instrum. **74**(11), 4794–4798 (2003)
13. Demir, A., SangiovanniVincentelli, A.L.: Simulation and modeling of phase noise in open-loop oscillators. In: Proceedings of the IEEE 1996 Custom Integrated Circuits Conference, pp. 453–456, 516 (1996)
14. Enz, C., Vittoz, E.A.: Charge-based MOS Transistor Modeling: the EKV Model for Low-power and RF IC Design. Wiley, Chichester (2006)
15. Gruschke, O.G., Baxan, N., Clad, L., Kratt, K., von Elverfeldt, D., Peter, A., Hennig, J., Badilita, V., Wallrabe, U., Korvink, J.G.: Lab on a chip phased-array MR multi-platform analysis system. Lab on a Chip **12**(3), 495–502 (2012)
16. Gualco, G., Anders, J., Sienkiewicz, A., Alberti, S., Forro, L., Boero, G.: Cryogenic single-chip electron spin resonance detector. J. Magn. Reson. **247**, 96–103 (2014)
17. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Applied Mathematical Sciences, 5th edn. Springer, New York (1997)
18. Ha, D., Paulsen, J., Sun, N., Song, Y.Q., Ham, D.: Scalable nmr spectroscopy with semiconductor chips. Proc. Nat. Acad. Sci. U.S.A **111**(33), 11955–11960 (2014)

19. Hajimiri, A., Lee, T.H.: A general theory of phase noise in electrical oscillators. IEEE J. Solid-State Circuits **33**(2), 179–194 (1998)
20. Ham, D., Hajimiri, A.: Virtual damping and Einstein relation in oscillators. IEEE J. Solid-State Circuits **38**(3), 407–418 (2003)
21. Handwerker, J., Schlecker, B., Wachter, U., Radermacher, P., Ortmanns, M., Anders, J.: 28.2 a 14ghz battery-operated point-of-care esr spectrometer based on a 0.13m CMOS asic. In: 2016 IEEE International Solid-State Circuits Conference (ISSCC), pp. 476–477 (2016)
22. Hoult, D.I.: The principle of reciprocity in signal strength calculations - a mathematical guide. Concepts Magn. Reson. **12**(4), 173–187 (2000)
23. Kaertner, F.X.: Analysis of white and f-alpha noise in oscillators. Int. J. Circuit Theory Appl. **18**(5), 485–519 (1990)
24. Khalil, H.K.: Nonlinear Systems, 3rd edn. Prentice Hall, Upper Saddle River (2002)
25. Lax, M.: Classical noise .v. noise in self sustained oscillators. Phys. Rev. **160**(2), 290 (1967)
26. Leeson, D.B.: A simple model of feedback oscillator noise spectrum. Proc. Inst. Electr. Electron. Eng. **54**(2), 329–330 (1966)
27. Leidich, S., Braun, M., Gessner, T., Riemer, T.: Silicon cylinder spiral coil for nuclear magnetic resonance spectroscopy of nanoliter samples. Concepts Magn. Reson. Part B Magn. Reson. Eng. **35B**(1), 11–22 (2009)
28. Mathis, W., Thiessen, T.: On noise analysis of oscillators based on statistical mechanics. Noise Anal. Oscill. Based Stat. Mech. **56**(4), 357–366 (2010)
29. Narkowicz, R., Suter, D.: Tuner and radiation shield for planar electron paramagnetic resonance microresonators. Rev. Sci. Instrum. **86**(2) (2015)
30. Narkowicz, R., Suter, D., Niemeyer, I.: Scaling of sensitivity and efficiency in planar microresonators for electron spin resonance. Rev. Sci. Instrum. **79**(8) (2008)
31. Papoulis, A., Pillai, S.U.: Probability, Random Variables, and Stochastic Processes, 4th edn. McGraw-Hill, Boston (2002)
32. Poole, C.P.: Electron Spin Resonance: A Comprehensive Treatise on Experimental Techniques. Dover Publications, Mineola, N.Y. (1996)
33. Razavi, B.: A 300-ghz fundamental oscillator in 65-nm CMOS technology. IEEE J. Solid-State Circuits **46**(4), 894–903 (2011)
34. Ryan, H., Song, S.H., Zass, A., Korvink, J., Utz, M.: Contactless NMR spectroscopy on a chip. Anal. Chem. **84**(8), 3696–3702 (2012)
35. Spengler, N., Hofflin, J., Moazenzadeh, A., Mager, D., MacKinnon, N., Badilita, V., Wallrabe, U., Korvink, J.G.: Heteronuclear micro-helmholtz coil facilitates mu m-range spatial and sub-hz spectral resolution NMR of NL-volume samples on customisable microfluidic chips. Plos One **11**(1) (2016)
36. Stratonovich, R.L.: Topics in the Theory of Random Noise. Mathematics and its applications, rev. English edn. Gordon and Breach, New York (1963)
37. Summanen, K.T.: A theory of the bloch-siegert shift. Phys. Lett. A **155**(4–5), 335–336 (1991)
38. Sun, N., Yoon, T.J., Lee, H., Andress, W., Weissleder, R., Ham, D.: Palm NMR and 1-chip NMR. IEEE J. Solid-State Circuits **46**(1), 342–352 (2011)
39. Twig, Y., Dikarov, E., Blank, A.: Ultra miniature resonators for electron spin resonance: sensitivity analysis, design and construction methods, and potential applications. Mol. Phys. **111**(18–19), 2674–2682 (2013)
40. Twig, Y., Suhovoy, E., Blank, A.: Sensitive surface loop-gap microresonators for electron spin resonance. Rev. Sci. Instrum. **81**(10) (2010)
41. Weil, J.A., Bolton, J.R.: Electron Paramagnetic Resonance: Elementary Theory and Practical Applications, 2nd edn. Wiley-Interscience, Hoboken, N.J. (2007)
42. Yalcin, T., Boero, G.: Single-chip detector for electron spin resonance spectroscopy. Rev. Sci. Instrum. **79**(9) (2008)

Effect of Nonlinearity and Boiler Dynamics in Automatic Generation Control of Multi-area Thermal Power System with Proportional-Integral-Derivative and Ant Colony Optimization Technique

K. Jagatheesan, B. Anand, K. Baskaran, N. Dey, A.S. Ashour and V.E. Balas

Abstract This work presents the automatic generation control (AGC) of a multiarea interconnected power system. The investigated multiarea power system is prepared with three equal reheat thermal power systems with suitable governor unit, turbine unit, generator unit, speed regulator unit, tie-line in each unit, and secondary proportional-integral-derivative (PID) controller. During nominal loading conditions, the power generating unit offers good quality of power to consumers. Nevertheless, the occurrence of sudden load disturbance in the interconnected power generating unit affects the entire performance (consistency in system frequency and voltage) and system stability. In order to moderate this big pose, the PID controller is introduced as a secondary controller. Jointly with the proper selection of the controller parameters (proportional gain (KP), integral gain (KI), and derivative gain (KD)) a good quality

K. Jagatheesan (✉)
Paavai Engineering College, Namakkal, Tamilnadu, India
e-mail: jaga.ksr@gmail.com

B. Anand
Hindusthan College of Engineering and Technology, Coimbatore, Tamilnadu, India
e-mail: b_anand_eee@yahoo.co.in

K. Baskaran
Government College of Technology, Coimbatore, Tamilnadu, India
e-mail: drbaskaran@gct.ac.in

N. Dey
Techno India College of Technology, Kolkata, India
e-mail: neelanjandey@gmail.com

A.S. Ashour
Tanta University, Tanta, Egypt
e-mail: amirasashour@yahoo.com

V.E. Balas
Aurel Vlaicu University of Arad, Arad, Romania
e-mail: valentina.balas@uav.ro

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_5

of power supply is crucial in a power system for generating. An artificial intelligence (AI) based ant colony optimization (ACO) technique is considered for tuning the control parameters. Further, in the current chapter, nonlinearity and boiler dynamics effects are considered to evaluate the performance of the investigated power system. The nonlinearities are generation rate constraints (GRC) and governor dead band (GDB). The drum-type oil-fired boiler system is considered in this work. The nonlinearity effect and boiler dynamics in the investigated power systems are derived by considering different scenarios: (a) GRC in all areas and two percent step load perturbation (2% SLP) in area 1, (b) GDB in all areas and two percent step load perturbation (2% SLP) in area 1 (c) GRC and GDB in all areas and two percent step load perturbation (2% SLP) in area 1 and (d) GRC, GDB, and boiler dynamics (BD) in all areas and two-percent step load perturbation (2% SLP) in area 1. Time-domain specification analysis is used for the evaluation of nonlinearity and boiler dynamics effect.

Keywords Ant colony optimization · Artificial intelligence · Automatic generation control · Nonlinearity · Time domain specification

1 Introduction

Nowadays, load frequency control (LFC) of automatic generation control (AGC) plays a major role in large-scale interconnected power generating units. The power generating units are interconnected through tie-lines for power transfer during sudden load demand conditions. In [1, 2], various previous studies related to LFC/AGC of single/multiarea interconnected power systems were reported. The performance of the power system and controllers were affected by introducing nonlinearities and boiler dynamics into the power generating unit. The response of the power system gets more damping oscillations with affected system stability by adding nonlinearity and the boiler dynamics into the power generating unit. Numerous studies related to nonlinear power systems with boiler dynamics were reported in [1–43].

Saikia et al. [1] introduced the generation rate constraint (GRC) in the AGC of three-area interconnected hydrothermal power systems and investigated power systems equipped with fuzzy integral double derivative (FIDD) controllers. The controller gain values were optimized by considering the bacterial foraging (BF) optimization technique. In [2], the AGC of a multiarea thermal power system under a deregulated environment was presented with GRC nonlinearity. The system performance was improved by introducing a fractional order PID (FOPID) controller, which employed the bacterial foraging (BF) optimization technique for tuning the controller parameters.

Sahu et al. [3] presented the LFC of a two-area reheat thermal power system with governor dead band (GDB) nonlinearity. The performance of the nonlinear power systems was improved by implementing a differential evolution (DE) algorithm that optimized a two-degree-of-freedom PID (2DOF-PID) controller. The integral time absolute time error (ITSE), integral square error (ISE), and integral time square error (ITSE) were used as objective functions. Shabani et al. [4] proposed the LFC of nonlinear power system with GRC nonlinearity and imperialist control algorithm (ICA) optimized PID controller. The LFCs of a two-area interconnected power system with GDB nonlinearity and an SMES unit-equipped power system were presented in [5]. Additionally, the cuckoo search (CS) designed proportional-integral (PI) controller was implemented in the same investigated power system.

The GRC and GDB nonlinearity were considered in the LFC of a two-area interconnected reheat thermal power system with appropriate fuzzy logic controller (FLC). The results proved the superiority of the FLC compared to the conventional PI controller in [6]. The AGC of a two-area interconnected reheat thermal power system was presented in [7], with GRC nonlinearity and fuzzy logic controller (FLC). The GRC nonlinearity was considered in the adaptive LFC of a two-area interconnected hydrothermal power system in [8] with adaptive implicit hybrid self-tuning controller.

Parida et al. [9]introduced the AGC of a deregulated hydrothermal power system employing the GRC nonlinearity and artificial neural network (ANN) based two-area reheat thermal power system with GDB nonlinearity. Robust decentralized frequency stabilizers were designed for a three-area interconnected reheat thermal power system in [10], by considering GRC, GDB, BD, SMES storage units and integral controller in all three interconnected power systems. The effects of the speed governor dead band nonlinearity on the performance of the system frequency and tie-line power flow between interconnected power systems were analyzed in [11]. The investigated power system consists of two power-generating units.

Concordia et al. [12] applied an ANN technique based on μ-synthesis to the LFC of an interconnected power system. The authors considered the GRC nonlinearity to examine the robustness performance of the proposed controller. The GDB and BD nonlinearity equipped power system was studied in [13] by considering the effect of a superconducting magnetic energy storage (SMES) unit and integral controller. The LFC of the interconnected reheat thermal power system parameters were optimized using a Lyapunov technique with the effect of governor dead band (backlash) nonlinearity in the power system [14].

The AGC of an interconnected two-area hydrothermal power system was presented by considering continuous and discrete modes with GRC nonlinearity in [15]. Nanda et al. [16] suggested the AGC of a two-area reheat thermal power system with nonlinearity. The authors optimized the speed regulation parameters based on an optimum selection method. The effect of a battery energy storage system into the LFC of an interconnected power system was presented in [17] by considering GRC and GDB nonlinearity effects. The LFC of a single-area thermal power system with adaptive controller and RC nonlinearity effect was discussed in [18].

A bacterial foraging (BF) optimization based integral controller was designed and implemented in the AGC of a three-area interconnected reheat thermal power system in [19]. The investigated power system was equipped with GRC nonlinearity to analyze the performance of the proposed technique. The AGC of a two-area interconnected hydrothermal power system was studied with several classical controllers in [20]. The investigated thermal power system was equipped with different steam configuration-based turbines and a hydropower system equipped with an electric governor. The LFC of a two-area hydrothermal power system was studied with GDB, GRC nonlinearity and boiler dynamics in [21]. The effects of nonlinearity, boiler dynamics, and load disturbances were overcame by developing a proper fuzzy logic controller.

Sample data automatic generation control of a two-area interconnected reheat thermal power system was discussed in [22], by considering GRC nonlinearity with classical controllers. Several classical controllers were implemented in AGC of a three-area interconnected reheat thermal power system in [23] using GRC nonlinearity. The classical controller gain values were optimized using an evolutionary computational-based bacterial foraging optimization technique. The effect of GDB nonlinearity was considered in AGC of the two-area interconnected reheat thermal power system studied in [24], and integral controller gain, frequency bias parameter values were optimized using the Lyapunov technique.

The AGC control of a three-area hydropower system with classical controllers was presented in [25]. The effects of GRC nonlinearity and controller gain values were optimized based on the bacterial foraging technique. Decentralized biased controller-based LFC of an interconnected power system was presented in [26], by considering the effects of GDB nonlinearity with an integral square error objective function-based PI controller. The cuckoo search (CS) algorithm-based AGC of a two-area interconnected reheat thermal power system was designed in [27]. The authors used a superconducting magnetic energy storage (SMES) unit with GRC nonlinearity.

A fractional order PID (FOPID) controller-based LFC of a two-area interconnected power system was designed considering the effect of dead zone and GRC nonlinearity in [28]. The FOPID controller gain values were optimized using the chaotic multiobjective optimization-based technique. Several flexible alternating current transmission system (FACTS) devices and 2-DOF controllers (2DOF-PI, 2DOF-PD, 2DOF-IDD, 2DOF-IDD) based three-area reheat thermal power systems were studied in [28]. The authors included the GRC nonlinearity effect in the power system as well as optimizing the controller gain values based on the cuckoo search (CS) algorithm.

A cascade PD-PID (proportional-derivative (PD)–(proportional-integral-derivative (PID))) controller-based AGC of a three-area reheat thermal power system was designed with GRC nonlinearity in [29]. In addition, the gain values of the cascade controller were optimized using the bat algorithm. The AGCs of a multisource interconnected power system with GRC and GDB nonlinearity were presented in [30] with an improved particle swarm optimization (IPSO) technique. Sahu et al. [31] presented a hybrid firefly algorithm and pattern search technique-based AGC of a multiarea power system considering the effect of GRC nonlinearity with a PID controller. An LFC of an unequal three-area thermal power system was designed by considering GRC and GDB nonlinearity with a PI/PID controller in [32]. The gain values were optimized based on the firefly algorithm (FA).

The performance of the cuckoo search (CS) algorithm was analyzed by implementing several CS optimized 2DOF controllers (2DOF-PI, 2DOF-PD, 2DOF-IDD, 2DOF-IDD) in a three-area power system with GRC nonlinearity [33]. The AGC of three unequal interconnected reheat thermal power systems was designed by considering the time delay, boiler, and GRC nonlinearity using the online wavelet filter in [34]. The AGC of a multiarea multiunit thermal power system was designed in [35] using the thyristor controlled series compensator (TCSC) and SMES energy storage unit. Automatic generation control of a two-are reheat thermal power system was studied by considering GRC nonlinearity in [36]. The performance of the discrete optimum integral controller was compared with optimum proportional-integral controllers.

2 Three-Area Interconnected Reheat Thermal Power System with Nonlinearity and Boiler Dynamics

The load frequency control (LFC) of an investigated multiarea power system is incorporated with three equal reheat thermal power systems and all the areas are connected through a tie-line. Each power generating unit is equipped with appropriate governor, turbine, and reheater unit, generate rate constraint (GRC), governor dead band (GDB) nonlinearity, and boiler dynamics (BD). The transfer function model of the three-area interconnected power system is illustrated in Fig. 1. In three-area interconnected power systems, area 1, area 2, and area 3 are similar power rated thermal power systems, and all the three areas are interconnected through a tie-line [37–39, 41–43]. In Fig. 1, R_1, R_2, and R_3 represent the self-regulation parameters for the governor in areas 1, 2, and 3, respectively in $T_p = (Hz/puMW)$; T_{g1}, T_{g2}, and T_{g3} refer to the speed governor time constants of areas 1, 2, and 3, respectively, (in seconds); T_{r1}, T_{r2}, and T_{r3} are the reheat time constants in areas 1, 2, and 3, respectively (in seconds); K_{r1}, K_{r2}, and K_{r3} are the reheater coefficients of areas 1, 2, and 3, respectively; T_{t1}, T_{t2}, and T_{t3} are the steam chest time constants of areas 1, 2,

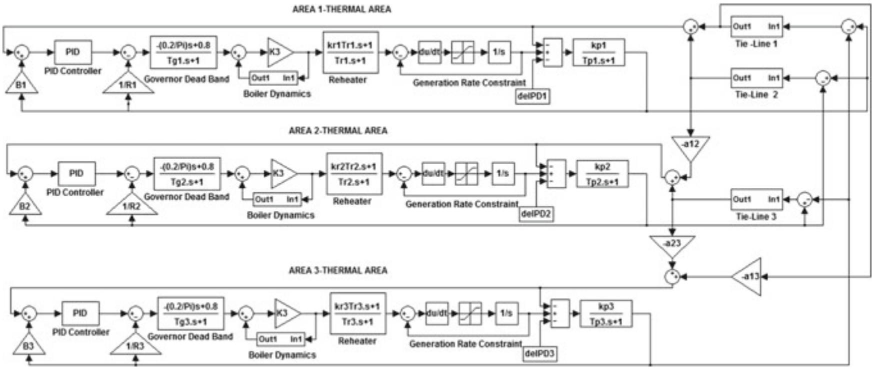


Fig. 1 Three-area interconnected power system with nonlinearity and boiler dynamics

Table 1 Transfer function model of power system components and their nominal parameters

Component	Transfer function	Nominal parameters
Governor	$\frac{1}{T_{gi}S+1}$	$\frac{1}{0.2S+1}$
Reheater	$\frac{K_{ri}Tri_{S}+1}{T_{ri}S+1}$	$\frac{3.33S+1}{10S+1}$
Turbine	$\frac{1}{T_{ii}S+1}$	$\frac{1}{0.3S+1}$
Power System	$\frac{K_p}{T_{Pi}S+1}$	$\frac{120}{20S+1}$

and 3, respectively (in seconds); T_{p1}, T_{p2}, and T_{p3} are the power system time constants of areas 1, 2, and 3, respectively given by ($T_p = 2H/f * D$) (in seconds); K_{p1}, K_{p2}, and K_{p3} are the power system gains of areas 1, 2, and 3, respectively ($K_p = 1/D$); B_1, B_2, and B_3 are the frequency bias constants; f-nominal frequency (Hz), H-Inertia constants of each area; $D - \Delta_{PD}/\Delta f$ (pu/Hz); delPtie is the incremental tie-line power change between connected areas (pu MW); $delF_1$, $delF_2$, and $delF_3$ are the incremental frequency deviations of areas 1, 2, and 3, respectively (in Hz); ACE_1, ACE_2, and ACE_3 stand for the area control error of areas 1, 2, and 3, respectively (in pu). The transfer function of the power system and its values are given in Table 1.

When a sudden load demand or disturbance occurs in any part of an interconnected power system, it will affect the total system frequency, tie-line power flow control between the control areas, and the system's stability. The primary speed governor system takes necessary control action to restore the system parameters. Nevertheless, it is not sufficient to restore the system's stability. In order to overcome this big pose secondary, controllers are introduced.

In this study, the proportional-integral-derivative (PID) controller is introduced as a secondary controller to generate the necessary control signal. The control signal is then given to the power system as input signal. The input of the PID controller is the area control error (ACE), which defined as a linear combination of frequency deviation and tie-line power deviation.

The effective performance of the LFC of an interconnected power system and secondary controller performances are affected by the nonlinearity and boiler dynamics in the power system. The nonlinearity in the power system affects the amplitude, damping oscillations, and settling time of the system response with a particular time period of 20 s. In this work, the GRC, GDB, and boiler dynamics are considered.

The GDB is defined as the total magnitude of a sustained speed change with no change in the valve position. The speed governor dead band produces a great effect on the performance of the power system. Due to including the GDB in the investigated system, it becomes nonlinear. The GDB decreases the speed before the valve position changes and the system exhibits an oscillatory response [41]. The transfer function model of the GDB nonlinearity is termed a governor with the dead band

$$Gg = \frac{-\frac{0.2}{\Pi} + 0.8}{T_g S + 1}. \tag{1}$$

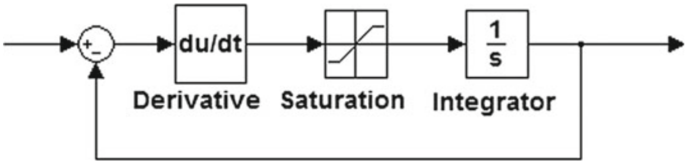


Fig. 2 Structure of the generation rate constraint

For this investigation, a backlash nonlinearity of 0.05% is considered for a thermal power system.

2.1 Generation Rate Constraint (GRC)

In practice, there are maximum and minimum limits for rate of change in generating power. Due to a sudden power drop, it would draw excessive steam from the boiler to cause steam condensation. The boiler keeps the steam pressure constant up to a maximum power generating limit. After the power generation reaches its upper limit, the turbine should be restricted by GRC [41]. The model of GRC is shown in Fig. 2

A generation rate constraint of 0.0017 p.u. MW sec^{-1} is considered for the thermal power system.

2.2 Boiler Dynamics (BD)

A boiler is a device that produces steam under pressure. There are three types of boiler system: gas or oil fired, coal fired well tuned, and coal fired poorly tuned. A gas or oil fired boiler is guaranteed to provide a quick response compared to other systems when load demand occurs. In this work, a drum-type boiler is considered. It is incorporated with long-term dynamics of fuel, steam flow on the boiler drum pressure, and combustion controls. The transfer function model of a drum-type boiler dynamics model is shown in Fig. 3 [41].

An oil/gas fired boiler system responds much more quickly during sudden load demand than a coal fired unit. A drum-type boiler is also called a recirculation boiler. In boiler dynamics, four different control strategies or modes are available: boiler leading, turbine leading coordinated boiler turbine, and sliding pressure control. In this study, the boiler leading or turbine following modes of operation are considered.

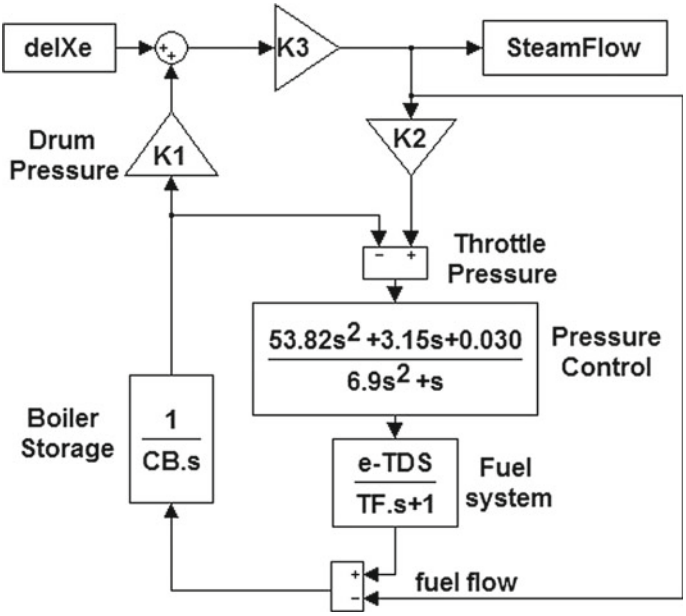


Fig. 3 Structure of the boiler dynamics

3 Design of a Proportional-Integral-Derivative (PID) Controller and Ant Colony Optimization (ACO) Technique

3.1 Proportional-Integral-Derivative (PID) Controller

In the current study, a proportional-integral-derivative (PID) controller is considered for load frequency control (LFC) of a three-area interconnected reheat thermal power system [37–39, 41–43]. The structure of the PID controller is shown in Fig. 4, where K_p, K_i, and K_d are proportional, integral and derivative gain values, respectively. The control input of PID controllers is the area control errors (ACEs), and the outputs of PID controllers (u_1,u_2, and u_3) are the input of the interconnected power system shown in Fig. 4:

$$ACE_i = B_i \Delta f_i + \Delta P_{tie,ij}. \tag{2}$$

The control outputs of the PID controllers are

$$U_1 = -K_i.ACE_1 - \frac{K_p}{T_i}\int ACE_1 - K_d T_d \frac{d}{dt} ACE_1, \tag{3}$$

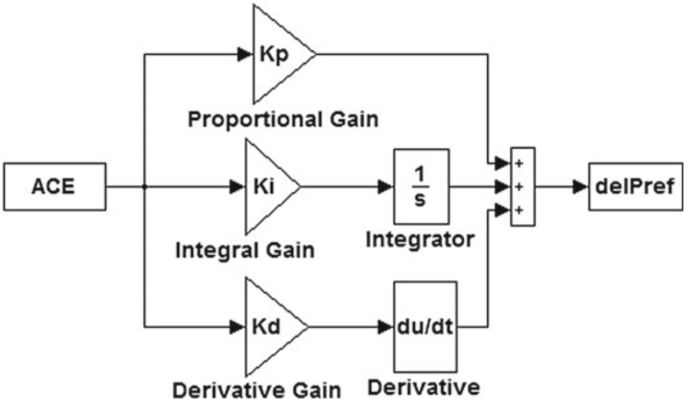


Fig. 4 Structure of the boiler dynamics

$$U_2 = -K_i.ACE_2 - \frac{K_p}{T_i}\int ACE_2 - K_d T_d \frac{d}{dt} ACE_2, \tag{4}$$

$$U_3 = -K_i.ACE_3 - \frac{K_p}{T_i}\int ACE_3 - K_d T_d \frac{d}{dt} ACE_3. \tag{5}$$

The transfer function of the PID controller is given by

$$U(S) = K_p E(S) + \frac{K_p}{T_i S} E(S) + K_d E(S). \tag{6}$$

During the design of an artificial intelligence (AI) based PID controller, the first objective function is defined based on the desired specification and constraints. In this study, the integral time absolute error (ITAE) criterion is considered as an objective function, since many literature surveys used this criterion because it reduces the settling time effectively. The expression for the ITAE objective function is depicted in the following equation [42]:

$$J = \int_0^{t_{sim}} t.[|\Delta f_1| + |\Delta f_2| + |\Delta P_{tie}|]dt. \tag{7}$$

In this study, the problem constraints are the PID controller parameters' limit. The design problem can be formulated as the minimization of performance index J within minimum and maximum values of PID controller parameters $K_{pmin} \leq K_p \leq K_{pmax}$, $K_{imin} \leq K_i \leq K_{imax}$, and $K_{dmin} \leq K_d \leq K_{dmax}$. The minimum and maximum values of the PID controller parameters are considered to be 0 and 1.

3.2 Ant Colony Optimization (ACO) Technique

Recently, artificial intelligence (AI) techniques have played a major role based on their capability for solving combinatorial optimization (CO) problems. These techniques are being used in many fields such as planning, operation, and control. In this work, AI based techniques are used to tune the controller parameters in the power system. Some AI based optimization techniques are genetic algorithm (GA), particle swarm optimization (PSO), stochastic particle swarm optimization (SPSO), and firefly algorithm.

In this study, the ant colony optimization (ACO) technique was proposed to tune the PID controllers in a three-area interconnected power system [38, 38, 42, 43]. The ACO algorithm was introduced by Dorigo and his colleague in the early 1990s to solve many combinatorial optimization problems. The natural behavior of real ants has inspired many researchers to develop the ACO algorithm. During the food searching process of real ants, initially all the ants are spread randomly around the nest searching good-quality food. As the ants return, they store a pheromone chemical on the ground based on the quality and quantity of the food and food source, the shortest path and the paths to the best quality of food having higher concentration of pheromones, enabling the other ants to follow the optimal path. In the ACO algorithm, there are three main phases: initialization, constructing the ant solution, and pheromone updating. The flow of the ACO algorithm for the PID controller is as follows:

ACO algorithm for the PID controller:

```
Step 1:  Start simulation
Step 2:  Parameter initialization
            No of ants
            Pheromone
            Evaporation parameter
            Number of iteration
Step 3:  Run the process model
Step 4:  Evaluate the objective function
Step 5:  Update pheromone and probability
Step 6:  calculate optimum PID controller parameters
Step 7:  Check whether maximum number of iterations has been reached
Step 8:  If YES: Stop the simulation process
Step 9:  if NO: Go to step 3 and repeat the procedure.
```

The ACO optimized PID controller gain values with different scenarios are given in Table 2. An interest point has locality in space with no spatial extent. The existence of interest points can significantly reduce the required computation time. A foremost interest point detector is a contour curvature based technique. Typically, these were relevant to piecewise constant regions, line drawings, extracting geomet-

Table 2 ACO optimized PID controller parameters for different scenarios

Criteria	Kp1	Ki1	Kd1	Kp2	K21	Kd2	Kp3	Ki3	Kd3
Without nonlinearity and BD	0.95	1	0.24	0.75	1	0.53	0.91	0.94	0.72
With GRC nonlinearity	1	1	0.29	0.77	0.97	0.66	0.97	0.87	0.49
With GDB nonlinearity	0.36	0.88	0.96	0.29	0.75	0.55	0.15	0.03	0.94
With GRC and GDB nonlinearity	0.16	0.81	0.76	0.07	1	0.12	0.49	0.02	0.97
With GRC, GDB nonlinearity and BD	0.6	0.92	0.96	0.58	0.92	0.77	0.29	0.36	0.98

rically important corners [11], etc. The following sections will discuss these different feature detectors with regard to concept and applications.

4 Simulation Results and Discussion

4.1 Simulation Results

Simulations of the investigated power system were carried out using a 4TH GEN INTEL, Core i3–4GB RAM-1TB HDD-WINDOWS (8.1) Laptop in the MATLAB 7.5.0.342 (R2007b) environment. The transfer function model of the investigated power system in this study was designed and developed under the MATLAB/SIMULINK environment. The ant colony optimization (ACO) is written in a separate file (a . m file). Initially, the PID controller parameters are optimized without considering the effect of nonlinearity (generation rate constraint (GRC) and governor dead band (GDB)) and boiler dynamics with two percent (2% SLP) step load perturbation in area 1. The objective function is calculated up to 180 s. Afterward, the PID controller parameters are optimized with different scenarios, and controller gain values of different scenarios are shown in Table 2.

Scenario A:

Consider the GRC in all areas and two-percent step load perturbation (2% SLP) in area 1. In this scenario, GRC nonlinearity is considered in the investigated power system. Figures 5 and 6 show the frequency deviations in area 1 and area 3. Figure 7 illustrates the tie-line power flow deviations in area 2. Figure 8 shows the area control error of area 2. In all figures, the solid lines indicate the response of the system without GRC nonlinearity; the dashed lines show the response of the power system with GRC nonlinearity. The time domain specification values of this scenario are shown in the Table 3.

Fig. 5 Deviation of frequency in area 1

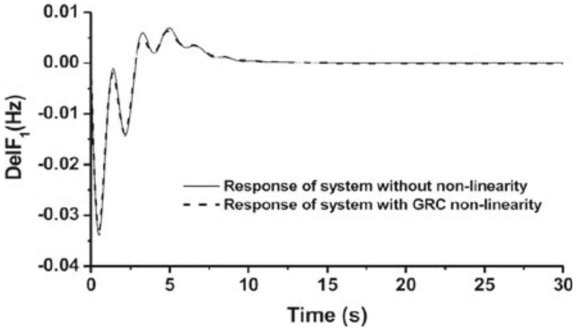


Fig. 6 Deviation of frequency in area 3

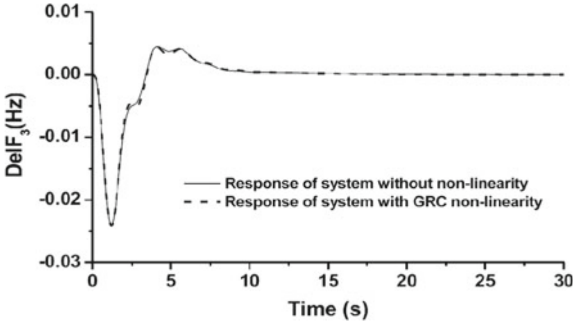


Fig. 7 Deviation of tie-line power in area 2

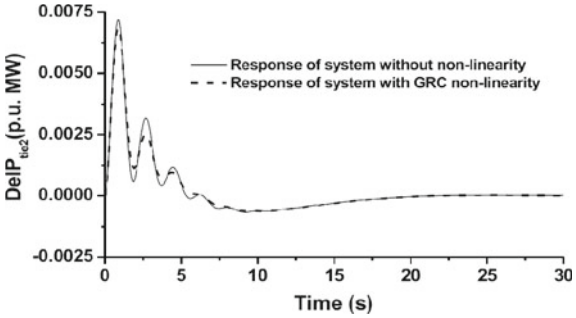


Fig. 8 Deviation of area control error in area 2

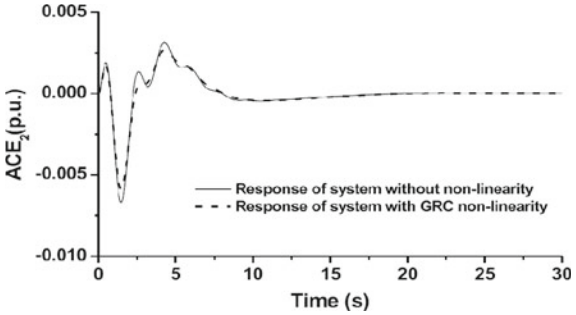
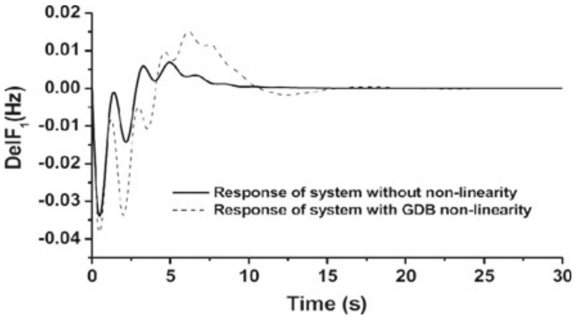


Table 3 Comparisons of time-domain parameters for different scenarios

Criteria	Parameters	delF1	Delptie3	ACE2
Without nonlinearity and BD	Settling time	12.92	19.9	19.15
	Peak overshoot	0.0052	0.007	0.0019
	Peak undershoot	0.033	0.0006	0.0064
GRC nonlinearity	Settling time	13.3	29.55	20.44
	Peak overshoot	0.004	0.0066	0.0014
	Peak undershoot	0.032	0.00061	0.0056
GDB nonlinearity	Settling time	20.48	22.7	22.66
	Peak overshoot	0.0085	0.00797	0.002
	Peak undershoot	0.037	0.0013	0.011
GRC and GDB nonlinearity	Settling time	17.9	32.36	34.16
	Peak overshoot	0.016	0.0088	0.0021
	Peak undershoot	0.039	0.0006	0.015
GRC, GDB nonlinearity and BD	Settling time	23.88	33.7	22.58
	Peak overshoot	0.011	0.0079	0.0021
	Peak undershoot	0.037	0.00048	0.011

Fig. 9 Deviation of frequency in area 1



It has been established from the previous figures that adding nonlinearity into the power system leads to increasing the settling time response compared to the system response without GRC nonlinearity.

Scenario B:

We consider GDB in all areas and two-percent step load perturbation (2% SLP) in area 1. In this scenario, the governor dead band (GDB) nonlinearity is added to the power system. Frequency deviations in area 1 and area 3, tie-line power flow deviations in area 2 and area control error of area 2 are shown in Figs. 9, 10, 11 and 12, respectively. In the response figures, the solid lines show the response of the power system without nonlinearity, while the dashed lines show the response of the power system with nonlinearity. It is observed from the response that adding the nonlinearity to the power system yields more damping oscillations with large

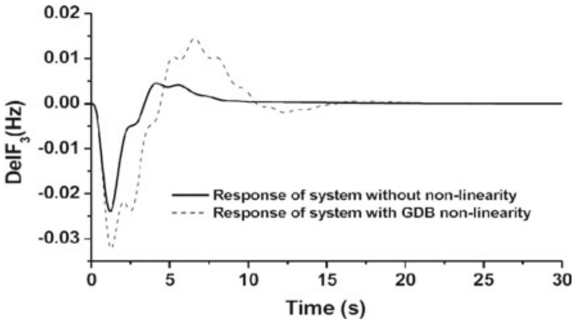


Fig. 10 Deviation of frequency in area 3

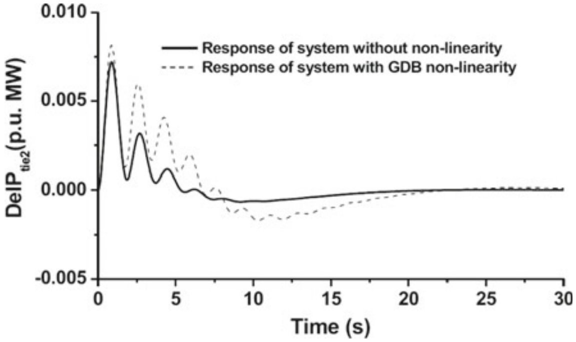


Fig. 11 Deviation of tie-line power in area 2

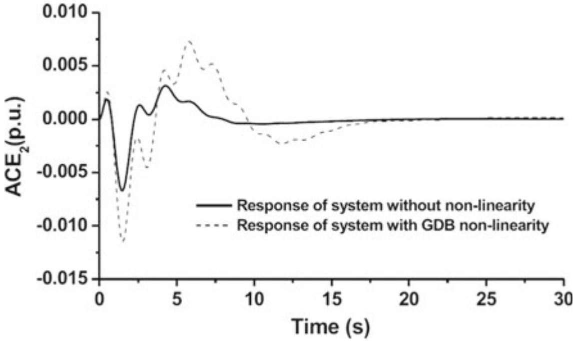


Fig. 12 Deviation of area control error in area 2

peak overshoots/undershoots with large settling time. The numerical time-domain specification values are given in Table 3.

Scenario C:

We consider GRC and GDB in all areas and a two-percent step load perturbation (2% SLP) in area 1. In this scenario, the generation rate constraint and governor dead band nonlinearity are considered in all three areas of the investigated power system. The response of power systems are shown in Figs. 13, 14, 15 and 16. The solid lines show the response of the power system without any nonlinearity, and the dashed

Fig. 13 Deviation of frequency in area 1

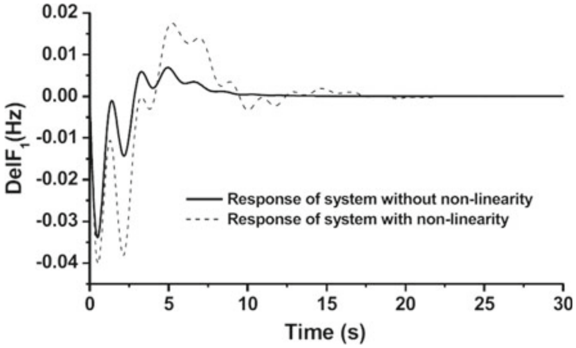


Fig. 14 Deviation of frequency in area 3

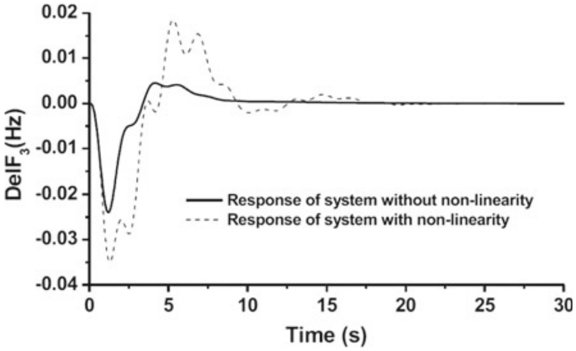
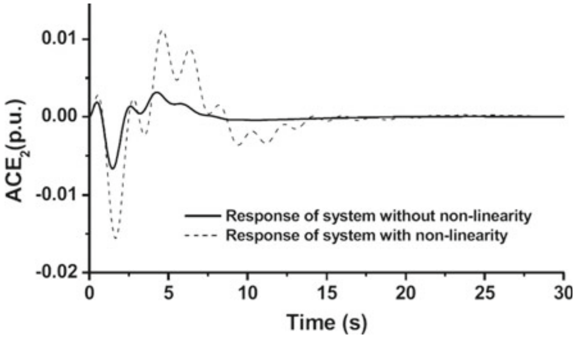


Fig. 15 Deviation of area control error in area 2



lines show the response of the power system with GRC and GDB nonlinearity. It is established that for scenario C, adding nonlinearity to the power system results in more damping oscillations with large peak over- and undershoots approximately up to 20 s, compared to the system response without nonlinearity. The numerical values of the time domain specification are given in Table 3.

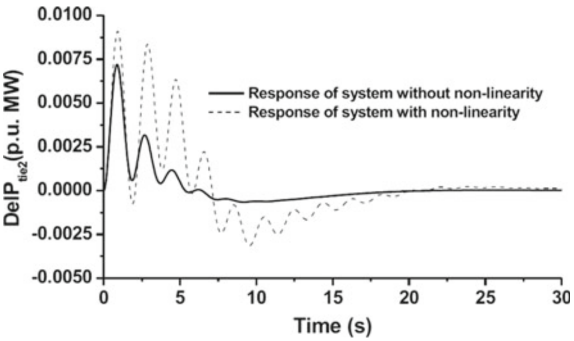


Fig. 16 Deviation of tie-line power in area 2

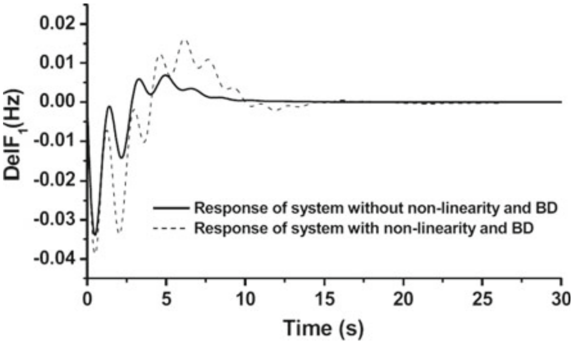


Fig. 17 Deviation of frequency in area 1

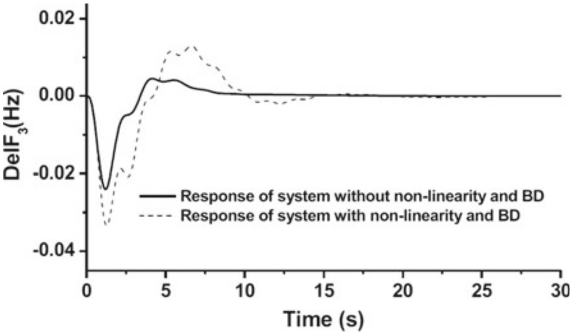


Fig. 18 Deviation of frequency in area 3

Scenario D:

We consider GRC, GDB, and boiler dynamics (BD) in all areas and two-percent step load perturbation (2% SLP) in area 1.

In this scenario, the GRC, GDB nonlinearity and boiler dynamics are considered in the investigated multiarea thermal power system. Frequency deviations in areas 1 and 3, tie-line power flow deviation in area 2, and area control error of area 2 are shown in Figs. 17, 18, 19 and 20, respectively. The solid lines show the response of the power system without nonlinearity, and the dashed lines show the response of the

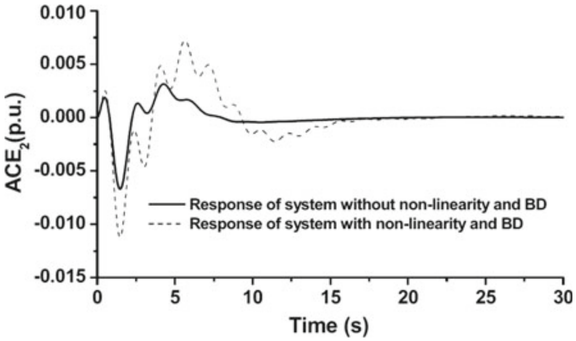


Fig. 19 Deviation of area control error in area 2

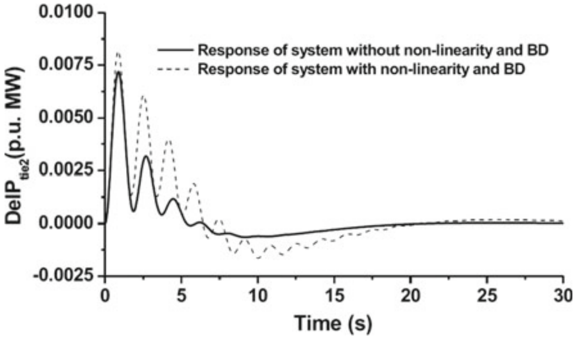


Fig. 20 Deviation of tie-line power in area 2

power system response with GRC, GDB, and BD in all areas. From the response of scenario D, it is evident that the power system responses with GRC, GDB, and BD yield more damping oscillations with large peak overshoot and undershoot compared to the system without GRC, GDB, and BD in the investigated power system. The values of settling time and peak overshoot and undershoot values are given in Table 3.

4.2 Discussion

The performance of a multiarea thermal power system is demonstrated in Sect. 4.1, while the results are discussed in Sect. 4.2. The values of settling time, peak overshoot, and peak undershoot for different scenarios (Scenario A, Scenario B, Scenario C, and Scenario D) are given in Table 3.

Table 3 shows that the proposed system without nonlinearity and BD outperforms all the other cases in terms of achieving the minimum settling time, while considering GRC nonlinearity achieves superior performance in terms of the minimum peak overshoot and undershoot, while without nonlinearity and BD, it is superior to the remaining cases.

Figure 21 shows a comparison of the settling times for the different scenarios.

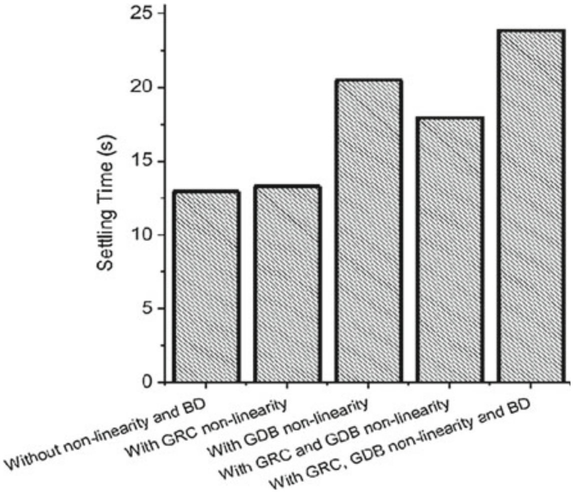


Fig. 21 Comparisons of settling times in different scenarios

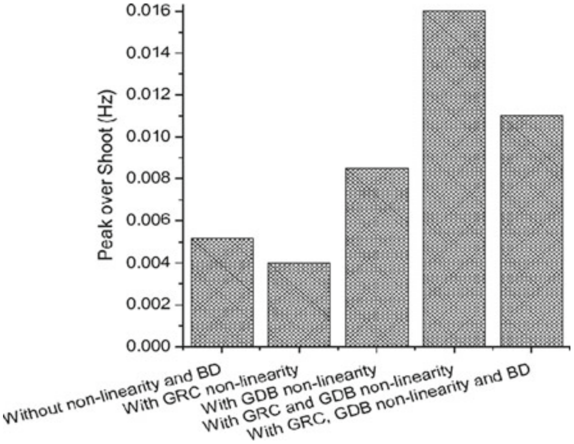


Fig. 22 Comparisons of peak overshoot in different scenarios

From Fig. 21 it is evident that with nonlinearity and boiler dynamics implemented into the power system, the performance of the system is affected and takes more time to settle compared to the performance of the system without any nonlinearity and boiler dynamics. The performance of the system with both nonlinearity and BD takes more settling time compared to all remaining scenarios. Figures 22 and 23 show the peak overshoot and undershoot values of the investigated power system in the different scenarios.

Figures 22 and 23 clearly demonstrate that system performance depends on the parameters of the power system's components. Adding any nonlinear parameters to

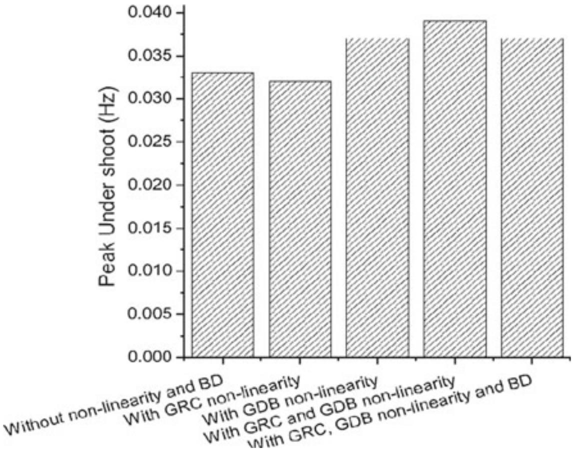


Fig. 23 Comparisons of peak undershoot in different scenarios

the power system affects the system response, and the response exhibits more peak over- and undershoots in the damping oscillations. When GRC and GDB nonlinearity is considered in the system, that leads to more peak shoots in the response compared to all other scenarios.

5 Conclusion

In this study, the load frequency control (LFC) of three-area interconnected reheat thermal power systems is considered with a proportional-integral-derivative (PID) controller. The optimal values of the PID controller are obtained using an artificial-intelligence-based ant colony optimization (ACO) technique. The performance of the LFC of an interconnected power system was studied by adding nonlinearity and boiler dynamics. Different scenarios are tested: (i) scenario A: GRC nonlinearity is considered in the investigated power system; (ii) scenario B: GDB nonlinearity is considered in the power system; (iii) scenario C: both GRC and GDB nonlinearities are considered in the power system; (iv) scenario D: GRC and GDB nonlinearities and boiler dynamics are considered in the power system.

The simulation result for the different scenarios established that when GRC nonlinearity is added to the power system (scenario A), it takes more time to settle compared to the system response without nonlinearity, while in scenario B, when GDB nonlinearity is added to the system, it produces more damping oscillations with large peak shoots and large settling time compared to a system without nonlinearity. In addition, when both nonlinearities are considered (scenario C), the system exhibits more damping oscillations with peak shoots and takes more settling time compared to systems without any nonlinearity. Finally, when nonlinearity and BD are added to the power system (scenario D), the system exhibits more damping oscillations with peak shoots with more settling time compared to system response without any nonlinearity.

Generally, it is observed that nonlinearity and boiler dynamics produce a significant effect in the response of the power system under investigation (it produces more damping oscillations, large peak shoots, and large settling time).

References

1. Saikia, L.C., Sinha, N., Nanda, J.: Maiden application of bacterial foraging based fuzzy IDD controller in AGC of a multi-area hydrothermal system. Electr. Power Energy Syst. **45**, 98–106 (2013)
2. Debbarma, S., Saikia, L.C., Sinha, N.: AGC of a multi-area thermal system under deregulated environment using a non-integer controller. Electr. Power Syst. Res. **95**, 175–183 (2013)
3. Sahu, R.K., Panda, S., Rout, U.K.: DE optimized parallel 2-DOF PID controller for load frequency control of power system with governor dead-band nonlinearity. Electr. Power Energy Syst. **49**, 19–33 (2013)
4. Shabani, H., Vahidi, B., Ebrahimpour, M.: A robust PID controller based on Imperialist competitive algorithm for load frequency control of power systems. ISA Trans. **52**, 88–95 (2013)
5. Ramesh Kumar, S., Ganapathy, S.: Cuckoo search optimization algorithm based load frequency control of interconnected power systems with GDB nonlinearity and SMES units. Intern. J. Eng. Innov. **2**(12), 23–28 (2013)
6. Subha, S.: Load frequency control with fuzzy logic controller considering governor dead band and generation rate constraint non-linearities. World Appl. Sci. J. **29**(8), 1059–1066 (2014)
7. Nanda, J., Sakkaram, J S.: Automatic generation control with fuzzy logic controller considering generation rate constraint. In: Proceedings of the 6th International Conference on Advances in Power System Control, Operation and Management, APSCOM 2003, Hong Kong, pp. 770–775 (2003)
8. Swain, A.K., Mohanty, A.K.: Adaptive load frequency control of an interconnected hydro thermal system considering generation rate constraint. J. Inst. Eng. (India) **76**, 109–114 (1995)
9. Parida, M., Nandha, J.: Automatic generation control of a hydro-thermal system in deregulated environment. In: Proceedings of the Eighth International Conference, Electrical Machines and Systems, 2005, ICEMS 2005, vol. 2, pp. 942–947 (2005)
10. Demiroren, A., Sengor, N.S., Lale Zeynelghi, H.: Automatic generation control by using ANN technique. Electr. Power Compon. Syst. **29**, 883–896 (2001)
11. Ngamroo, I.: Robust decentralized frequency stabilizers design for SMES taking into consideration system uncertainties. Electr. Power Energy Syst. **74**, 281–292 (2005)
12. Concordia, C., Kirchmayer, L.K., Szymanski, E.A.: Effect of speed-governor dead band on tie-line power and frequency control performance. IEEE Trans. Power Appar. Syst. **76**(3), 429–434 (1957)
13. Shayeghi, H., Shayanfor, H.A.: Application of ANN technique based μ-synthesis to load frequency control of interconnected power and energy systems. Electr. Power Energy Syst. **28**, 503–511 (2006)
14. Tripathy, S.C., Balasubramaniam, R., Chandramohan Nair, P.S.: Effect of superconducting magnetic energy storage on automatic generation control considering governor dead and boiler dynamics. IEEE Trans. Power Syst. **7**(3), 1266–1273 (1992)
15. Tripathy, S.C., Hope, G.S., Malik, O.P.: Optimization of load-frequency control parameters for power systems with reheat steam turbines and governor dead band nonlinearity. IEE Proc. **129**(1), 10–16 (1982)
16. Nanda, J., Kothari, M.L., Satsangi, P.S.: Automatic generation control of an interconnected hydrothermal system in continuous and discrete modes considering generation rate constraints. IEE Proc. **130**(1), 17–27 (1983)

17. Hari, L., Kothari, M.L., Nanda, J.: Optimum selection of speed regulation parameters for automatic generation control in discrete mode considering generation rate constrains. IEE Proc. C **138**(4), 401–406 (1991)
18. Lu, C.-F., Liu, C.-C., Wu, C.-J.: Effect of battery energy storage system on load frequency control considering governordead band and generation rate constraint. IEEE Trans. Energy Convers. **10**(3), 555–561 (1995)
19. Pan, C.T., Liaw, C.M.: An adaptive controller for power system load-frequency control. IEEE Trans. Power Syst. **4**(1), 122–128 (1989)
20. Nanda, J., Mishra, S., Sailkia, L.C.: Maiden application of bacterial foraging based optimization technique in multi area automatic generation control. IEEE Trans. Power Syst. **24**(2), 602–609 (2009)
21. Nanda, J., Mangla, A., Suri, S.: Some new findings on automatic generation control of an interconnected hydrothermal system with conventional controllers. IEEE Trans. Energy Convers. **21**(1), 187–194 (2006)
22. Anand, B., Jeyakumar, E.: Fuzzy logic load frequency control of hydro-thermal system with non-linearities. Int. J. Elec. Power Eng. **3**(2), 112–118 (2009)
23. Nanda, J., Saikia, L.C.: Comparison of performances of several types of classical controller in automatic generation control for an interconnected multi-area thermal system. In: Proceedings of the 2008 Australasian Universities Power Engineering Conferenc (AUPEC08), pp. 1–6 (2008)
24. Tripathy, S.C., Bhatti, T.S., Jha, C.S., Malik, O.P., Hope, G.S.: Sampled data automatic generation control analysis with reheat steam turbines and governor dead-band effects. IEE Trans. Power Appar. Syst. **103**(5), 1045–1051 (1984)
25. Saikia, L.C., Bharali, A., Diixit, O., Malakar, T., Sharma, B., Kouli, S.: Automatic generation control of multi-area hydro system using classical controllers. In: 1st International Conference on Power and Energy in NERIST (ICPEN), Nirjuli, pp. 1–6, 28–29 December 2012
26. Chidambaram, I.A., Velusami, S.: Decentralize biased controllers for load frequency control of interconnected power systems considering governor dead band non-linearity. In: IEEE Indicon 2005, Chennai, India, pp. 521–525, 11–13 Dec 2005
27. Chaine, S., Tripathy, M.: Design of an optimal SMES for automatic generation control of two-area thermal power system using cuckoo search algorithm. J. Electr. Syst. Inf. Technol., xx: xxx–xxx (2015) (In Press)
28. Appl. Softw. Comput. Fractional-orer load frequency control of interconnected power systems using chaotic multi-objective optimization. **29**, 328–344 (2015)
29. Dash, P., Saikia, L.C., Sinha, N.: Automatic generation control of multi area thermal system using bat algorithm optimized PD-PID cascade controller. Electr. Power Energy Syst. **68**, 364–372 (2015)
30. Zare, K., Hagh, M.T., Morsali, J.: Effective oscillation damping of an interconnected multi-source power system with automatic generation control and TCSC. Electr. Power Energy Syst. **65**, 220–230 (2015)
31. Sahu, R.K., Panda, S., Padhan, S.: A hybrid firefy algorithm and pattern search technique for automatic generation control of multi area power systems. Electr. Power Energy Syst. **64**, 9–23 (2015)
32. Padhan, S., Sahu, R.K., Panda, S.: Application of firefly algorithm for load frequency control of multi-area interconnected power system. Electr. Power Compon. Syst. **42**(13), 1419–1430 (2014)
33. Dash, P., Saikia, L.C., Sinha, N.: Comparison of performances of several Cuckoo search algorithm based 2DOF controllers in AGC of multi-area thermal system. Electr. Power Energy Syst. **55**, 429–436 (2014)
34. Naidu, K., Mokhlis, H., Bakar, A.H.A., Terzija, V., Ilias, H.A.: Application of firefly algorithm with online wavelet filter in automatic generation control of an interconnected power reheat thermal power syste. Electr. Power Energy Syst. **63**, 401–413 (2014)
35. Padhan, S., Sahu, R.K., Panda, S.: Automatic generation control with thyristor controlled series compensator including superconducting magnetic energy storage units. Ain Shams Eng. J. **5**, 759–774 (2014)

36. Kothari, M.L., Satsangi, P., Nanda, J.: Sampled-data automatic generation control of interconnected reheat thermal systems considering generation rate constraint. IEEE Trans. Power Appar. Syst. **100**(5), 2334–2342 (1981)
37. Jagatheesan, K., Anand, B., Ebrahim, M.A.: Stochastic particle swarm optimization for tuning of PID controller in load frequency control of single area reheat thermal power system. Int. J. Electr. Power Eng. **8**(2), 33–40 (2014)
38. Jagatheesan, K., Anand, B.: Automatic generation control of three area Hydro-Thermal power systems considering electric and mechanical governor with conventional and ant colony optimization technique. Adv. Nat. Appl. Sci. **8**(20), 25–33 (2014)
39. Jagatheesan, K., Anand, B., Dey, N.: Automatic generation control of Thermal-Thermal-Hydro power systems with PID controller using ant colony optimization. Int. J. Serv. Sci. Manag. Eng. Technol. **6**(2), 18–34 (2015)
40. Dey, N., Samanta, S., Yang, X.S., Chaudhri, S.S., Das, A.: Optimization of scaling factors in electrocardiogram signal watermarking using Cuckoo search. Int. J. Bio-Inspired Comput. (IJBIC) **5**(5), 315–326 (2014)
41. Anand, B., Ebenezer, J.A.: Load frequency control with fuzzy logic controller considering Non-Linearities and boiler dynamics. ACSE **8**, 15–20 (2012)
42. Jagatheesan, K., Anand, B.: Dynamic performance of multi-area hydro thermal power systems with integral controller considering various performance indices methods. In: Proceedings of IEEE International Conference on Emerging Trends in Science, Engineering and Technology, pp. 474–478 (2012)
43. Jagatheesan, K., Jeyanthi, S., Anand, B.: Conventional load frequency control of an interconnected multi-area reheat thermal power systems using HVDC link. Int. J. Sci. Eng. Res. **5**(5), 88–92 (2014)

Part III
Transportation

A Review of Traffic Light Control Systems and Introduction of a Control Concept Based on Coupled Nonlinear Oscillators

Jean Chamberlain Chedjou and Kyandoghere Kyamakya

Abstract This chapter provides an in-depth overview of the state of the art on traffic light control and optimization. Several classical control systems, methods, concepts, tools, and strategies are described, and their related pros and cons are discussed. Four different types of control strategies are considered, pretimed, actuated, adaptive, and self-organized, and it is demonstrated that some strategies can be appropriate for local control, area control, or both (local and area). Further, the chapter develops a system of coupled nonlinear oscillators, that is used for traffic light control and optimization both at isolated junctions (i.e., local control) and in a network of coupled traffic junctions (i.e., area control). The case of an isolated traffic junction is modeled by coupled oscillators, each of which represents a specific phase group of the traffic light at the junction. The case of a network of coupled traffic junctions is also considered, and we show the possibility of modeling the traffic light at each junction by a single oscillator. The system developed is viewed as a modified version of the self-organized Kuramoto model for traffic light control due to some important features that are common to both systems (i.e., Kuramoto model and the system developed). The main advantage of the system developed is the possibility of dynamically monitoring the signals' phases (delays), signal-splits, or signal timings according to the dynamic variation of the traffic demand in all conflicting approaches of the traffic junctions under investigation. The system of coupled nonlinear oscillators is considered a flexible platform, which is appropriate for modeling the four types of traffic light control strategies. Another advantage of the system developed is the possibility of an easy hardware implementation using electronic devices. This allows a straightforward possibility for designing appropriate electronic prototypes of traffic light controllers (appropriate for both local and area controls).

J.C. Chedjou (✉) · K. Kyamakya
Transportation Informatics Group (TIG) Institute of Smart Systems Technologies,
Alpen-Adria University of Klagenfurt, 9020 Klagenfurt, Austria
e-mail: jean.chedjou@aau.at

K. Kyamakya
e-mail: kyandoghere.kyamakya@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_6

Keywords Modeling · Simulation · Optimization · Nonlinear oscillators · Coupled oscillators · Isolated traffic junctions · Coupled traffic junctions · Pretimed control · Actuated control · Adaptive control · Self-organized control · Signal timings · Phase groups · Green/red signals · Kuramoto model

1 Introduction

1.1 Motivation for Developing New Traffic Light Control Systems and Concepts

The growing population density is a major factor explaining the increasing number of traffic participants, and this results in some well-known endemic traffic problems (e.g., congestion, accidents, and pollution) if the structures of roads are not modified. Nevertheless, it is very difficult and costly to modify road structures; it can also be impossible to construct new roads or to reconstruct a deficient street network. These facts justify the tremendous attention devoted to the development of modern (i.e., efficient, robust, and high-performance) traffic control systems and concepts, since such systems are good alternatives for overcoming well-known endemic traffic problems [1].

Traffic signals at road junctions are generally used to optimize the junction's throughput through an efficient management of conflicting traffic movements. This leads to safety improvements and also congestion avoidance, less energy consumption, and less pollution, to name just a few factors. Generally, the management of conflicting movements is performed through the signal-splitting process (or green signal sharing within a cycle time), whereby individual traffic movements are assigned portions of green signals. In order to optimize the control strategy, several nonconflicting movements are grouped into a single movement and are further assigned a common "traffic signal," called a "phase group." It should be noted that a "phase group" is the fundamental unit of any traffic control scheme [1].

The tremendous attention devoted to the development of new traffic light control systems in recent decades has resulted in the design and implementation of multiple traffic control strategies such as pretimed, actuated, adaptive, and self-organized. However, due to the complex characteristics of the traffic dynamics (e.g., high nonlinearity, stochasticity, extreme sensitivity to perturbations, chaos, unpredictability) previously developed systems, concepts, and tools do not satisfactorily capture all insights into the complex dynamics of traffic flows at the microscopic, macroscopic, and mesoscopic level of detail. Some limitations of most system concepts are low accuracy, lack of robustness, lack of adaptivity, very time-consuming, not real-time, lack of proactivity, do not accommodate the dynamic change of traffic demand. It has been observed that existing traffic control systems, concepts, and tools appear complementary, since each can solve a problem not solved by its counterpart. This statement clearly justifies the importance of exploring further research issues with the specific aim of developing new systems, concepts, or paradigms for traffic light control at junctions.

1.2 Traffic Light Control Strategies: Classification and Description

Traffic signals at junctions are generally used to optimize the junction's throughput through an efficient management of conflicting traffic movements. Generally, the management of conflicting movements is performed through the signal splitting process (or green signal sharing within a cycle time), whereby individual traffic movements are assigned portions of the total green time available within a cycle.

The spatial extent of traffic light control is classified into (a) local control and (b) area control. Local control is applied in the case of isolated traffic junctions, while area control is used in a network of coupled traffic junctions (also called nonisolated traffic junctions). Further, local control encompasses (a) pretimed, (b) actuated, and (c) adaptive control strategies. Area control involves (a) pretimed, (b) actuated, (c) adaptive, and (d) self-organized control strategies.

Pretimed control: This represents a fixed-time control; it does not depend on the current real traffic demand. At every time of the day and day of the week, a precomputed timing plan is selected from a set of plans. These plans are developed offline based on historic traffic data, the so-called AADT (average annual daily traffic). AADT is the average calculated over a year of the number of vehicles passing a point in a given counting section each day (usually expressed in vehicles per day or hour of the day), with different day types incorporated in the timing plan and interpretation.

Actuated control: This is a detection-actuated control; it depends on the traffic demand. Traffic detectors indicating the presence or absence of vehicles are used to adapt the control scheme (generally a fixed-time one) to fluctuations in traffic demand. If the gap between vehicles is larger than some maximum gap, the controller may decide to stop the green phase.

Adaptive control: The main difference between actuated control and adaptive control is that adaptive control first verifies (or considers) the state of the complete intersection (i.e., all phase groups of the junction) before deciding to which phase group priority should be assigned. Generally, the metric of decision is the current traffic demand in all phase groups (arrivals in phase groups). In essence, the effective traffic load continuously sensed by detectors is used to continuously optimize the signal plans. However, adaptive control processes a huge amount of data and is therefore very time demanding (i.e., very slow) and thus cannot be used efficiently for real-time applications.

Self-organized control: Traditionally, artificial self-organized systems use internal (or local) interactions among a large number of agents in order to solve control problems [2]. Artificial self-organizing systems are commonly inspired by natural systems that are regulated by their intrinsic internal processes. Examples of natural systems are swarm behaviors of insects, fishes, and birds [3]. In essence, systems of self-organizing traffic signals use a decentralized optimization principle that allows global coordination of the traffic flows in a road network. Using the decentralized principle, traffic signals at each intersection are controlled by an agent. This agent makes a decision that does not depend on real-time traffic data. This means that the agent makes a decision based on the traffic conditions that exist at road segments connected to an intersection. The self-organizing strategy is very flexible with respect to local traffic demands and more robust to dynamic changes of the traffic flows, as compared to the adaptive strategy. Further, traffic control systems based on a self-organizing strategy are likely to respond to actual real-time traffic conditions without using a predetermined signalization schedule. These schedules are generally based on average traffic characteristics (e.g., average speed, average flow, average density) [4]. in essence, the self-organized control strategy overcomes the drawbacks of adaptive control, since it is an online processing strategy. The self-organized control strategy does not require a huge amount of data and therefore executes fast (or real-time) processing. This technique can be considered an alternative that may help to overcome some limitations of standard central as well as decentralized controllers and make an adaptive and purely demand-driven traffic light control practically applicable.

1.3 Commonly Used Metrics for Traffic Light Control and Optimization

The general theory of traffic signal control based on an optimization procedure can be developed by providing a clear definition of some performance metrics (to be minimized) of junctions such as the total delay D_{lay}, the throughput T_{put}, and the number of stops S_{Top}, which are further considered during the optimization procedure. The total delay is the time experienced by all vehicles crossing the traffic junction, while the throughput expresses the number of vehicles that cross the junction during the green signals. The number of stops corresponds to all vehicles that were unable to cross the junction during the cycle time or cycle length.

The overall optimization procedure exploits analytical expressions (mathematical formulas) of the traffic performance metrics (D_{lay}, T_{put}, and S_{Top}). Related formulas have been proposed in the literature (see Webster (1958) [5], Miller (1968) [6], Hale (TRANSYT-7F, 2006) [7], Akcelik (SIDRA, Australia, 1980, 2011) [8], just to name a few) as functions of the following traffic parameters and state variables: n is the number of phase groups; c is the cycle length (i.e., the cycle time); q_i is the flow demand in a phase group with index i. This demand is measured by magnetic sensors (i.e., inductive loops); $u_i = \phi_i / c$ is the proportion of green for each phase group with index i; ϕ_i corresponds to the effective green signal allocated to the corresponding phase group; $y_i = (q_i/S_i)$ is the flow factor, whereby S_i is the saturation flow (expressed by the formula in Eq. (1) (Kimber et al. 1992) **[312]**; $Q_i = S_i\phi/c$ is the capacity of a lane group with index i; $x_i = (q_i/Q_i)$ is the degree of saturation; $z_i = (x_i - 1)$ is a characteristic coefficient, which is used to depict the traffic states (e.g., undersaturation ($z_i < 0$), at saturation ($z_i = 0$), and oversaturation ($z_i > 0$); T_i is the demand period:

$$S_i = \frac{S_0}{1 + \frac{4.92}{R_i}}, \tag{1}$$

where S_0 corresponds to the nominal value of the saturation flow (this value is approximated in practice to 2100 vehicles/lane) and R_i is the radius of the trajectory followed by vehicles.

1.4 Challenges Faced by Traffic Light Control Systems and Related Limitations

Traffic is a typical example of a complex system [9]. Traffic undergoes nonlinear dynamics [10], time-varying dynamics [11], and stochastic dynamics [12].

Regarding local traffic control at a junction, the traditional approaches are (a) pretimed [13], (b) pseudoadaptive (actuated or semiactuated). The traditional methods are at best "reactive" [14]. Further traditional methods are based on fixed phase group sequences, and none of the traditional methods can efficiently consider all traffic states (i.e., undersaturation, at saturation, and oversaturation) [11]. Finally, there is a crucial lack of comprehensive benchmarking between the traditional control concepts [1].

Further, the traditional traffic control schemes are not fully adaptive; they cannot perform efficient control under real-time constraints [15]; they are not proactive [11]; and they cannot cope with the imperfection of sensors. Further, the traditional controllers do not consider the dynamic sequencing of signal-timing, and do not consider dynamic cycle times [11]. Finally, the traditional traffic controllers cannot ensure good performance at all traffic states (i.e., undersaturation, saturation, and oversaturation) [11], and they do not enforce the "fairness" principle (see [11, 16]).

The main objective of this work regarding traffic control is to contribute to the enrichment of the current state of the art by developing modern traffic control methods (or concepts) that can efficiently address and solve the above-mentioned limitations of the traditional traffic control concepts. Specifically, the following performance metrics are at stake:

Performance:

- Precision, fairness, speed (or real-time operation), high performance in all system states (undersaturation, saturation, and oversaturation), simulative projections.

Robustness:

- Proactive intelligence involving an adaptive (or dynamic) black box traffic model forecasting capability;
- Adaptivity to system-model time variations, for example due to weather and other external events;
- Adaptive sequencing of signal phase groups;
- Low sensitivity to sensors' imperfections.

1.5 Contribution and Organization

The main objective of the work presented (in this chapter) is twofold: first, to provide an overview of the classical traffic light control schemes. Different types of control strategies are considered and their pros and cons are discussed. Some inherent challenges faced by the traffic light control systems are addressed, and the possibility of overcoming their related limitations is analyzed. In essence, useful information is collected (from the state of the art) and is systematically classified and presented as a summary of research achievements in the field of traffic light control.

The second objective consists in using the nonlinear oscillatory paradigm to develop a new concept for traffic light control and optimization. The strong features (or characteristics) of the concept developed are as follows:

- Flexibility: Modeling of both isolated junctions and coupled junctions.
- Adaptivity: real-time sensing of the time-varying traffic demand (i.e., temporal variation of traffic flow) and optimization of traffic signal timings at junctions accordingly.
- Implementability: The concept developed is easily implementable in hardware and thus could be used to manufacture concrete electronic prototypes of traffic light controllers.

The remainder of this chapter is organized as follows. Section 2 presents a general overview of the traditional traffic light control systems. Several pioneering contributions are described, and their related pros and cons are discussed. Section 3 introduces a new concept for traffic light control and optimization. The concept developed exploits the paradigm of coupled nonlinear oscillators, and the resulting mathematical model is expressed in the form of coupled nonlinear ordinary differential equations (ODEs). The parameters (or coefficients) of the resulting ODEs are expressed in terms of both the fundamental parameters of traffic flow (i.e., flow, speed, and density) and the traffic junctions' parameters (e.g., geometry of the junction, number of phase groups, type of control, dynamics of overflow queue, saturation flow factor, degree of saturation). Finally, the coupled nonlinear ODEs are used to optimize the throughput of traffic junctions. The last section (Sect. 4) is devoted to concluding remarks and outlooks. Further, challenging and unsolved research issues are discussed.

2 Review of Traffic Light Control Systems: Related Works

2.1 State-of-the-Art of Traffic Light Control

The past decades have witnessed a tremendous focus on the development of methods, concepts, paradigms, algorithms, and simulation tools for traffic light control (local and area control). Several contributions have been published so far: cellular automata [17–22], Petri nets [17–22], multiagent systems [4, 23–31], analytical concepts [1, 5, 6], coupled oscillatory paradigm [30, 32–34], simulation tools [7, 8, 11, 14, 35, 36], just to name a few.

Deterministic and stochastic cellular automata (CA) models have been intensively developed for traffic simulation in the frame of traffic light control [17]. The deterministic CA models are computationally less time demanding when dealing with traffic light control. However, deterministic CA models do not present (or provide) a good trade-off between accuracy and computing time [17]. In fact, the computing time increases with increasing accuracy and vice versa [17].

Traffic light control/optimization was also investigated using deterministic and/or stochastic cellular automata (CA) models [22, 37–40]. Reference [22] develops a CA model for traffic light control in a network of coupled traffic junctions. The CA model combines the Nagel–Schreckenberg model for highway traffic [37] with the Biham–Middleton–Levine (BML) model for city traffic [22, 39]. Using the synchronization strategy of traffic signals at junctions, it is demonstrated that the cycle length (or cycle time) of junctions significantly affects the capacity of the network of coupled traffic junctions. In essence, the CA model in [22] calculates the optimal parameters of a traffic signal coordination plan that maximize the flow in a road network.

Reference [38] develops both deterministic and stochastic CA models of a signal-controlled traffic stream. Rules of the CA models are defined in terms of saturation flow, traffic composition, turning movement, free flow speed, geometry of the junction, etc. The stochastic CA model uses the randomization rule of the NaSch model [37]. Further modifications of the randomization rule are envisaged in order to obtain a more realistic CA model. The parameters of the CA models are expressed in terms of the saturation flow rates at a simulated intersection, and thus the strong dependence between the parameters of the CA models and the saturation rates of simulated junctions is revealed. Specifically, the stochastic CA model allows any value of the saturation flow to be obtained by adjusting a deceleration probability parameter.

Reference [40] develops a dynamic traffic signal control strategy based on the cell-transmission model (CTM). The control strategy is based on the optimization of the delays of traffic junctions. This is achieved by transforming the CTM to a set of mixed-integer linear constraints. The transformed CTM is further used for optimizing the dynamic signal control using the delays experienced by traffic junctions as a key metric. This implies an optimal choice of the signal phases such that the delays at junctions are minimized.

Further interesting contributions encompass the work in [41] that develops a deterministic CA model of city traffic. The CA model is used to simulate large road networks of coupled traffic junctions. Another contribution can be found in [42]. Here, rings of cells are used as a simplified representation of intersections in order to simulate a road network of coupled traffic junctions.

Contributions have been published showing the application of both deterministic and stochastic Petri net (PN) models for traffic light control [43–58]. In general, Petri nets have been developed in the form of coupled algebraic mathematical models for traffic light control [43]. The algebraic mathematical model is suitable for the analysis of deterministic and stochastic system states, continuous and discrete states, and distributed and parallel states [43].

Deterministic-time Petri net (DTPN) models for local traffic light control were proposed by Febbraro and Giglio [44–46]. These models are described by (a) some fundamental parameters (e.g., road links/lanes and intersections, occupancy, turning rate, crossing sections, and signal timing plan) and (b) a key feature (e.g., an integration of DTPN submodels expressed in the form of algebraic states). However, in order to avoid undesirable deadlock states that may occur in applying PNs to some specific case studies of traffic light control, Febbraro and Giglio introduced stochastic-timed PNs (STPNs) [47] as an alternative solution for a suitable modeling of (1) the interarrival time of vehicles in the traffic network, (2) the minimum throughput of a junction, and (3) the minimum time to travel between two neighboring junctions [43]. These good features of stochastic STPN models allow the avoidance of deadlocks. Another good feature of STPNs is the possibility of estimating queues at a traffic junction.

Further significant contributions for microscopic or local traffic light control are (1) the PN-based models combining deterministic-time (D) and stochastic-time (S) Petri net (DSPN) models introduced by Badamchizadeh and Joroughi [48], and Makela et al. [49]. The advantage of the DSPN models is the easy calculation of the queue lengths [49] and the average delays experienced by vehicles at junctions [48]; (2) the colored PN models (CTPN) introduced by Basile et al. [50] and Dotoli et al. [51, 52] as an alternative solution to the “uncolored” PN models. CTPN models can solve some of the unsolved problems by “uncolored” PNs such as (a) the successful modeling of the different attributes of vehicles, (b) modeling of the different vehicles’ behavior and parameters [43]. In essence, CTPN has been proven efficient for modeling vehicle routing and occupancy.

Regarding macroscopic or area traffic light control based PN models, several ideas have been proposed. Kutil and Hanzalek developed a traffic light control model based on constant-speed continuous Petri nets called CCPNs [53]. The proposed model is suitable for both the handling of conflicts in the network of coupled traffic junctions and the simulation of flow rate in traffic junctions [43]. Febbraro et al. proposed a hybrid Petri net model (HPN) for area traffic light control [54, 55]. The HPN model is suitable for the modeling and simulation of traffic flow and queue dynamics in a network of coupled junctions. Further, a modified version of HPN was developed in the form of first-order HPN (FOHPNs) by Dotoli et al. [56]. The FOHPN is suitable for the modeling of both lane cells and intersection/junction cells and thus can clearly describe the full dynamics of the network of coupled traffic junctions. Zhang and Jia [57] combined a colored Petri net (CPN) with a discrete PN to obtain a new system suitable for the macroscopic modeling of a junction. Finally, Wang et al.

[58] developed a stochastic-timed version of Petri nets (STPN) that is suitable for the modeling of traffic flow and traffic control at signalized junctions. The proposed STPN model is suitable for the evaluation of the performance criteria (or metrics) of traffic junctions such as queue lengths and time delay.

Multiagent techniques/algorithms have been intensively used for traffic light control [4, 23–30]. A.L.C. Bazzan [23] used the multiagent technique to develop a distributed approach for the coordination of traffic signal agents. The proposed control strategy is a decentralized type aiming at overcoming the limitations of the commonly used strategies based on a central traffic-responsive control system, which are very time-consuming and also difficult to implement. The technique based on game theory consists in modeling intersections (in an arterial) in the form of individual agents (or players) taking part in a dynamic process describing the global control policies (or strategies) at traffic junctions. The proposed technique allows a good coordination and cooperation of traffic signal phases at junctions. Arel et al. [25] combined multiagent technology and neural networks to develop a new concept, which is a multiagent system based on reinforcement learning (RL) for area traffic light control. As case study, a network of five coupled traffic junctions is considered in which each junction is controlled by an autonomous intelligent agent. Using a multiagent setting procedure, the RL algorithm [26] is exploited to control the different intersections in the network of coupled traffic junctions. The proposed system aims at ensuring a good scheduling and coordination of traffic signals in the network of coupled traffic junctions through an efficient traffic light control policy. Fundamental or key performance metrics of traffic junctions are optimized such as minimization of the average delay, throughput, and number of stops. De Oliveira and Camponogara [27] developed a framework based on distributed agents for area traffic control. The framework uses the model predictive control (MPC) approach in the network of coupled traffic junctions. Each junction is considered a subsystem and is managed by a specific agent. Each agent senses and controls a specific traffic junction, and the good handling of communication between agents leads to a coordination of the coupled traffic junctions. The overall process is an optimization whereby the global optimum is obtained using the MPC approach. Stefan Laemmer et al. [4], inspired by the observation of self-organized oscillations of pedestrian flows at bottlenecks, developed a self-organization approach for traffic light control. They view the problem as a multiagent problem with interactions between vehicles and traffic lights. Specifically, this approach assumes a priority-based control of traffic lights by the vehicle flows themselves, taking into account short-sighted anticipation of vehicle flows and platoons. Gershenson et al. [30] proposed a self-organized concept using what they call “simple rules” and no direct communication; traffic lights become able to self-organize and adapt to changing traffic conditions. The mathematical foundation of this approach is, however, weak and not clearly formulated in the paper. Further, the authors refer to and cite none of the self-organization-based approaches from the Kuramoto model family; this is, of course, rather strange. Further, no performance comparison to adaptive approaches of the second and third generations (i.e., 2G and 3G) has been conducted. The following contributions involving traffic signal scheduling (area control) are also worth mentioning: The combination of RL with the dynamic programming algorithm [28] to obtain an adaptive traffic signal controller used to dynamically monitor the delays of traffic junctions. A system combining the Markov model and Robertson's platoon dispersion traffic model [29] to obtain a robust framework for the optimization and coordination of traffic lights in a network of coupled junctions.

Considering the traffic signal control as an optimization problem, specific fundamental performance metrics are generally defined (e.g., delay, throughput, number of stops) in order to characterize traffic junctions. Thus, the aim of the optimization is generally to achieve the optimal values of the performance metrics of given junctions.

The technological and societal importance of developing traffic control systems is justified by the huge number of scientific contributions provided by the state of the art.

Adaptive traffic systems are classified into three generations. The first generation (1G) of traffic-adaptive systems comprises traffic-responsive signal control systems. The main limitation of traffic-responsive signal control systems is their poor coordination. The second generation (2G) of traffic-adaptive systems use an online strategy that implements signal timing plans based on real-time surveillance data and predicted values. In practice, the optimization process is repeated every five or ten minutes. Some adaptive traffic control prototypes of the second generation used in practice are: SCATS, SCOOT, and MOTION (method for optimization of traffic signals in online-controlled networks) [35]. Third-generation (3G) systems are similar to the second-generation ones, but differ with respect to the frequency of revising the signal plans. The third generation control allows the parameters of the signal plans to change permanently in terms of the real-time measurements of traffic variables; this enables an "acyclic" operation. The third generation encompasses the following system examples: OPAC (optimized policies for adaptive control) [59], PRODYN (programmation dynamique) [60], RHODES (real-time, hierarchical, optimized, distributed, and effective system) [61], UTOPIA/SPOT (urban traffic optimization by integrated automation / signal progression optimization technology) [62], and TUC (traffic-responsive urban control) [63].

The self-organized principle for traffic light control has also been given a great deal of attention [4, 30, 30, 31, 33, 34, 64]. The most promising way to realize self-organized traffic lights is to use simple models of networks of coupled oscillators (where each oscillator represents a signalized intersection) that are adapted to the specific features and requirements of traffic lights. An ideal starting point for such investigations is the phase oscillator model introduced in 1975 by Yoshiki Kuramoto [33], which later became famous as the Kuramoto model with numerous applications in various scientific disciplines [34]. Some selected additional interesting contributions are proposed by Sekiyama et al. presented in 2000 [31], Nishikawa et al. [34], Laemmer et al. [4, 31], Akbas et al. [30], Gershenson et al. [30]. These contributions are all inspired from the basic principle of the Kuramoto model for traffic light control.

2.2 Kuramoto Model for Traffic Light Control

The Kuramoto oscillatory model has been intensively used for traffic signals control in a network of coupled junctions [30, 33]. The model is based on the self-organization principle (generally referred to as local synchronization). The achievement of self-organization is an insight of the perfect coordination of all traffic signals involved in a network of coupled traffic junctions. The Kuramoto model is a concept that considers the traffic signal at a junction as a control problem (see [30, 33]). This is a robust traffic control scheme that uses traffic demands, cycle times, offsets, and some parameters of the junction as inputs to the traffic control process. The output of this process is expressed in terms of some output system parameters that are obtained under varying conditions. The output system parameters considered by the Kuramoto model are the phases of the traffic signals of a set of coupled junctions. Indeed, the traffic signals of the coupled junctions are assumed to be periodic (i.e., an oscillator) functions with phases (the phases are solutions of the coupled ODEs). The state variables of the ODEs represent the phases of signals belonging to the coupled junctions under investigation [30]. These phases vary in real time according to traffic demand, offsets, and cycle time adjustment [30]. The real-time variation is expressed by the temporal evolution of all phases of the coupled junctions as solutions of the resulting coupled ODEs. The self-organization observed in the Kuramoto oscillator is referred to as internal (or local) synchronization [30]. Indeed, self-organization is observed when the coupled oscillators with different natural frequencies (i.e., different cycles times) are phase locked [30]. The phase-locked process is observed with some phase delays caused by their mutual couplings (interactions) [30]. We have clearly described in two of our published papers (see [9, 65]) the different types of synchronization. We have also provided in published papers the analytical criteria for the depiction of synchronization (or its occurrence). The main strength of the Kuramoto oscillator/model is that of providing a robust self-organizing control scheme that internally or locally adapts (without any external action or contribution) to dynamical demand changes [30].

2.3 Pros and Cons of the Various Traffic Light Control and Optimization Strategies

An isolated (or single) traffic junction is managed by a basic traffic signal system called a “local controller.” Such a system is likely to be used (or to operate) as a pretimed (fixed-time), actuated, or adaptive controller. However, the main limitation of the pretimed controller is the lack of adaptivity to changes in the traffic volume. Another weakness of the pretimed controller is that the signal-splits (i.e., green signals per phase groups, red signals, and the lost time) are fixed; they are assigned a constant duration within a cycle time, which is also generally constant. Regarding the pretimed control strategy, Webster’s formula [5, 66] is most commonly used to calculate the optimal cycle time and the optimal signal duration to be assigned to all phase groups. This calculation is generally conducted by assuming that the arrival flow rates (i.e., the traffic demand in all phase groups) are constants. In essence, Webster's formula minimizes the average delay for all passing vehicles. However, the above-mentioned assumption (under which this formula is used) does not fit with the real traffic dynamics (the traffic dynamics is essentially nonlinear and stochastic). Further, Webster's formula does not consider the traffic dynamics in queues (waiting areas in front of traffic signals). This is a serious limitation, since the traffic dynamics in queues experiences stochasticity. In contrast to the pretimed control, the actuated control strategy takes into account the traffic dynamics in queues, and this leads to the optimization of the green signals assigned to different phase groups. Indeed, all phase groups are equipped with detectors (inductive loops) with the essential task to manage the dynamics in queues. An example (just for illustration) of the management process in queues is that when a given queue is empty, the priority is immediately removed and is sent to queues in conflicting movements (in case they are occupied by vehicles). Regarding the traffic adaptive control strategy, the main difference with the actuated control is that the adaptive control first verifies (or considers) the state of the complete intersection (i.e., all phase groups of the junction) before deciding to which phase group priority should be assigned. Generally, the metric of decision is the traffic demand in phase groups (arrivals in the different phase groups). As already mentioned, the adaptive control strategy processes a huge amount of data and is therefore very time-demanding (i.e., very slow) and thus cannot be efficiently used for real-time applications. Moreover, the optimization algorithm to be thereby executed is complex as well. The self-organized control strategy was therefore proposed as an alternative to the adaptive control strategy, since it is an online process that requires much less data and is therefore suitable for real-time applications. But no serious benchmarking has shown how well the self-organized control strategy performs compared to the adaptive strategy.

Overall, the traffic signal control concept based on the optimization procedure presents the following advantages:

- The mathematical expressions of the delay, throughput, number of stops, etc., are explicit enough, since they are clearly expressed in terms of the traffic parameters as well as the parameters of the traffic junction.
- The mathematical expressions allow a straightforward possibility of performing splitting (signals phases or signal timings) at traffic junctions.
- The mathematical expressions offer the possibility of constructing a white box system model for traffic signals splitting and throughput optimization.
- The mathematical expressions can be used to design a real-time optimizer of traffic signals at the junction.
- The mathematical expressions are suitable for the microscopic analysis and optimization of the junction.

Regarding the limitations of the traffic signal control concept based on the optimization procedure, one can underscore the following:

- The mathematical expressions of the delay, throughput, number of stops, etc., are generally empirical expressions.
- None of these expressions can perform efficiently under all possible traffic states (i.e., undersaturation, at saturation, and oversaturation). They are mostly good enough only for the undersaturation state.
- The mathematical expressions are generally too complex and thus difficult to handle (or solve).
- The optimization algorithms are generally not robust and are very time-consuming as well.

At this point, it is important to mention the lack of comparative benchmarks among the traffic-adaptive systems. Benchmark studies are needed to ascertain the added value of each of the proposed systems with respect to a reference. In most papers reporting on field tests, it is the current system at that time against which it is benchmarked. Since it is possible that the system has not been well maintained, it is hard to judge the new system appropriately. In cases in which simulation is used as a benchmark, it is in principle possible to benchmark against an optimized system. However, several authors are skeptical with regard to the real (or optimum) value of the "optimized systems" used for benchmarking in many reported cases. Moreover, simulation presents only a partial view of the reality. An example (for illustration) is the frequent inaccurate modeling of both side-street parking and pedestrian activity. Nevertheless, a comprehensive simulation framework may be a good and cost-effective way for a systematic benchmarking of the most promising traffic control concepts.

2.4 Description of Commonly Used Traffic Light Control Strategies

- **Local control**

The local control is applied to isolated (or single) traffic junctions. The isolated junctions, also called uncoupled junctions, do not interact with any other junction. Hence, isolated traffic junctions are independent of one another. In other words, the state of an isolated traffic junction does not influence the states of its counterparts (i.e., neighboring junctions) and vice versa. A network of isolated traffic junctions is characterized by the lack of coordination between the traffic junctions.

Regarding the pretimed local control strategy: The pretimed control is commonly modeled by Webster's formula [5, 66]. Webster's formula calculates the optimal cycle time and the optimal signal duration to be assigned to all phase groups. This calculation is generally conducted by assuming that the traffic demand in all phase groups is constant. This assumption does not fit with the real traffic dynamics, which is essentially time-varying and stochastic. Further, Webster's formula does not consider the traffic dynamics in queues (waiting areas in front of traffic signals). This is a serious limitation.

Regarding actuated local control strategy: In contrast, the actuated control strategy takes into account the traffic dynamics in queues, and this leads to the optimization of green signals assigned to different phase groups. Indeed, all phase groups are equipped with detectors (inductive loops) with the essential task of providing information on the traffic arrival process.

Regarding adaptive local control strategy: Regarding the adaptive control strategy, the main difference with the actuated control is that the adaptive control first verifies (or considers) the state of the complete intersection (i.e., all phase groups of the junction) before deciding to which phase group priority should be assigned and then fixes appropriate timings. Generally, the metric of decision is the traffic demand in phase groups (arrivals in phase groups). Some commonly used schemes for ensuring an adaptive local traffic control are MOVA (microprocessor optimized vehicle actuation) [36], CRONOS (control or networks by optimization of switchovers) [67], and SPPORT (signal priority procedure for optimization in real time) [68], to name a few.

- **Area control**

The coordination between neighboring junctions significantly affects the overall performance. Vehicles leaving a queue at a traffic signal generally travel in a platoon that disperses as vehicles travel further downstream. However, considering the case of neighboring and signalized junctions, a platoon of vehicles released from a junction will not completely disperse before arriving at the next junction. Thus, a suitable coordination of all traffic signals in a network allows the platoons of vehicles to move within a network of coupled traffic junctions without significant dispersion. Signal coordination enhances the overall traffic operation. Regarding signal coordination, four strategies are presented by the current state of the art: *(a) pretimed*, *(b) actuated*, *(c) adaptive*, and *(d) self-organized*. The approaches (a), (b), and (c) are generally focused on the timing plans of one arterial road. Approach (d) considers the synchronization of interwoven traffic streams.

Regarding pretimed area control, two approaches are used to compute timing plans. The first approach is the progression-based method (see MAXBAND [13], PASSER II [69], and PASSER IV [70] as selected examples of progression-based systems). This method is based on the optimization (maximization) of the bandwidth of the progression. The second approach is the disutility-based method (see TRANSYT-7F [7] (Traffic Network Study Tool) and SYNCHRO). This is also an optimization approach aiming at minimizing some key performance metrics, namely the overall delay and the number of stops. These two approaches, which converge to different objectives can easily result in significantly different timing plans under identical traffic conditions.

Regarding actuated area control, coordination is achieved using the same principle as in the case of pretimed area control. However, in order to make sure that traffic-actuated controllers will return to the coordinated phase (in time), a mechanism must be developed to force phases that are uncoordinated to complete (or to terminate). Two types of mechanisms are generally used, which are called force-off modes: (a) the floating force-off and (b) the fixed force-off [71]. A force-off point for each uncoordinated phase is the point where the respective phase must terminate to make sure that the controller returns to the coordinated phase at the right (or proper) time in the cycle.

Regarding adaptive area control, the adaptive strategy uses real-time traffic information in order to permanently update signal timings to fit the current traffic demand. The leading adaptive control schemes are SCOOT [14] and SCATS [72]. SCOOT is an acronym of split, cycle, and offset optimization technique, while the acronym SCATS corresponds to Sydney coordinated adaptive traffic system. These two adaptive techniques have been intensively and successfully used since the 1970s and have led to very good performances in various cities around the world. Further leading schemes that have been successfully used in practice are SCATS, SCOOT, and MOTION (method for optimization of traffic signals in online-controlled networks) [35], OPAC (optimized policies for adaptive control) [59], PRODYN (programmation dynamique) [60], RHODES (real-time, hierarchical, optimized, distributed, and effective system) [61], UTOPIA/SPOT (urban traffic optimization by integrated automation / signal progression optimization technology) [62], and TUC (TRAFFIC-RESPONSIVE URBAN CONTROL) [63].

Regarding traffic self-organized area control, this technique can be considered an alternative that may help to overcome some limitations of standard central as well as decentralized controllers and make an adaptive and purely demand-driven traffic light control practically applicable. Various self-organization mechanisms of conflicting flows have been proposed in the recent relevant literature. The online optimization of a whole traffic network (see [23], based on centralized control principle) is computationally expensive, since it requires vast amounts of data to be collected and processed. To tackle this problem, an alternative method consists in using the decentralized control principle, which consists in placing (or positioning) decentralized (or distributed) controllers at the individual intersections, which may decide autonomously about the operation of the associated traffic lights (see [73], based on the decentralized control principle). However, simple models of networks of coupled oscillators are appropriate to realize self-organized and decentralized traffic light controllers.

2.5 Self-organized Traffic Lights: Models Based on Decentralized Control

The self-organized and decentralized area traffic signal control technique can be considered an alternative that may help to overcome some limitations of standard central as well as decentralized controllers and make an adaptive and purely demand-driven traffic light control practically applicable. Various self-organization mechanisms of conflicting flows have been proposed in the recent relevant literature. The on-line optimization of a whole traffic network (approach followed in adaptive control structures of the second generation [23]) requires vast amounts of data to be collected and processed. As a consequence of the complexity of the corresponding optimization problem, a corresponding control can hardly be operated online, and it is thus very inflexible in the case of exceptional events and natural disasters, but also in reaction to short-term fluctuations in the traffic volume during "normal" conditions. These factors have resulted in a call for the development of a third generation of control methods, as already mentioned above, based on decentralized/distributed controllers placed at individual intersections, which may decide autonomously about the operation of the associated traffic lights. Obviously, such controllers are much more flexible and computationally far less demanding [73]. The result is better adaptivity to the current local traffic conditions, a potentially higher robustness with respect to failures, and the possibility of an online operation due to the low complexity of the corresponding optimization approaches. In contrast, there is the disadvantage that there is no guarantee that the sum of locally optimized strategies also yields a globally optimal service at all intersections. In particular, the capability of decentralized approaches to deal with highly congested networks should be a benchmark for every corresponding method.

Regarding self-organized and decentralized traffic light control models, the Kuramoto model is a typical example that can be used efficiently for real-time applications. This model has especially motivated several studies on possible decentralized approaches for traffic light control. Some selected interesting models are worth mentioning: Sekiyama et al. presented in 2000 [31] the first conceptual study explicitly using the Kuramoto model for the development of a decentralized control strategy for traffic lights. Starting from the standard setting of the Kuramoto model as a system of mutually coupled oscillators described by simple linear interactions, Sekiyama et al. described the demand-dependent self-organized adaptation of the cycle frequency and timing for the case of traffic lights at simple four-armed intersections assuming that these can be operated by only two consecutive service phases. Neglecting pedestrian traffic and related effects in the real-world system, the full service of such an intersection typically requires, however, more than two phases, since there is always one direction of turning traffic that would collide with the straight-on traffic moving in the opposite direction. The model of Sekiyama et al. neglects these separate turning phases; there are, however, some road networks within which this solution has been realized.

Nishikawa et al. [34] have developed an alternative model for an areawide control of traffic signals based on coupled phase oscillators.

Laemmer et al. [31] studied the emergence of phase synchronization in another Kuramoto-type model for traffic lights and proposed some extensions. In the related formalism, all characteristic periods are expressed in terms of phase angles.

Akbas et al. [30] adapted the nonlinear oscillator model to the traffic signal system of a two-way arterial road. The control methodology is based on measurements of the microscopic occupancy parameters for incoming flows at intersections that have a four-way geometric structure with four green splits. The desired signal parameters such as cycle times, green splits, and offsets are adjusted dynamically according to local traffic data. Thus, the desired signal patterns are self-organized through the mutual interactions among the signals.

Some other approaches mimicking the self-organization paradigm have been presented in the literature; it is, however, not clear whether these are really better than those based on the coupled nonlinear oscillator models. A major problem of the related publications is that they do not provide any comparative performance evaluation with respect to those approaches based on the Kuramoto model.

3 A New Concept Involving an Extended Version of the Kuramoto Model for Traffic Light Control and Optimization in a Network of Coupled Traffic Junctions

The aim of this section is to introduce a new concept for traffic light control. The new concept, which is a modified version of the Kuramoto model, exploits the self-organized and decentralized control principle.

This section develops a new concept inspired by the traditional Kuramoto model for traffic control in a network of interacting traffic junctions. The advantages of the new concept are demonstrated by the possibility of solving some of the problems left unsolved by the traditional Kuramoto model. In particular, the new concept provides the following significant contributions in the area of traffic control: The new concept developed provides a detailed and systematic analytical framework to understand how coupled oscillators can be used in traffic control. The current state of the art (i.e., the Kuramoto analytical concept) does not provide a detailed and in-depth analytical study, as is the case in this section. This section demonstrates the possibility of modifying (or extending) the Kuramoto oscillator model to the analysis and control of traffic at isolated junctions. The traditional Kuramoto model is valid only for a network of coupled oscillators, each of which represents a traffic junction. Thus, the Kuramoto model does not consider isolated junctions and consequently cannot model the impacts or effects of the signal phase groups at traffic junctions (this is a microscopic view of traffic junctions).

The above-mentioned points are necessary for the development of a traffic control system that can be efficiently used in practice.

3.1 Architecture of a Road Network and Dynamics of Vehicles at Junctions

A traffic network is generally modeled as a bidirected graph. In such a graph, nodes (or vertices) represent the traffic junctions (or crossings), and the edges (or links) represent the bidirectional segments of roads connecting neighboring junctions. The architecture of a road network is illustrated in Fig. 1. According to this figure, a road network generally involves coupled traffic junctions. A junction handles conflicting directions of traffic flows. The dynamics of traffic at a junction is illustrated in Fig. 2. According to this figure, the junction is made up of four approaches, each of which serves traffic in three directions (this is the general view). Thus, the outgoing flows (from an approach denoted by ϵ_{ij}; see Fig. 2) can be expressed mathematically as the sum of the through traffic (ϵ_{TH}), the left-turn traffic (ϵ_{LT}), and the right-turn traffic (ϵ_{RT}). The relationship among the different traffic directions can be expressed as follows:

$$\epsilon_{ij} = \epsilon_{TH} + \epsilon_{RT} + \epsilon_{LT}, \tag{2a}$$

where

$$\epsilon_{TH} = \alpha_{TH} * \epsilon_{ij}, \tag{2b}$$

$$\epsilon_{RT} = \alpha_{RT} * \epsilon_{ij}, \tag{2c}$$

$$\epsilon_{LT} = \alpha_{LT} * \epsilon_{ij}, \tag{2d}$$

$$\alpha_{TH} + \alpha_{RT} + \alpha_{LT} = 1, \tag{2e}$$

where α_{TH}, α_{RT}, and α_{LT} represent the respective proportions of traffic demand in each direction. Figure 1 shows the coupling topology between traffic junctions, each of which is characterized by a signal; as shown in Fig. 1, the center junction with index i (characterized by the signal S_i) is coupled to its neighboring junctions characterized by signals S_{ij} $(i = 1, 2, 3, 4)$. It is important to mention here that the notation S_{ij} expresses the neighboring signals S_j with respect to the signal S_i.

Figure 3 shows the representation, in complex coordinates $[\Im m(\phi), \Re e(\phi)]$, of the phases of all traffic signals involved in the network of coupled junctions in Fig. 1. The phases of traffic signals (ϕ_i, and ϕ_{ij} $(j = 1, 2, 3, 4)$) are characterized by vectors represented in complex coordinates by their amplitudes and corresponding angular deviations (rotation). The amplitudes of these vectors depend on the traffic demand (or flow) on each road connecting the junction S_i to the junction S_j. This demand, denoted by ϵ_{ij}, is expressed mathematically as follows:

$$\epsilon_{ij} = \frac{1}{\rho_{im} T_i} \int_t^{t+T_i} \rho_{ij}(t) dt; \tag{3}$$

where ρ_{im} is the road capacity, and $\rho_{ij}(t)$ is the traffic density on the road connecting S_i to S_j in the time window $[t, t + T_i]$; T_i is the cycle of the signal S_i. The indices $1 \leq i \leq N$ and $1 \leq j \leq n_i$ are integers; N corresponds to the number of junctions involved in the coupled network, and n_i is the number of neighboring signals S_j connected to S_i.

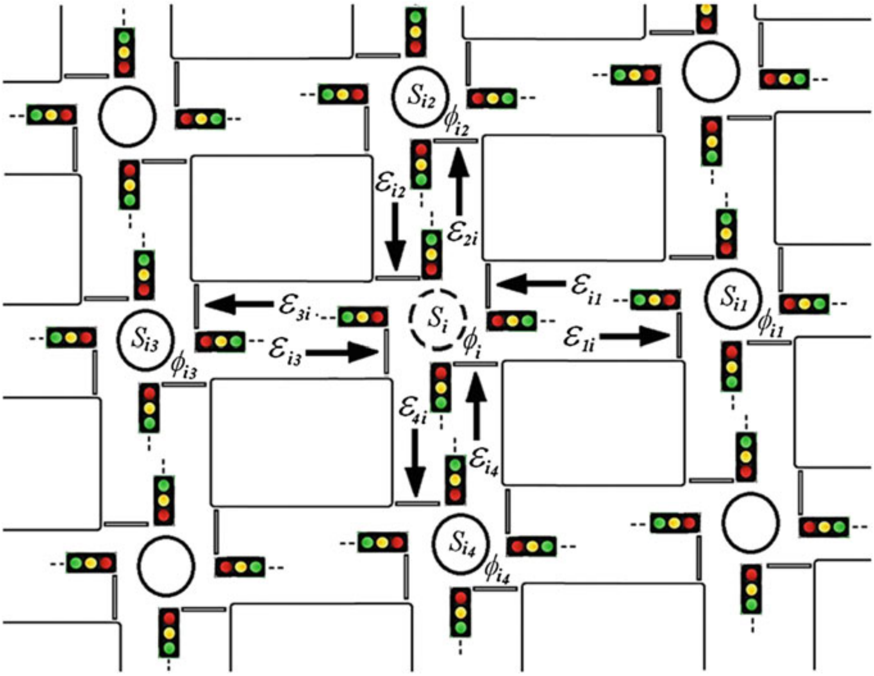


Fig. 1 Synoptic representation of a network of coupled neighboring traffic junctions' S_{in} represent the neighboring signals S_n coupled to S_i; ϵ_{in} is the traffic demand served by signals S_i and S_n; ϕ_i is the *green* signal duration for each of the signal groups of the junction; and ϕ_{in} denote the *green* signal durations to be allocated to each of the four signal groups ($n = 1, 2, 3, 4$)

3.2 Architecture of a Road Network and Dynamics of Vehicles at Junctions

We describe here a technique based on traffic signal timing used at a traffic junction to assign (or remove) right-of-way (priority). Signal timing provides the green duration at each approach (see the green signals ϕ_1, ϕ_2, ϕ_2, and ϕ_4 (phase groups) in Fig. 2) of the junction as well as the green duration for pedestrians, among other things. The traffic signals generally involve green, yellow, and red colors, each of which assigns (or removes) right-of-way. The split is generally adjusted within a cycle according to the traffic demand (i.e., incoming traffic in approaches). Thus, according to Fig. 2, the cycle length (or cycle time), denoted by $T_i(n)$, is expressed as follows:

$$T_i(n) = \sum_{i=1}^{4}(\phi_i + L_i), \tag{4a}$$

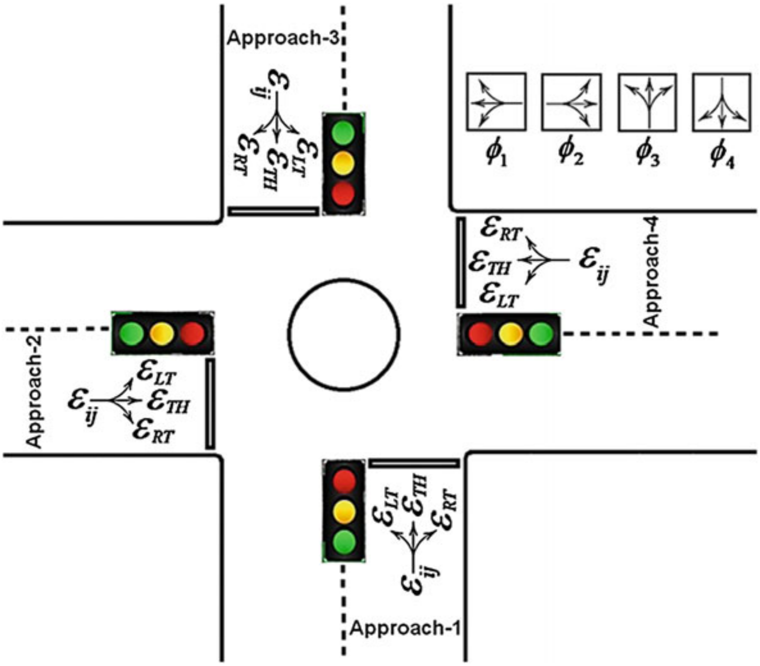


Fig. 2 Illustration of the general structure of a traffic junction: ϵ_{ij} is the outgoing traffic flow from an approach; ϵ_{TH} is the straight traffic; ϵ_{RT} is the *right-turn* traffic; ϵ_{LT} is the *left-turn* traffic; ϕ_1, ϕ_2, ϕ_2, and ϕ_4 are the signal phase groups (i.e., the *green* signal durations assigned to each approach)

where n corresponds to the index of the cycle, ϕ_i denotes the green signal duration for each of the signal groups of the junction in Fig. 2, and L_i is the time needed by each signal group to clear the junction; L_i is also called "lost time." In practice, the yellow and red durations are generally approximated by 3 and 2 s respectively. Thus, Eq. (4a) is written in the form

$$T_i(n) = 4L_i + \sum_{i=1}^{4} \phi_i(n). \tag{4b}$$

Considering a pretimed control (this is a control strategy based on constant or fixed signal timing or splitting), the expression in Eq. (4b) holds, since the splitting in this context does not depend on the variation in the traffic demand. However, in order to consider the demand changes, the actuated control appears to be the appropriate control scheme. This scheme amends the signals' timing according to changes in the traffic demand. Thus, a new form of Eq. (4b) is needed to perform a dynamic control at a junction. This new form is derived by transforming Eq. (4b) as follows:

$$\phi_1(n) = \alpha_1 * [T_i(n) - 4L_i], \tag{5a}$$

$$\phi_2(n) = \alpha_2 * [T_i(n) - 4L_i], \tag{5b}$$

$$\phi_3(n) = \alpha_3 * [T_i(n) - 4L_i], \tag{5c}$$

$$\phi_4(n) = \alpha_4 * [T_i(n) - 4L_i], \tag{5d}$$

where α_i is the proportion of traffic demand in the approach with index i. If we denote by $\epsilon_{ij}(ap_k)$ the traffic demand (within a cycle) in an approach with index i, the total demand $\epsilon_{ij}(total)$ in the junction during a cycle time $T_i(n)$ is expressed as follows (ap_k stands for "approach number k (see approaches illustrated in Fig. 2)):

$$\epsilon_{ij}(total) = \sum_{k=1}^{4} \left[\epsilon_{ij}(ap_k)\right] = \epsilon_{ij}(ap_1) + \epsilon_{ij}(ap_2) + \epsilon_{ij}(ap_3) + \epsilon_{ij}(ap_4). \tag{6}$$

Using Eq. (6), the coefficients α_i are obtained as follows:

$$\alpha_1 = \frac{\epsilon_{ij}(ap_1)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]}, \tag{7a}$$

$$\alpha_2 = \frac{\epsilon_{ij}(ap_2)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]}, \tag{7b}$$

$$\alpha_3 = \frac{\epsilon_{ij}(ap_3)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]}, \tag{7c}$$

$$\alpha_4 = \frac{\epsilon_{ij}(ap_4)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]}. \tag{7d}$$

Therefore, substituting Eq. (7) into (5) leads to the following expression for the green signal splits in terms of the traffic demand variations observed in all approaches of the traffic junction:

$$\phi_1^*(n) = \frac{\epsilon_{ij}(ap_1)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]} * [T_i(n) - 4L_i], \tag{8a}$$

$$\phi_2^*(n) = \frac{\epsilon_{ij}(ap_2)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]} * [T_i(n) - 4L_i], \tag{8b}$$

$$\phi_3^*(n) = \frac{\epsilon_{ij}(ap_3)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]} * [T_i(n) - 4L_i], \tag{8c}$$

$$\phi_4^*(n) = \frac{\epsilon_{ij}(ap_4)}{\sum_{k=1}^{4}[\epsilon_{ij}(ap_k)]} * [T_i(n) - 4L_i]. \tag{8d}$$

The quantities $\phi_1^*(n)$, $\phi_2^*(n)$, $\phi_3^*(n)$, and $\phi_4^*(n)$ represent the expected (or the desired) green signal durations to be allocated to each of the four signal groups of the traffic junction indexed by i. The parameter n corresponds to the index of the cycle. According to Eq. (8), the desired green splits depend on the traffic volume. This traffic volume is evaluated at the end of the previous cycle. Therefore, according to the traffic demand variation, the green signal splits must be updated. Sekiyama et al. (2001) [31] have proposed the following analytical expression for updating the desired green signals:

$$\phi_i(n+1) = \phi_i(n) + \gamma[\phi_i^*(n) - \phi_i(n)], \tag{9a}$$

where γ is an updating constant parameter (i.e., a parameter that is monitored during updating). We consider the basic Kuramoto model in Eq. (9b) (see [33, 34]),

$$\frac{d\phi_i}{dt} = \omega_i + \frac{K}{n_i}\sum_{j=1}^{n_i}\epsilon_{ij}\Gamma(\phi_i, \phi_{ij}), \tag{9b}$$

and for the sake of simplification, the periodic function $\Gamma_i(\phi_i, \phi_{ij})$ with period 2π is expressed in the sinusoidal form as shown in Eq. (9c). Here n_i is the number of coupled phase oscillators (here represented by signals S_i in Fig. 1), and K is a coupling constant that depends on the traffic volume in the network (here denoted by ϵ_{ij}):

$$\Gamma_i(\phi_i, \phi_{ij}) = sin(\phi_{ij} - \phi_i) = \frac{e^{i(\phi_{ij}-\phi_i)} - e^{-i(\phi_{ij}-\phi_i)}}{2i}. \tag{9c}$$

Thus, substituting Eq. (9c) into (9b) leads to the following relation:

$$\frac{d\phi_i}{dt} = \omega_i + \frac{K}{2i}\left[e^{-i\phi_i}\sum_{j=1}^{n_i}\left(\frac{\epsilon_{ij}}{n_i}e^{i\phi_{ij}}\right) - e^{i\phi_i}\sum_{j=1}^{n_i}\left(\frac{\epsilon_{ij}}{n_i}e^{-i\phi_{ij}}\right)\right]. \tag{10}$$

We now envisage the existence of two fundamental parameters, $\bar{\phi_i}$ (the weighted phase average expressed by Eq. (13d)) and σ_i (the amplitude of the weighted vector corresponding to $\bar{\phi_i}$). Geometrically (see Fig. 3), the amplitude σ_i is obtained as the vectorial sum of a_i and b_i. The first term (a_i) corresponds to the sum (in the real-axis direction $\Re e$) of the amplitudes of all neighboring signals coupled to the signal S_i. The second term (b_i) corresponds to the sum (in the imaginary-axis direction $\Im m$) of the amplitudes of all neighboring signals coupled to the signal S_i. This geometric interpretation of the representation in complex coordinates in Fig. 3 leads to the following expressions:

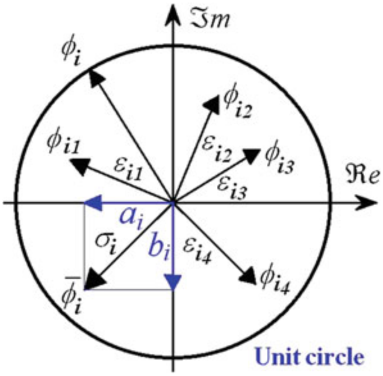


Fig. 3 Representation in complex coordinates of the phases of coupled neighboring traffic junctions. The symbol $\bar{\phi}_i$ denotes the weighted phase average of signal S_i

$$\sigma_i e^{i\bar{\phi}_i} = \sum_{j=1}^{n_i}\left(\frac{\epsilon_{ij}}{n_i}e^{i\phi_{ij}}\right) = a_i + ib_i, \tag{11a}$$

$$\sigma_i e^{-i\bar{\phi}_i} = \sum_{j=1}^{n_i}\left(\frac{\epsilon_{ij}}{n_i}e^{-i\phi_{ij}}\right) = a_i - ib_i, \tag{11b}$$

where $\bar{\phi}_i$ stands for the average phase of the signal S_i, and σ_i is the corresponding amplitude (see the representation in complex coordinates in Fig. 3). The aim of this transformation is to provide a simplified form of Eq. (10). This simplified form is obtained in Eq. (12) by substituting Eq. (11) into (10):

$$\frac{d\varphi_i}{dt} = \omega_i + \frac{\sigma_i K}{2i}\left[e^{i(\bar{\varphi}_i - \varphi_i)} - e^{-i(\bar{\varphi}_i - \varphi_i)}\right] = \omega_i + \sigma_i K sin(\bar{\varphi}_i - \varphi_i). \tag{12}$$

In Eq. (12), the problem is related to the determination of the analytical forms of the new variables ($\bar{\varphi}_i$ and σ_i), which have been introduced for the sake of simplification. Thus, the following relationships are derived according to Eq. (11):

$$a_i = \sum_{j=1}^{n_i}\frac{\epsilon_{ij}}{n_i}cos(\varphi_{ij}), \tag{13a}$$

$$b_i = \sum_{j=1}^{n_i}\frac{\epsilon_{ij}}{n_i}sin(\varphi_{ij}), \tag{13b}$$

$$\sigma_i = \sqrt{a_i^2 + b_i^2} = \frac{1}{n_i}\sqrt{\left[\sum_{j=1}^{n_i}\epsilon_{ij}cos(\varphi_{ij})\right]^2 + \left[\sum_{j=1}^{n_i}\epsilon_{ij}sin(\varphi_{ij})\right]^2}, \tag{13c}$$

$$\bar{\phi_i} = arctan\left[\frac{b_i}{a_i}\right] = arctan\left[\frac{\sum_{j=1}^{n_i} \epsilon_{ij} sin(\varphi_{ij})}{\sum_{j=1}^{n_i} \epsilon_{ij} cos(\varphi_{ij})}\right]. \tag{13d}$$

The quantities a_i, b_i, $\bar{\phi_i}$, and σ_i are represented in complex coordinates in Fig. 3. The advantage of the new set of equations (see Eq. (13)) is that it allows a physical interpretation of the functioning principle of the Kuramoto model and thus a clear understanding of the self-organization phenomenon within the coupled oscillator. For further elaboration, full details are provided in the sections below.

3.3 Self-organization and Offset Settings in the Kuramoto Oscillator

In order to illustrate the general applicability of the Kuramoto oscillator for traffic signal control and also demonstrate the achievement of self-organization (i.e., local synchronization), we consider the synoptic representations in Figs. 1 and 3. These figures illustrate the coupling between neighboring junctions as well as the representation of the phases of signals in the complex plane. Each signal is characterized by a frequency ω_i (ω_{ij}) and a corresponding phase ϕ_i, and ϕ_{ij} ($i = 1, 2, 3, \ldots, 4$) for the neighboring junctions. These phases are represented in the complex coordinate system, denoted by [$\Im m(\phi_i)$,$\Re m(\phi_i)$] (see Fig. 3), in the form of vectors, each of which is characterized by a specific amplitude and a corresponding angle deviation. As shown in Fig. 3, the phase ϕ_i and its neighboring phases ϕ_{ij} ($i = 1, 2, 3, 4$) are expressed in the form of vectors in complex space. Thus, in order to demonstrate the occurrence of self-organization in the Kuramoto oscillator, we introduce a relative phase $\psi_i(t)$ expressed as follows:

$$\psi_i(t) = (\bar{\phi_i} - \phi_i). \tag{14a}$$

Combining Eq. (14a) with Eq. (12) leads to

$$\frac{d\psi_i}{dt} = \frac{d\bar{\phi_i}}{dt} - \omega_i - \sigma_i K sin(\psi_i). \tag{14b}$$

Further, let us denote the weighted phase average $\bar{\phi_i}$ by

$$\bar{\phi_i} = \Omega_i t + \phi_i(t_0), \tag{14c}$$

where Ω_i denotes the natural frequency of the weighted phase, and $\phi_(t_0)$ is the initial phase. Thus, combining Eq. (14b) with Eq. (14c) leads to the following expression:

$$\frac{d\psi_i}{dt} = \Omega_i - \omega_i - \sigma_i K sin(\psi_i). \tag{14d}$$

Self-organization (or local synchronization) is achieved at equilibrium of the model describing the behavior of the Kuramoto oscillator model. This equilibrium point (denoted by ψ_i^*) is expressed as follows:

$$\begin{cases} \psi_i^* = arcsin\left(\frac{\Omega_i - \omega_i^*}{\sigma_i^* K}\right) \\ \qquad\qquad and \\ \qquad \left|\frac{\Omega_i - \omega_i^*}{\sigma_i^i K}\right| \leq 1. \end{cases} \tag{15}$$

Thus when the above inequality is satisfied, phase-locking occurs with a phase difference corresponding to ψ_i^*. Further, ψ_i^* is a constant parameter, since it is determined at equilibrium. However, according to Eq. (15), ψ_i^* is constant if and only if σ_i^* is constant. This condition leads to the derivation of the following expression for σ_i^* (obtained through algebraic manipulation of Eq. (11):

$$\sigma_i^* = \sqrt{a_i^2 + b_i^2} = \frac{1}{n_i}\sqrt{\left[\sum_{j=1}^{n_i} \epsilon_{ij} cos(\varphi_{ij} - \bar{\varphi}_i)\right]^2 + \left[\sum_{j=1}^{n_i} \epsilon_{ij} sin(\varphi_{ij} - \bar{\varphi}_i)\right]^2}. \tag{16}$$

Further, according to Eq. (14a), when the oscillators are phase-locked, ψ_i^* is constant, and this leads to the following expression:

$$\frac{d\bar{\phi}_i}{dt} = \frac{d\phi_i}{dt}. \tag{17}$$

Thus the following characteristic equation is obtained, from which the phase differences between the neighboring S_{ij} signals and the center signal S_i must be dynamically adjusted regarding variation observed in traffic demand:

$$\sigma_i^* = \sqrt{a_i^2 + b_i^2} = \frac{1}{n_i}\sqrt{\left[\sum_{j=1}^{n_i} \epsilon_{ij} cos(\Delta\varphi_{ij})\right]^2 + \left[\sum_{j=1}^{n_i} \epsilon_{ij} sin(\Delta\varphi_{ij})\right]^2}. \tag{18}$$

The next section will provide a full description of the adjustment principle leading to an appropriate use of Eq. (18).

3.4 Signal Timing and Offset Settings Adjustment: Illustration of both Desired Offset and Desired Relative Phase

The desired offset and desired relative phase are two important concepts that must be well understood in the framework of this analysis. In order to provide a good understanding of these two concepts, we consider the representation of signal splitting in Fig. 4.

We now want to determine the appropriate offset settings of consecutive intersections (or consecutive signals) using the phase difference in Eq. (18). The desired offsets here are obtained as the phase shift between the signals in both primary and secondary directions.

According to Fig. 4, the desired offset Δt_{ij}^{*} between the signals S_{i-1} and S_i is defined as the travel time experienced by a vehicle in moving from signal S_{i-1} to signal S_i. Let us denote by $d_{s_{i-1}\rightarrow s_i}$ the distance between S_{i-1} and S_i, and by v_{ij} the speed on the road. Assuming the case that a speed limit v_{ij}^{max} is imposed on the road, the expected travel time is expressed as follows:

$$\Delta t_{ij}^{*} = \frac{d_{s_{i-1}\rightarrow s_i}}{v_{ij}^{max}}. \tag{19}$$

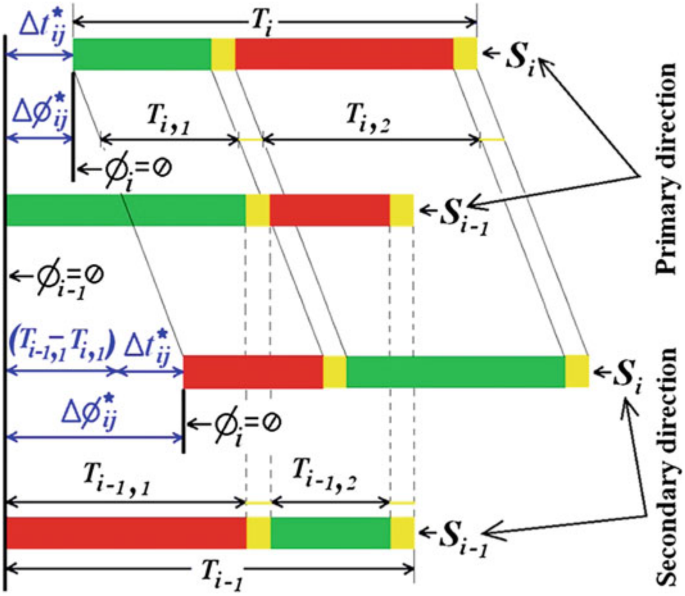


Fig. 4 Illustration of signal splitting in both the primary and secondary directions. The desired offset Δt_{ij}^{*} (in the primary direction) and desired relative phase $\Delta\phi_{ij}^{*}$ (in the secondary direction) are clearly illustrated

The expected time Δt_{ij}^{*} is used to calculate the desired offset between the signals S_{i-1} and S_i (in the primary direction of the flow). The calculation is conducted as follows. Using complex coordinates, the phase corresponding to the duration Δt_{ij}^{*} is expressed as $\left[\omega^{*}\Delta t_{ij}^{*}+\phi_i(t_0)\right]$, where $\omega_i=2\pi/T_i$ is the frequency of the signal S_i, and $\phi_i(t_0)=0$ is the starting phase (initial phase) at the appearance of the green signal S_i in the primary direction. Thus, the desired offset $\Delta\phi_{ij}^{*}$ (in the primary direction) between S_{i-1} and S_i is expressed as follows:

$$\Delta\varphi_{ij}^{*}=\omega*\Delta t_{ij}^{*}=\frac{2\pi}{T_i}\Delta t_{ij}^{*}=\left(\frac{2\pi}{T_i}\right)\left(\frac{d_{s_{i-1}\rightarrow s_i}}{v_{ij}^{max}}\right). \tag{20}$$

Remark 1 The desired offset is obtained by considering the flow in the primary direction. The desired offset corresponds to the interval (or duration) between the appearance of the green signal in S_{i-1} and the appearance of the green signal in S_i.

Remark 2 The desired relative phase is obtained by considering the oscillators in both primary and secondary directions. According to Fig. 4, the desired relative phase is obtained (in the secondary direction of the flow) as the interval (or duration) between the appearance of the red signal in S_{i-1} and the appearance of the red signal in S_i.

According to the above remarks, the desired relative phase is expressed as follows:

$$\Delta\phi_{ij}^{*}=\omega*\Delta t_{ij}^{*}=\frac{2\pi}{T_i}\Delta t_{ij}^{*}=\left(\frac{2\pi}{T_i}\right)\left(T_{i-1,1}-T_{i,1}+\Delta t_{ij}^{*}\right). \tag{21}$$

Equation (21) represents the expression of the desired relative phase. However, this expression does not depend on the traffic demand. Thus, derivation of a second expression that does depend on the traffic demand is necessary. We now want to derive the second expression of the desired relative phase in terms of the traffic demand. This is achieved by considering Eqs. (17) and (18) and also by assuming the occurrence of self-organization. These two conditions lead to the following expression:

$$\bar{\varphi}_i^{*}=arctan\left[\frac{b_i}{a_i}\right]=arctan\left[\frac{\sum_{j=1}^{n_i}\epsilon_{ij}sin(\varphi_{ij}^{*})}{\sum_{j=1}^{n_i}\epsilon_{ij}cos(\varphi_{ij}^{*})}\right]. \tag{22}$$

Further, applying the conditions Eqs. (17) and (18) to Eq. (22) leads to

$$\bar{\varphi}_i^{*}=arctan\left[\frac{\sum_{j=1}^{n_i}\epsilon_{ij}sin(\Delta\varphi_{ij}^{*})}{\sum_{j=1}^{n_i}\epsilon_{ij}cos(\Delta\varphi_{ij}^{*})}\right]. \tag{23}$$

This expression (Eq. (23)) corresponds to the desired relative phase, since at local synchronization (i.e., self-organization), all the neighboring oscillators have the same frequency. The expression in Eq. (23) is the characteristic equation that is used to adjust the desired offsets (i.e., traffic timing of signals splitting) with regard to the variation of the traffic demand.

The next section is concerned with the establishment of analytical conditions showing that even the cycle times of the signals' frequencies can be adjusted using the Kuramoto oscillator model.

3.5 Cycle Time Adjustments

The preceding results (see Eq. (23)) are now used to demonstrate the possibility of adjusting the cycle time (or cycle length) as the flow demand varies. Using Eq. (15), the following expression is obtained:

$$\omega_i^* = \Omega_i - \sigma_i^* K sin(\psi_i^*). \tag{24}$$

Further, according to Eq. (23), we have $\psi_i^* = \bar{\phi_i}^*$ at local synchronization and at the appearance (or start) of the green signal in the primary direction (for $\phi_i = 0$). Thus Eqs. (15) and (23) are combined to obtain the following characteristic system, from which the cycle time is adjusted in terms of the variations of the traffic demand:

$$\begin{cases} \omega_i^* = \Omega_i - \sigma_i^* K sin(\bar{\phi}_i^*), \\ \bar{\phi}_i^* = arctan\left[\frac{\sum_{j=1}^{n_i} \epsilon_{ij} sin(\Delta\phi_{ij}^*)}{\sum_{j=1}^{n_i} \epsilon_{ij} cos(\Delta\phi_{ij}^*)}\right]. \end{cases} \tag{25}$$

Thus, using Eq. (25), the desired natural frequency ω_i^* of the signal S_i can be updated according to the following rule proposed by Sekiyama et al. (2001) [31]:

$$\omega_i(n+1) = \omega_i(n) + \alpha[\omega_i^* - \omega_i(n)] + \beta[\bar{\omega}_i^* - \omega_i(n)], \tag{26}$$

where α and β are binary parameters. The threshold frequency $\bar{\omega}_i^*$ is introduced for each signal in order to avoid the occurrence of frequency entrainment between the natural frequencies of the signals involved in the coupled traffic network.

Derivation of $\bar{\omega}_i$ in terms of the demand flow: In order to derive the expression of $\bar{\omega}_i$, we consider Eq. (14a). Further, we consider that the local synchronization of the signal S_i with its neighboring signals S_{ij} is effective. We also consider the original equation of the Kuramoto oscillator model with the nonlinear function in sinusoidal form (Eq. (10)).

According to Eq. (10), the behavior of the neighboring oscillators to S_i is modeled mathematically by the following expression:

$$\frac{d\phi_{ij}}{dt} = \omega_{ij} + \frac{K}{n_i}\epsilon_{ij} sin(\phi_i - \phi_{ij}) + F_{ij}, \tag{27}$$

where ω_{ij} is the frequency of neighboring signals S_{ij} coupled to the signal S_i; n_i stands for the number of neighboring signals S_{ij} coupled to signal S_i; F_{ij} represents the mutual influence of the neighboring oscillators/signals (except S_i) on S_{ij}. The quantity F_{ij} is the frequency of signals (except S_i) that are directly coupled to S_{ij}. Equation (27) can be rewritten as the following expression:

$$\begin{cases} \frac{d\phi_{ij}}{dt} = \omega_{ij} + K\epsilon_{ij} sin(\phi_i - \phi_{ij}) + \Delta F_{ij}, \\ \\ \Delta F_{ij} = F_{ij} - \frac{n_i - 1}{n_i} K\epsilon_{ij} sin(\phi_i - \phi_{ij}). \end{cases} \tag{28}$$

Considering the occurrence of local synchronization, the following relation expresses the phase entrainment of the coupled oscillators:

$$\frac{d\phi_{ij}}{dt} = \frac{d\phi_i}{dt} = \bar{\omega_i}, \tag{29}$$

where $\bar{\omega_i}$ is the compromise frequency of the oscillators when they are phase-locked. Combining Eq. (28) with Eq. (29) leads to the following expression:

$$\omega_{ij} + K\epsilon_{ij} sin(\phi_i - \phi_{ij}) + \Delta F_{ij} = \omega_i + \sum_{j=1}^{n_i} \frac{K}{n_i} \epsilon_{ij} sin(\phi_{ij} - \phi_i) = \bar{\omega_i}. \tag{30}$$

Thus, using Eq. (30), the following set of equations is derived:

$$\begin{cases} \omega_{ij} + K\epsilon_{ij} sin(\phi_i - \phi_{ij}) + \Delta F_{ij} = \bar{\omega_i}, \\ \\ \omega_i + \frac{K}{n_i} \sum_{j=1}^{n_i} \epsilon_{ij} sin(\phi_{ij} - \phi_i) = \bar{\omega_i}. \end{cases} \tag{31}$$

From Eq. (31), some algebraic manipulations are performed, and the following expression of the compromise frequency of the oscillators is obtained (at steady state and when the signals are phase-locked):

$$\bar{\omega_i} = \frac{\left[\omega_i + \frac{\Delta F_{ij}}{n_i} \sum_{j=1}^{n_i} \frac{\epsilon_{ij}}{\epsilon_{ji}} + \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\epsilon_{ij}}{\epsilon_{ji}} \omega_{ij}\right]}{\left[1 + \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\epsilon_{ij}}{\epsilon_{ji}}\right]}. \tag{32}$$

Finally, assuming that the mutual interaction between the signals does not vary significantly ($\Delta F_{ij} \approx 0$), Eq. (32) leads to the following simplified expression:

$$\bar{\omega_i} = \frac{\left[\omega_i + \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\epsilon_{ij}}{\epsilon_{ji}} \omega_{ij}\right]}{\left[1 + \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\epsilon_{ij}}{\epsilon_{ji}}\right]}. \tag{33}$$

The expression in Eq. (33) is fundamental to the signal control procedure based on coupled oscillators (i.e., Kuramoto), since it offers the possibility of adjusting the frequencies of the oscillators when changes occur in the traffic demand.

3.6 Mathematical Modeling of the Traffic Signal Control Strategy at Junctions Using a Modified Version of the Kuramoto Oscillator Model

This subsection demonstrates the possibility of modifying the Kuramoto concept and then extending it to local traffic control. The extended form of the Kuramoto model is a modified form that can be used to model the dynamics of the green signal of all phase groups involved in a local traffic junction.

The analysis in this section considers the case of a traffic junction involving N phase groups each of which is considered a phase oscillator with amplitude a_k and phase ϕ_k. The amplitude of each phase oscillator represents the traffic demand in the phase group, while the phase of the oscillator represents the green signal duration assigned to the phase group.

Each phase oscillator is assumed to be directly coupled to a central oscillator. The fundamental parameters of the central oscillator are the amplitude a_i and phase ϕ_i. The amplitude a_i represents the total traffic demand at an isolated junction during a cycle time/length, while the phase ϕ_i represents the time cycle length of the traffic junction.

A master–slave coupling materializes the interaction between the coupled phase oscillators and the central oscillator. Considering the case of an isolated traffic junction with N phase groups, this interaction can be modeled mathematically as follows:

$$\begin{cases} \frac{d\phi_1}{dt} = \omega_1 \pm \epsilon_1 sin(\phi_i - \phi_1), \\ \frac{d\phi_2}{dt} = \omega_2 \pm \epsilon_2 sin(\phi_i - \phi_2), \\ \frac{d\phi_3}{dt} = \omega_3 \pm \epsilon_3 sin(\phi_i - \phi_3), \\ \frac{d\phi_4}{dt} = \omega_4 \pm \epsilon_4 sin(\phi_i - \phi_4), \\ \qquad \qquad \ldots\ldots\ldots\ldots \\ \qquad \qquad \ldots\ldots\ldots\ldots \\ \qquad \qquad \ldots\ldots\ldots\ldots \\ \frac{d\phi_N}{dt} = \omega_N \pm \epsilon_N sin(\phi_i - \phi_N), \end{cases} \tag{34}$$

where the cycle length is expressed as

$$\phi_i = (\phi_1 + \phi_2 + \phi_3 + \phi_5 + \cdots + \phi_N) + N\delta, \tag{35}$$

and ϕ_k and ϵ_k represent the green signal and the traffic demand in a phase group with index k, respectively; δ stands for the lost time experienced by each of the N phase groups. In this context we assumed that the lost times experienced by all phase groups are identical. This assumption is commonly used even in practice. Equations (34) and (35) are combined to obtain the following set of coupled ODEs:

$$\begin{cases} \frac{d\phi_1}{dt} = \omega_1 \pm \epsilon_1 sin\left[\phi_2 + \phi_3 + \phi_4 + \cdots + \phi_N + N\delta\right], \\ \frac{d\phi_2}{dt} = \omega_2 \pm \epsilon_2 sin\left[\phi_1 + \phi_3 + \phi_4 + \cdots + \phi_N + N\delta\right], \\ \frac{d\phi_3}{dt} = \omega_3 \pm \epsilon_3 sin\left[\phi_1 + \phi_2 + \phi_4 + \cdots + \phi_N + N\delta\right], \\ \frac{d\phi_4}{dt} = \omega_4 \pm \epsilon_4 sin\left[\phi_1 + \phi_2 + \phi_3 + \cdots + \phi_N + N\delta\right], \\ \qquad \ldots\ldots\ldots \\ \qquad \ldots\ldots\ldots \\ \qquad \ldots\ldots\ldots \\ \frac{d\phi_N}{dt} = \omega_N \pm \epsilon_N sin\left[\phi_1 + \phi_2 + \phi_3 + \cdots + \phi_{N-1} + N\delta\right]. \end{cases} \tag{36}$$

Thus, the coupled system in Eq. (36) expresses the interaction between the green signals at isolated traffic junctions. This coupled system can be suitably (or appropriately) used by exploiting the following statements (or conclusions) proposed by Kuramoto (see [31, 34]) when analyzing synchronization of coupled nonlinear oscillators. The concluding remarks proposed by Kuramoto are concerned with the fundamental parameters (i.e., ϵ_k, ω_k, and ϕ_k) of the coupled oscillators and the effects of those parameters on the occurrence of both frequency and/or phase entrainments:

- ***Statement by Kuramoto regarding frequency entrainment***: When the natural frequencies ω_k of the coupled oscillators are different, the frequency entrainment can occur only by monitoring the coupling terms ϵ_k.
- ***Statement by Kuramoto regarding phase entrainment***: When the natural frequencies ω_k of the coupled oscillators are identical, there always exists a unique stable solution of the phase entrainment with a constant phase shift $\Delta\phi_k = (\phi_i - \phi_k)$. This solution is obtained at the equilibrium point of the coupled system.

The analysis carried out here exploits the second statement (proposed by Kuramoto) by assuming that the natural frequencies of the coupled oscillators are identical. Thus, when this condition is fulfilled, phase entrainment occurs at equilibrium. Therefore, we now seek the equilibrium point of the system. This point is obtained and corresponds to a unique set of values assigned to the green signals ϕ_k of all phase groups involved in the isolated junctions. Using the preceding equations, the green signals are obtained as solutions of the following equations:

$$\begin{cases} \phi_2 + \phi_3 + \phi_4 + \cdots + \phi_N = -N\delta \pm arcsin\left[\frac{\omega_1}{\epsilon_1}\right], \\ \phi_1 + \phi_3 + \phi_4 + \cdots + \phi_N = -N\delta \pm arcsin\left[\frac{\omega_2}{\epsilon_2}\right], \\ \phi_1 + \phi_2 + \phi_4 + \cdots + \phi_N = -N\delta \pm arcsin\left[\frac{\omega_3}{\epsilon_3}\right], \\ \phi_1 + \phi_2 + \phi_3 + \cdots + \phi_N = -N\delta \pm arcsin\left[\frac{\omega_4}{\epsilon_4}\right], \\ \qquad\qquad\qquad\qquad\qquad\qquad \dots\dots\dots\dots \\ \qquad\qquad\qquad\qquad\qquad\qquad \dots\dots\dots\dots \\ \qquad\qquad\qquad\qquad\qquad\qquad \dots\dots\dots\dots \\ \phi_1 + \phi_2 + \phi_3 + \cdots + \phi_{N-1} = -N\delta \pm arcsin\left[\frac{\omega_N}{\epsilon_N}\right]. \end{cases} \tag{37}$$

Thus the general solution of Eq. (37) is expressed as follows:

$$\phi_k = \left(\frac{-1}{N-1}\right)\left[(N-2)a_k - \sum_{j=1, j\neq k}^{N} a_j\right], \geq 0 \tag{38a}$$

where

$$a_k = -N\delta \pm arcsin\left[\frac{\omega_k}{\epsilon_k}\right] \quad (k = j) \quad \text{and} \quad -1 \leq \left[\frac{\omega_k}{\epsilon_k}\right] \leq 1 \tag{38b}$$

and ϕ_k corresponds to the specific value allocated to the green signal of each phase group. The cycle time/length c is expressed in terms of the traffic parameters as follows:

$$c = \frac{\sum_{k=1}^{N} a_k}{(N-1)}. \tag{38c}$$

Considering (for illustration) the case of an isolated traffic junction with four phase groups as shown in Fig. 2, the allocated green signals are expressed analytically by the following expressions:

$$\begin{cases} \phi_1 = \frac{1}{3}\left\{-4\delta \pm 2arcsin\left[\frac{\omega_1}{\epsilon_1}\right] \pm arcsin\left[\frac{\omega_2}{\epsilon_2}\right] \pm arcsin\left[\frac{\omega_3}{\epsilon_3}\right] \pm arcsin\left[\frac{\omega_4}{\epsilon_4}\right]\right\}, \\ \phi_2 = \frac{1}{3}\left\{-4\delta \pm arcsin\left[\frac{\omega_1}{\epsilon_1}\right] \pm 2arcsin\left[\frac{\omega_2}{\epsilon_2}\right] \pm arcsin\left[\frac{\omega_3}{\epsilon_3}\right] \pm arcsin\left[\frac{\omega_4}{\epsilon_4}\right]\right\}, \\ \phi_3 = \frac{1}{3}\left\{-4\delta \pm arcsin\left[\frac{\omega_1}{\epsilon_1}\right] \pm arcsin\left[\frac{\omega_2}{\epsilon_2}\right] \pm 2arcsin\left[\frac{\omega_3}{\epsilon_3}\right] \pm arcsin\left[\frac{\omega_4}{\epsilon_4}\right]\right\}, \\ \phi_4 = \frac{1}{3}\left\{-4\delta \pm arcsin\left[\frac{\omega_1}{\epsilon_1}\right] \pm arcsin\left[\frac{\omega_2}{\epsilon_2}\right] \pm arcsin\left[\frac{\omega_3}{\epsilon_3}\right] \pm 2arcsin\left[\frac{\omega_4}{\epsilon_4}\right]\right\}. \end{cases} \tag{39}$$

The traffic parameters are monitored (as control parameters) to adjust the splits (i.e., green signal sharing among all phase groups) according to the following fundamental expression of the cycle time/length in terms of the input traffic parameters:

$$c = -\frac{16}{3}\delta + \frac{1}{3}\left[\pm arcsin\left[\frac{\omega_1}{\epsilon_1}\right] \pm arcsin\left[\frac{\omega_2}{\epsilon_2}\right] \pm arcsin\left[\frac{\omega_3}{\epsilon_3}\right] \pm arcsin\left[\frac{\omega_4}{\epsilon_4}\right]\right]. \tag{40}$$

The fundamental Eq. (40) shows the phase splits in a traffic junction (see Fig. 2) for a fixed (or constant) cycle time c. An important remark to be underscored is that the duration of the green signal in each phase group is expressed in terms of the total traffic demand in all lane groups, each of which is controlled by a specific phase group (to which is assigned a signal phase). Equation (40) can easily be extended to a network of coupled traffic junctions. In this case, the total traffic demand corresponds to the throughput of each traffic junction involved in the complete network of coupled junctions. The throughput corresponds to the number of vehicles going through the junction during a unit of time (1 h, cycle length, fixed period/duration, etc.).

4 Concluding Remarks and Outlook

This chapter has provided an in-depth review of traffic control and optimization methods, concepts, systems, and tools. A comprehensive overview of the state of the art of traditional traffic control strategies has been presented. The pros and cons of those traditional control schemes have been discussed. Further, the chapter has developed a new analytical concept for traffic control and optimization based on coupled nonlinear oscillators. The concept developed is based on the seminal Kuramoto model. The advantage of the proposed analytical concept is justified by the strong correlation between the signal phases and the traffic demand in lane groups. Thus, the concept developed is viewed as a typical example of an adaptive concept that is very sensitive to variation in traffic demands in lane groups. We have carried out an in-depth analytical modeling procedure, and some analytical expressions have been derived under which the control strategy is effective. Our main interest has been to provide the details of the mathematical modeling of the adaptive control strategy at isolated traffic junctions. The results achieved contribute significantly to the enrichment of the current state of the art, since new expressions or formulas have been proposed for an adaptive control strategy at isolated junctions. The mathematical formulas obtained are based on deterministic coefficients. However, a straightforward extension of the deterministic mathematical models to stochastic models is possible. This can be achieved by assigning probability distributions to deterministic coefficients. Finally, the concept developed is of high flexibility and can be easily extended to a network of coupled traffic junctions.

An interesting ongoing project is concerned with the extension of the analytical concept developed here to the case of stochastic models. Further, the application of the stochastic mathematical models to concrete traffic scenarios (defined in the form of case studies) is also under consideration as proofs of concept.

References

1. Chedjou, J.C., Kyamakya, K.: Cellular neural networks based local traffic signals control at a junction/intersection. In: Proceeding of the IFAC Conference on Embedded Systems, Computational Intelligence and Telematics in Control. International Federation of Automatic Control (IFAC-2012), Würzburg, pp. 2034–2040 (2012)
2. Gershenson, C., Heylighen, F.: When can we call a system self-organizing? Advances in Artificial Life, pp. 606–614. Springer, Berlin (2003)
3. Floreano, D., Mattiussi, C.: Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies. The MIT Press, Cambridge (2008)
4. Gershenson, C.: Self-organizing traffic lights. Complex Syst. **16**, 29–53 (2005)
5. Webster, F.V.: Traffic Signal Settings. Road Research, vol. 39. Her Majesty's Stationery Office, London (1958)
6. Miller, A.J.: A computer control system for traffic network. In: Proceedings of 2nd International Symposium on Theory of Road Traffic Flow, pp. 201–220 (1963)
7. Hale, D.K.: Traffic Network Study Tool, TRANSYT-7F. United States Version, Mc-Trans Center, University of Florida, Gamesville, Florida 32611-6585 (2006)
8. Akçelik, R.: Time-dependent expressions for delay, stop rate and queue length at traffic signals. Australian Road Research Board, Internal Report, AIR 367-1, Vermont South, Australia (1980 and 2011)
9. Chedjou, J.C., Kyamakya, K., Mathis, W., Moussa, I., Fomethe, A., Fono, A.V.: Chaotic synchronization in ultra-wide-band communication and positioning systems. J. Vib. Acoust. Trans. ASME **130**, 011012-1-12 (2008)
10. Chedjou, J.C., Kana, L.K., Moussa, I., Kyamakya, K., Laurent, A.: Dynamics of a quasiperiodically forced Rayleigh oscillator. J. Dyn. Syst. Meas. Control Trans. ASME **128**, 600–607 (2006)
11. Wong, S.C., Wong, W.T., Leung, C.M., Ton, C.O.: Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control. Transp. Res. Part B **36**, 291–312 (2002)
12. Kimber, R.M., Hollis, E.M.: Traffic queues and delays at road junctions. TRRL Report LR 909, Road Research Laboratory, Crowthorne (1979)
13. Little, J.D.C., Kelson, M.D., Gartner, N.H.: MAXBAND: a program for setting signals on arteries and triangular networks. Transp. Res. Record **796**, 40–46 (1981)

14. Robertson, D.I., Bretherton, R.D.: Optimizing networks of traffic signals in real time - the SCOOT method. IEEE Trans. Veh. Technol. **40**, 11–15 (1991)
15. Laemmer, S., Kori, H., Peters, K., Helbing, D.: Decentralised control of material or traffic flows in networks using phase-synchronisation. Phys. A Stat. Mech. Appl. **363**, 39–47 (2006)
16. Liu, H.X., Oh, J.-S., Recker, W.: Adaptive signal control system with online performance measure for a single intersection. J. Transp. Res. Board **1811**, 131–138 (2002)
17. PLaczek, B.: A traffic model based on fuzzy cellular automata. J. Cell. Autom. **8**, 261–282 (2013)
18. Schadschneider, A., Chowdhury, D., Brockfeld, E., Klauck, K., Santen, L., Zittartz, J.: A new cellular automata model for city traffic. In: Helbing, D., et al. (eds.) Traffic and Granular Flow '99: Social, Traffic, and Granular Dynamics. Springer, Berlin (2000)
19. Zhu, H.B., Ge, H.X., Dai, S.Q.: A density-dependent NaSch model for traffic flow controlled by a traffic light. In: Appert-Rolland, C., et al. (eds.) Traffic and Granular Flow '07, pp. 447–452. Springer, Berlin (2009)
20. Jin, W., Zheng, Y., Li, J.: Microscopic simulation of traffic flow at signalized intersection based on cellular automata. In: Proceedings of the IEEE International Vehicle Electronics Conference IVEC '99, IEEE, pp. 106–109 (1999)
21. He, H.D., Dong, L.Y., Dai, S.Q.: Simulation of traffic flow with traffic light strategies. J. Shanghai Univ. (English Edition) **10**, 189–191 (2006)
22. Brockfeld, E., Barlovic, R., Schadschneider, A., Schreckenberg, M.: Optimizing traffic lights in a cellular automaton model for city traffic. Phys. Rev. E **64**, 056132–056145 (2001)
23. Bazzan, A.L.C.: A distributed approach for the coordination of traffic signal agents. Autonom. Agents Multi-Agent Syst. **10**, 131–164 (2005)
24. Arel, I., Liu, C., Urbanik, T., Kohls, A.G.: Reinforcement learning- based multi-agent system for network traffic signal control. IET Intell. Transp. Syst. **4**, 128–135 (2010)
25. Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D. thesis, Cambridge University, Cambridge, UK (1989)
26. De Oliveira, L.-B., Camponogara, E.: Multi-agent model predictive control of signaling split in urban traffic networks. Transp. Res. **18C**, 120–139 (2010)
27. Cai, C., Wong, C.K., Heydecker, B.G.: Adaptive traffic signal control using approximate dynamic programming. Transp. Res. Part C **17**, 456–474 (2009)
28. Yu, X.-H., Recker, W.W.: Stochastic adaptive control model for traffic signal systems. Transp. Res. **14C**, 263–282 (2006)
29. Laemmer, S., Helbing, D.: Self-control of traffic lights and vehicle flows in urban road networks. J. Stat. Mech. Theory Exp. P04019 (2008)
30. Akbas, A., Ergun, M.: Dynamic traffic signal control using a nonlinear coupled oscillators approach. Can. J. Civ. Eng. **32**, 430–441 (2005)
31. Laemmer, S., Kori, H., Peters, K.: Decentralised control of material or traffic flows in networks using phase synchronisation. Phys. A **363**, 39–47 (2006)
32. Bielefeldt, C., Busch, F.: MOTION - a new on-line traffic signal network control system. In: Proceedings of the 7th International Conference on Road Traffic Monitoring and Control, pp. 55–59 (1994)
33. Strogatz, S.H.: From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. Phys. D **143**, 1–20 (2000)
34. Nishikawa, I., Nakazawa, S., Kita, H.: Area-wide control of traffic signals by a phase model. Trans. Soc. Instrum. Control Eng. **39**, 199–208 (2003)
35. Gartner, N.H., Pooran, F.J., Andrews, C.M.: Implementation of the OPAC adaptive control strategy in a traffic signal network. In: Proceedings of the IEEE Intelligent Transportation Systems Conference (2001)
36. Crabtree M.R., Henderson, I.R.: MOVA Traffic Control Manual, B edition (2006)
37. Nagel, K., Schreckenberg, M.: A cellular automaton model for freeway traffic. Journal de Physique I **2**, 2221–2229 (1992)
38. Spyropoulou, I.: Modelling a signal controlled traffic stream using cellular automata. Transp. Res. Part C Emerg. Technol. **15**, 175–190 (2007)

39. Biham, O., Middleton, A., Levine, D.: Self-organization and a dynamical transition in traffic-flow models. Phys. Rev. A **46**, 6124–6127 (1992)
40. Lo, H.K.: A cell-based traffic control formulation: strategies and benefits of dynamic timing plans. Transp. Sci. **35**, 148–164 (2001)
41. Rosenblueth, D., Gershenson, C.: A model of city traffic based on elementary cellular automata. Complex Syst. **19**, 305–322 (2011)
42. Wainer, G., Davidson, A.: Defining a traffic modeling language using cellular discrete-event abstractions. J. Cell. Autom. **2**(4), 291–343 (2007)
43. Ng, K.M., Reaz, M.B.I., Ali, M.A.M.: A review on the applications of Petri nets in modeling, analysis, and control of urban traffic. IEEE Trans. Intell. Transp. Syst. **14**, 858–870 (2013)
44. Febbraro, A.D., Giglio, D.: On representing signalized urban areas by means of deterministic-timed Petri nets. In: Proceedings of the IEEE Intelligent Transportation Systems Conference, Washington, DC, USA, pp. 372–377 (2004)
45. Febbraro, A.D., Giglio, D.: On adopting a Petri net-based switching modeling system to represent and control urban areas. In: Proceedings of the 8th International Conference on IEEE Intelligent Transportation Systems, Vienna, Austria, pp. 185–190 (2005)
46. Febbraro, A.D., Giglio, D.: Traffic-responsive signaling control through a modular/switching model represented via DTPN. In: Proceedings of the IEEE Intelligent Transportation Systems Conference, Toronto, ON, Canada, pp. 1430–1435 (2006)
47. Febbraro, A.D., Sacco, N., Giglio, D.: On using Petri nets for representing and controlling signalized urban areas: new model and results. In: Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, pp. 1–8 (2009)
48. Badamchizadeh, M.A., Joroughi, M.: Deterministic and stochastic Petri net for urban traffic systems, pp. 364–368. Proceedings of the IACSIT Singapore Conference, Singapore (2010)
49. Makela, M., Lahtinen, J., Ojala, L.: Performance analysis of traffic control systems using stochastic Petri nets, Helsinki University of Technology, Digital Systems Laboratory, Helsinki, Finland, Series B: Technical report 19 (1998)
50. Basile, F., Chiacchio, P., Teta, D.: A hybrid model for real-time simulation of urban traffic. Control Eng. Pract. **20**, 123–137 (2012)
51. Dotoli, M., Fanti, M.P.: An urban traffic network model via coloured timed Petri nets. Control Eng. Pract. **14**, 1213–1229 (2006)
52. Dotoli, M., Fanti, M.P., Iacobellis, G.: Validation of an urban traffic network model using colored timed Petri nets. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 1347–1352 (2005)
53. Kutil, M., Hanzalek, Z.: Light controlled intersection model based on the continuous Petri net. In: Proceedings of the 12th IFAC Symposium on Control in Transportation Systems, pp. 519–525 (2009)
54. Febbraro, A.D., Giglio, D., Sacco, N.: Modular representation of urban traffic systems based on hybrid Petri nets. In: Proceedings of the IEEE Intelligent Transportation Systems Conference, Oakland, CA, USA, pp. 866–871 (2001)
55. Febbraro, A.D., Giglio, D., Sacco, N.: Urban traffic control structure based on hybrid Petri nets. IEEE Trans. Intell. Transp. Syst. **5**, 224–237 (2004)
56. Dotoli, M., Fanti, M.P., Iacobellis, G.: An urban traffic network model by first order hybrid Petri nets. In: Proceedings of the IEEE Conference on SMC, pp. 1929–1934 (2008)
57. Zhang, L., Jia, L.: Intersection traffic system simulation based on hybrid CPN model. In: Proceedings of the 2nd International Conference on Intelligent Computation Technology and Automation, Zhangjiajie, China, pp. 644–647 (2009)
58. Wang, J.C., Jin, C., Deng, Y.: Performance analysis of traffic networks based upon stochastic timed Petri net models. Int. J. Softw. Eng. Knowl. Eng. **10**, 735–757 (2000)
59. Henry, J.J., Farges, J.L., Tuffal, J.: The PRODYN real time traffic algorithm. In: Proceedings of the 4th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Baden-Baden, Germany, pp. 307–312 (1983)
60. Mirchandani, P., Head, L.: A real-time traffic signal control system: architecture, algorithms and analysis. Transp. Res. Part C **9**, 415–432 (2001)

61. Mauro, V., Di Taranto, C.: UTOPIA. In Proceedings of the 2nd IFAC-IFIP-IFORS Symposium on Traffic Control and Transportation Systems, pp. 575–597 (1989)
62. Dinopoulou, V., Diakaki, C., Papageorgiou, M.: Applications of the urban traffic control strategy TUC. Eur. J. Oper. Res. **175**, 1652–1665 (2006)
63. Kuramoto, Y.: Self-entrainment of a population of coupled non-linear oscillators. In: Araki, H. (ed.) International Symposium on Mathematical Problems in Theoretical Physics. Lecture Notes in Physics, vol. 39, pp. 420–422. Springer, New York (1975)
64. Sekiyama, K., Nakanishi, J., Takagawa, I.: Self-organizing control of urban traffic signal networks. In: Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics, vol. 4, pp. 2481–2486 (2001)
65. Kengne, J., Chedjou, J.C., Kenne, G., Kyamakya, K.: Dynamical properties and chaos synchronization of improved Colpitts oscillators. Commun. Nonlinear Sci. Numer. Simul. **17**, 2914–2923 (2012). Elsevier
66. Hoque, S., Imran, A.: Modification of Webster's delay formula under non-lane based heterogeneous road traffic condition. J. Civ. Eng. (IEB) **35**, 81–92 (2007)
67. Boillot, F.: Group-based safety constraint description of an intersection. In: Proceedings of the 9th Mini-EURO Conference on Handling Uncertainty in the Analysis of Traffic and Transportation Systems, Budva, Yugoslavia (2002)
68. Dion, F., Hellinga, B.: A methodology for obtaining signal coordination within a distributed real- time network signal control system with transit priority. In: Proceedings of the 80th Annual Meeting of the Transportation Research Board (2001)
69. Texas Transportation Institute: PASSERTMIII-98 Application and User's Guide (1998)
70. TRB: Highway Capacity Manual. Transportation Research Board, National Research Council, Washington DC (2000)
71. Sunkari, S.R., Engelbrecht, R.J., Balke, K.N.: Evaluation of advanced coordination features in traffic signal controllers. FHWA/TX-05/0-4657-1, FHWA, September (2004)
72. Lowrie, P.R.: The sidney co-ordinated adaptive traffic system: principles, methodology, and algorithms. In: Proceedings of the IEE Conference on Road Traffic Signaling, vol. 207, pp. 67–70 (1982)
73. Brockfeld, E., Barlovic, R., Schadschneider, A.: Optimizing traffic lights in a cellular automaton model for city traffic. Phys. Rev. E **64**, 056132 (2001)

Neural-Network-Based Calibration of Macroscopic Traffic Flow Models

Nkiediel Alain Akwir, Jean Chamberlain Chedjou and Kyandoghere Kyamakya

Abstract This chapter proposes a neural-network-based calibration of macroscopic traffic flow models expressed in the form of nonlinear partial differential equations (PDEs). The calibration scheme/module developed aims at improving both accuracy and stability of the nonlinear PDE models in order to make them more realistic. In essence, the macroscopic nonlinear PDE models of traffic flow (proposed in the literature) are generally inaccurate, unstable, and unlikely to describe the realistic dynamics of traffic flow. To overcome these drawbacks we exploit the artificial neural network paradigm to build a concept called NN-PDE (neural network + PDE solver), which involves the macroscopic model (in the form of a nonlinear partial differential equation) on the one hand, and the calibration scheme/module-based artificial neural network (ANN) on the other. The calibration scheme/module is used to dynamically adjust/optimize all outputs of the nonlinear "PDE" model in order to obtain a realistic set of parameters that could be used by the PDE model to describe the real/realistic dynamics of traffic flow. Overall, specific attention is devoted to some relevant issues related to the modeling concept such as accuracy, stability, and overfitting, just to name a few. These issues generally occur due to the complexity of the PDE model at stake for the sake of an accurate traffic flow model. Finally, some simulation results are shown to demonstrate the effectiveness of the NN-PDE concept developed.

N.A. Akwir (✉) · J.C. Chedjou · K. Kyamakya
Transportation Informatics Group (TIG), Institute of Smart Systems
Technologies, Alpen-Adria University of Klagenfurt, 9020 Klagenfurt, Austria
e-mail: nkiediel.akwir@aau.at

J.C. Chedjou
e-mail: jean.chedjou@aau.at

K. Kyamakya
e-mail: kyandoghere.kyamakya@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_7

1 Introduction

1.1 Background of Traffic Flow Modeling

Transportation has always been a vector of economic development of society and thereby contributes to improvement in the quality of life. However, due to the growth of population density and urbanization, traffic congestion is increasing worldwide. The transportation-related socioeconomic and environmental impacts such as increasing travel time, fuel consumption, and pollution have been given a tremendous attention by decision-makers in recent decades [1].

Developing new control and optimization strategies to better manage vehicular traffic [2–4] can be viewed as the best realistic option to deal with issues related to congestion, pollution, fuel consumption, etc., since building new roads is very expensive and could be an unrealistic solution, since the availability of land is not always ensured [5].

Different modeling concepts have been developed in order to capture traffic dynamics: (1) mathematical models (in the form of ODEs or PDEs) [6–8], (2) graphical models (graphs and Petri nets) [9–12], and (3) software tools (Synchro, Vissim, and Visum, etc.) [13–15], for macroscopic and microscopic traffic simulation.

The mathematical modeling of traffic flow has always been a challenging issue in science and engineering due to the striking and complex nature of the dynamical behavior of traffic flows (e.g., nonlinearity, hysteresis, stiffness, saturation, oversaturation, jam, shock waves, rarefaction waves, stop-and-go waves, platoons, bottleneck, chaos, just to name a few) [16–19].

The use of ordinary and/or partial differential equations for traffic flow modeling depends on the level of detail (i.e., macroscopic, microscopic, and mesoscopic levels) [20]. For instance, microscopic traffic flow models distinguish and trace the behavior of each individual vehicle (e.g., car following model, lane change model) [20]. The use of ordinary differential equations for microscopic traffic flow modeling requires a huge amount of data for the optimization of corresponding ODE parameters. Further, the use of ODE models for microscopic traffic simulation is generally time-consuming (i.e., very slow). Macroscopic models aggregate vehicles allowing a description of traffic flow as a continuum (i.e., a global view that does not distinguish vehicles). The related macroscopic models are expressed in the form of partial differential equations, in which dependent variables are generally obtained as the average of the three fundamental traffic parameters (i.e., mean flow, mean density, mean speed). Macroscopic models are generally less costly computationally than microscopic models. However, the accuracy of macroscopic models is an issue when compared with the good accuracy of microscopic models. Mesoscopic models generally encompass both microscopic and macroscopic models and usually involve human intelligence (e.g., driver behavior) and/or artificial intelligence (e.g., sensors) [21, 22].

The mathematical traffic flow models provided in the literature (see state of the art on traffic mathematical modeling) are all derived using a series of specific assumptions (e.g., see analogy with fluid dynamics [23]; see also analogy with gas kinetics [24]). However, these assumptions often do not consider all practical constraints (i.e., those faced on the road network when one is considering real traffic dynamics). Thus, the available mathematical models reveal only a partial view of the reality (i.e., the real traffic dynamics and related traffic phenomena and scenarios). This underscores and justifies the statement that most traffic models presented in the literature are not realistic enough. Further, the literature does not provide a mathematical model that can be used as a general framework for traffic modeling in considering all practical realistic constraints. Among the great number of mathematical models published so far, none of them is likely to simultaneously describe traffic phenomena (all practical realistic constraints) such as shock waves, rarefaction waves, stop-and-go waves, platoons, bottlenecks, and jams, just to name a few. The cited phenomena are generally described by different types of PDEs (e.g., linear, nonlinear, convex, concave, hyperbolic, nonhyperbolic).

A seminal mathematical model for traffic flow was proposed by Lighthill–Whitham (1955) [6] and Richards (1956) [25], the so-called LWR model, expressed in the form of a partial differential equation. This model, also known as a first-order model, is based on the continuity equation from compressible dynamics theory, which expresses the conservation of a flowing quantity from one point to another. Despite the fact that the LWR model can reproduce some phenomena, it is based on a number of assumptions that are not always realistic (e.g., constant speed, no overtaking, no ramp). To solve this issue, Payne [26], Ross (1988) [27], and Del Castillo (1993) [28] have proposed mathematical models of second order, which take into account the speed dynamics. These latter models were later improved by Zhang [8], Jiang et al. [29], and Gupta et al. [7]. Even though these PDE-based macroscopic traffic models can reproduce the spatiotemporal behavior of traffic flow on a road segment relatively well, they suffer mainly from a lack of accuracy, and they are relatively slow (although they remain significantly much faster (for traffic simulation) than their microscopic counterparts). Compared to microscopic traffic simulators, the macroscopic ones are less accurate. Essentially, in their raw forms, none of the existing macroscopic traffic models and simulation concepts fulfills the necessary requirements (e.g., robustness, accuracy, realism, ultrafast simulation) for online traffic simulation. The calibration of the PDE models for traffic flow is of importance in addressing the requirement of online traffic simulation (i.e., fast computational speed and a practical constraints aware accuracy).

1.2 Background of Model Calibration

Calibration is a process aiming at fitting a given model using a set of empirical data. This process is very important in traffic modeling, since a traffic mathematical model generally involves a number of parameters that must be adjusted for some specific scenarios. Otherwise, there is often a big gap (errors) between theoretical models and experimental data.

Several traffic flow calibration algorithms have been proposed based on the optimization process [30]. Nelder and Mead (1965) [31] proposed an algorithm for multidimensional unconstrained optimization. This algorithm was proven to be suitable for problems modeled by objective functions (or cost function) expressed in nonlinear, discontinuous, or stochastic forms. Such an optimization algorithm requires a large number of terations to achieve convergence, however, without significant improvement in some cases; further, several updates of the optimization algorithm are also needed. Genetic algorithms (GA) that have been proposed by Golberg (1989) and Holland (1992) [32, 33] are part of the largest class of evolutionary algorithms. Genetic algorithms mimic evolution in biological processes using natural selection, mutation, and crossover techniques [32, 33]. Genetic algorithms are appropriate for various optimization problems such as discontinuous objective functions and nondifferentiable, stochastic, or highly nonlinear objective functions. Despite the fact that this algorithm is flexible in searching complex solution spaces, each iteration requires as many cost function evaluations as the population size; thus the algorithm is computationally costly (i.e., slow). Another very important algorithm is the cross entropy method proposed by Rubinstein and Kroese (2004) [34] and de Boer et al. (2005) [35]. This algorithm is applicable to discontinuous, nondifferentiable, or highly nonlinear objective functions. The main drawbacks of this algorithm are the high computational cost and the slow convergence, since it requires as many cost function evaluations as the size of the population.

The few optimization algorithms underscored so far (see [22]–[27]) have been used intensively during the past decades, in classical calibration processes and/or calibration modules. Generally, despite the fact that those algorithms may be applied to complex problems (e.g., discontinuous, nondifferentiable, or highly nonlinear cost functions), they have some inherent drawbacks, which are mostly related to the computational cost (i.e., too time-consuming) and lack of convergence.

1.3 Key Contribution and Organization

A novel calibration concept for nonlinear partial differential equation (PDE) models used for macroscopic traffic modeling is developed here. The concept developed uses artificial neural networks (ANN) for the calibration of partial differential equation (PDE) models of traffic flow. The main advantage of the concept developed is the possibility of overcoming some drawbacks of classical (or traditional) calibration concepts (e.g., flexibility, universality, robustness, and accuracy). Another advantage of the calibration concept developed is the straightforward and easy applicability to traffic flow modeling and an online simulation capability. The traffic models (PDEs) and the online simulation (using PDEs) are essential in analyzing, understanding, depicting, and controlling the realistic and spatiotemporal evolution of traffic flow dynamics while considering practical operational constraints (e.g., real-time constraint, realistic nature of solutions regarding real traffic constraints, correctness of solutions, stability and robustness). In essence, the PDE model helps to obtain the spatiotemporal evolution of the behavior of the traffic flow, while the calibration helps to improve both accuracy and correctness of the related PDE models.

The remainder of this chapter is organized as follows. Section 2 is concerned with the mathematical modeling of traffic flow. A PDE model is proposed for the traffic flow scenarios under investigation. Section 3 presents the calibration concept developed. Full details of all analytical steps toward development of the calibration module are provided. The calibration module is based on artificial neural networks (ANN). Section 4 develops a novel concept to which we have assigned the acronym NN-PDE. This concept combines PDE and ANN paradigms. The PDE model considers the spatiotemporal behavior of traffic flow, while the ANN paradigm is used for calibration to improve the accuracy of the PDE model. The offline ANN training process is performed using data from quasireal traffic provided by a relevant popular simulator tool called VISSIM. VISSIM is an accurate microscopic traffic simulation tool, which is able to configure several traffic scenarios while taking into account several practical conditions. Section 5 is devoted to numerical simulation. A case study of a concrete traffic scenario modeled by PDE is considered. The calibration of the resulting PDE model is conducted, and the results of the calibration obtained are discussed. Finally, the advantage of the calibration is clearly demonstrated. This chapter ends with some concluding remarks and an outlook (Sect. 6).

2 PDE Models for Macroscopic Traffic Flow

Several traffic flow models have been proposed in the literature to express the dynamics of traffic flow on arterial roads [6, 8, 26]. These models are expressed in the form of continuous differential equations and involve the three fundamental parameters of traffic, namely (1) the flow (number of cars crossing a section of the road per unit of time), (2) the density (number of cars per unit of length), and (3) the speed (rate of variation of the position over time).

As already stated above, the past decades have witnessed the derivation of several mathematical models for the analysis of traffic flow at the macroscopic level. Some interesting mathematical models corresponding to specific scenarios are provided by Eqs. (6), (8), and (9).

Traffic Scenario 1: *Here no overtaking and no ramp (see Fig. 1).*

This scenario shows a simple case of traffic flow in a single lane. It has been shown (see [6]) that this scenario is modeled mathematically by Eq. (1). The assumption made in [36] to obtain the model in Eq. (1) is the homogeneous nature of the traffic flow (i.e., all cars are assumed to be of the same type).

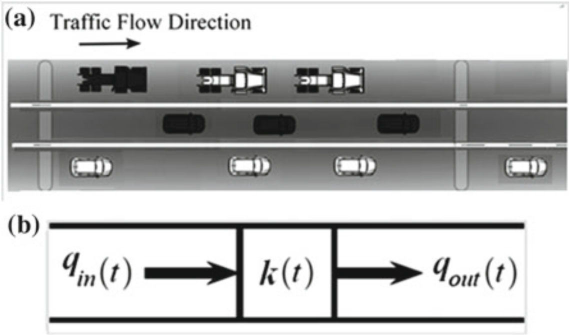


Fig. 1 General representation of the traffic flow on a road segment of finite length. **a** Illustration of the traffic flow without overtaking. **b** Synoptic representation of the traffic flow for the sake of modeling. (***source***: [37])

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{1}$$

In Eq. (1), the dependent variables are represented by k and q. These two variables stand for the traffic density k and traffic flow q on a road segment. The independent variables x and t are the spatial and temporal dimensions, respectively. For the sake of solvability of Eq. (1), a second analytical expression is envisaged in Eq. (2) in order to express the relationship among the three fundamental parameters

$$q = k \cdot u. \tag{2}$$

Equations (1) and (2) cannot be used to express the spatiotemporal evolution of the three fundamental parameters of traffic flow (k, q, and u). This justifies the choice of additional relationships expressing the empirical Greenshield's model (see Eqs. (3), (4), and (5)). Equations (3), (4), and (5) are used to obtain the fundamental diagrams shown in Fig. 2. This diagram is a three-dimensional (3D) representation of the interaction between the fundamental parameters of traffic flow. Figure 2 clearly emphasizes that the speed is high at low traffic density (e.g., weak interaction between vehicles), while at high traffic density (e.g., strong interaction between vehicles), the speed is low. This observation is important, since it could be used to express the monotonicity of the speed of vehicles on roads. This is important in depicting some phenomena such as shock waves, stop-and-go waves, and rarefaction waves, just to name a few.

Many mathematical expressions have been proposed in the literature to represent the 3D relationship between the fundamental parameters of traffic flow empirically. Some classical and commonly used empirical models are Greenshield's [38], Greenberg [39], Underwood [40], Wang et al. [41]. These empirical models are generally obtained through the use of data history, which corresponds to specific measurements of the three fundamental parameters: flow, speed, and density.

The present chapter exploits Greenshield's model expressed in a 3D system of coordinates as follows:

$$u = u(k) = u_f - \frac{u_f}{k_j}k, \tag{3}$$

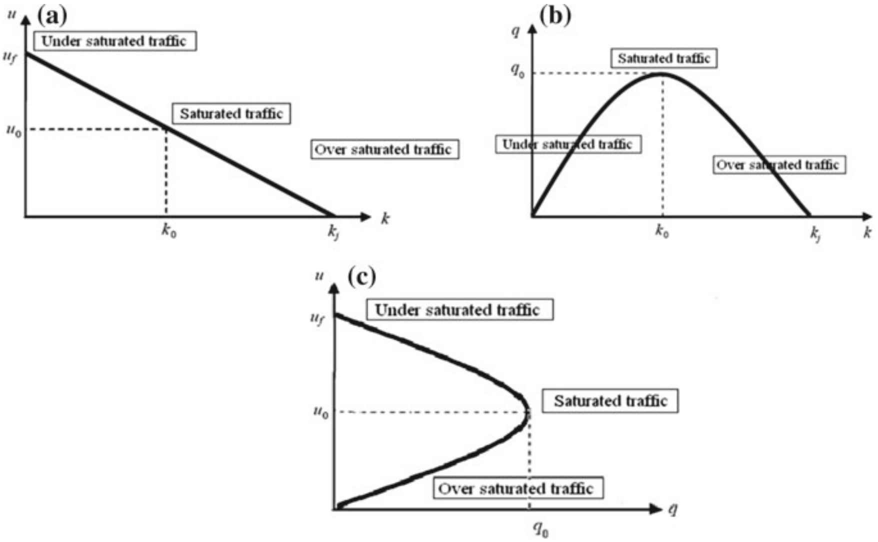


Fig. 2 Fundamental diagrams: **a** Speed versus density. **b** Flow versus density. **c** Speed versus flow

$$q = q(k) = ku_f - \frac{u_f}{k_j}k^2, \tag{4}$$

$$u = u(q) \Rightarrow u^2 = uu_f - \frac{u_f}{k_j}q. \tag{5}$$

The quantity u_f stands for the free flow speed. The free flow here expresses the possibility of driving without any interaction between vehicles. Thus, the driver could choose the speed (denoted here by u_f) at his convenience (it is the maximum allowed speed according to legal speed limit on a particular road segment). Equations (3), (4), and (5) are used to plot the fundamental diagrams of speed versus density in Fig. 2a, flow versus density in Fig. 2b, and speed versus flow in Fig. 2c. The diagrams in Fig. 2 reveal the three possible traffic states (i.e., undersaturation, saturation, and oversaturation). The quantities k_0 and u_0 are the density (critical) and speed at the capacity of the road. The quantity k_j stands for jam density, and the quantity u_f is the free flow speed, as mentioned before.

The classical and seminal model for traffic flow, the so-called LWR model, proposed by Lighthill, Whitham, and Richard [25] corresponds to Eq. (6), obtained by combining Eqs. (1) and (4):

$$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \\ q = q(k) = ku_f - \frac{u_f}{k_j}k^2. \end{cases} \tag{6}$$

Equation (6) is a first-order quasilinear hyperbolic partial differential equation in the dependent variable $k(x, t)$. The nonconservative form of Eq. (6) is given as follows:

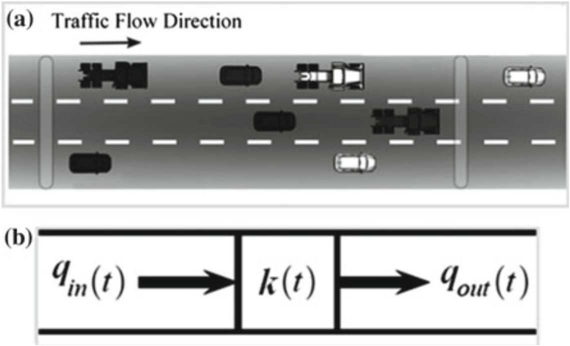


Fig. 3 General representation of the traffic flow on a road segment of finite length. **a** Illustration of the traffic flow with overtaking and without ramps and detectors. **b** Synoptic representation of the traffic flow for the sake of modeling. (***source***: [37])

$$\frac{\partial k}{\partial t} + (u_f - 2\frac{u_f}{k_j}k)\frac{\partial k}{\partial x} = 0. \tag{7}$$

Equation (7) is the traffic flow model that is considered in this chapter.

Traffic Scenario 2: *Here there is overtaking but no ramp (see Fig. 3).*

This scenario shows a simple case of traffic flow in multiple lanes with overtaking. It has been shown (see [42]) that this scenario is modeled mathematically by Eq. (8). Equation (8) is obtained by assuming a homogeneous traffic flow:

$$\begin{cases} \dfrac{\partial k_1}{\partial t} + \dfrac{\partial q_1}{\partial x} = \dfrac{k_2}{T_2^1} - \dfrac{k_1}{T_1^2}, \\ \dfrac{\partial k_j}{\partial t} + \dfrac{\partial q_j}{\partial x} = \dfrac{k_{j-1}}{T_{j-1}^j} - \dfrac{k_j}{T_j^{j-1}} + \dfrac{k_{j+1}}{T_{j+1}^j} - \dfrac{k_j}{T_j^{j+1}}, \\ \dfrac{\partial k_N}{\partial t} + \dfrac{\partial q_N}{\partial x} = \dfrac{k_{N-1}}{T_{N-1}^N} - \dfrac{k_N}{T_N^{N-1}}. \end{cases} \tag{8}$$

In (8), the subscripts $J = 2, \ldots, N-1$ refer to the interior lanes, and the subscripts 1 and N refer to the extreme lanes; $T_j^k = T_j^k(k_j, k_k)$ is the vehicle transition rate from two neighboring lanes (j to lane k).

Traffic Scenario 3: *Here there are both overtaking and ramps (see Fig. 4)* This scenario shows a simple case of traffic flow on multiple lanes with overtaking. It has been shown (see [43]) that this scenario is modeled mathematically by Eq. (9). The assumption of homogeneous traffic flow has been made to obtain Eq. (9):

$$\frac{\partial k}{\partial t} + \frac{\partial ku}{\partial x} = \pm\rho(x, t). \tag{9}$$

The quantity $\rho(x, t) \geq 0$ corresponds the rate of vehicles entering (+) the highway or leaving (−) the highway through ramps (respectively entrance and exit ramps).

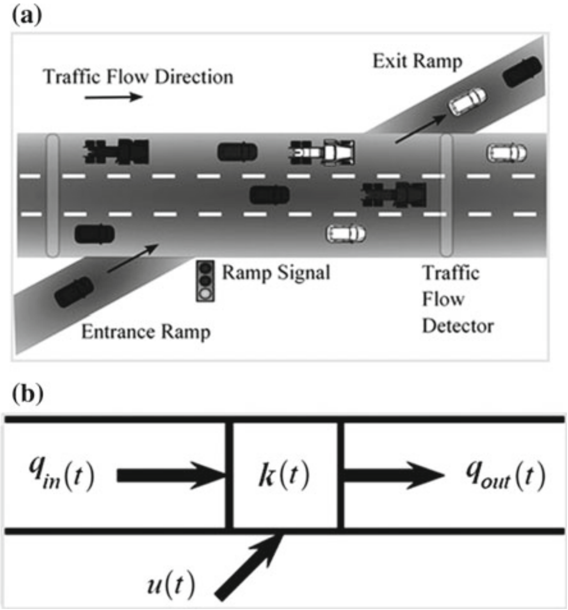


Fig. 4 General representation of traffic flow on a road segment of finite length. **a** Illustration of traffic flow and the influence of signal detectors and ramps. **b** Synoptic representation of the traffic flow for the sake of modeling. (***source***: [37])

3 Calibration Concept

3.1 General Principle

The mathematical models proposed by scientists and engineers may reflect the dynamics of systems and phenomena. However, some of these models might be inaccurate and unlikely to reflect the real dynamics of specific scenarios, especially when the model involves a large number of parameters. Therefore, without a careful prior model calibration, it would not be possible to rely on the simulated results. In order to ensure the validity of the models, they must ideally be adequately calibrated for the full range of conditions and scenarios of relevance. Several models have been developed for calibrating either macroscopic or microscopic models (see [45] and references therein).

Calibration is an optimization process involving techniques such as the genetic and memetic algorithm [44] and the cross-entropy method [45], just to name a few. The overall optimization procedure consists in estimating some parameters for minimizing a cost function that can be made, for instance, by an error (or discrepancy) between the mathematical model result and real data obtained from field measurements (or obtained from a realistic simulator like the VISSIM Simulator). This is not a trivial task, since the system of equations may be highly nonlinear in both parameters and state variables.

The parameter estimation problem can be formulated as a nonlinear least-squares output error problem that aims at minimizing the discrepancy between the model calculations and the real traffic data using the following cost function:

$$E(w) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}[k(n) - \bar{k}(n)]^2}, \tag{10}$$

where $E(w)$ is the root mean squared error (RMSE). In the calibration process, this error is minimized by finding the appropriate value of the weight w; $k(n)$ is the set of data resulting from the model, and $\bar{k}(n)$ stands for real traffic data from the field (or VISSIM Simulator).

3.2 Calibration Involving Artificial Neural Networks

An artificial neural network (ANN) [46] is a computational model implemented as a computer program aiming at emulating the key properties and operations of biological neural networks. ANNs are used to model unknown or unspecified functional relationships between the input and output of a "black box" system. In order to apply such a procedure to decision problems, a key requirement is ANN training to minimize the discrepancy between modeled and measured system output. Due to its principle or properties (e.g., training capability, usage phase, redimensioning of the network, etc.), the ANN paradigm is a good candidate for calibration tasks.

3.3 Mathematical Model of a Neural Network

A single neuron model consists of inputs, weights, an activation function, and outputs. The mathematical model of a single neuron is expressed as follows:

$$Y = f(\sum_{i} w_i x_i + b), \tag{11}$$

where the set of inputs is $X = [x_0, x_1, x_2, \ldots, x_n]$, the set of weights is $W = [w_0, w_1, w_2, w_3, \ldots, w_n]$, the output signal is $Y = (y_1, y_2, y_3, \ldots, y_n)$, the activation function is given by $f = f(W, X)$, and b is the bias.

An artificial neural network consists of neurons connected together. Several architectures of neural networks have been proposed in the literature, such as multilayer perceptron neural networks [47], recurrent neural networks [48], and Hopfield neural network [49], just to name few. In this chapter we exploit the multilayer perceptron (MLP), which consists of several layers of interconnected perceptrons. Several training methods such as the back-propagation algorithm [50] and the Levenberg–Marquardt algorithm [51, 52] are used to train this type of ANN.

Considering that we have $\bar{Y} = (\bar{y_1}, \bar{y_1}, \ldots, \bar{y_n})$ as target data (in our case they are real traffic flow data), the training is done using an optimization process iteratively in order minimize the discrepancy D between the ANN outputs and the target data:

$$minimize_w D(Y, \bar{Y}, w). \tag{12}$$

At the end of the optimization procedure, a neural network is obtained with new weight values $W^* = [w_0^*, w_1^*, w_2^*, w_3^*, \ldots, w_n^*]$ that minimize the discrepancy expressed by a suitable norm function:

$$D(Y, \bar{Y}) = \|Y - \bar{Y}\|. \tag{13}$$

Once the weight values are known, the ANN can be used to perform various tasks related to the assessment of further input data (i.e., the so-called test data).

4 The NN-PDE Concept

The NN-PDE concept is a combination of a PDE-based model of traffic flow and a calibration based on an artificial neural network. We use the PDE model, which has the capability to represent the traffic flow in the spatiotemporal domain. However, as already mentioned, the PDE models are generally inaccurate, and the calibration carried out aims at proposing a new and accurate model to which we have assigned the acronym NN-PDE. A calibration is exploited to adjust the result with reality (real traffic data), resulting in increasing the accuracy of the model. A neural network module for calibration is built offline by exploiting real traffic data (or quasireal data "from a microscopic simulator") and the data obtained from the PDE model, which takes into account some contextual information related to a specific traffic flow scenario. Let us note that real traffic data are obtained using VISSIM, which is an accurate microscopic traffic simulator that can propose and configure several scenarios. The block diagram in Fig. 5 illustrates the NN-PDE concept.

Block (a) in Fig. 5 is the PDE solver. The PDE model considered can be solved using some numerical methods such as the method of lines (MOL), finite difference methods (FDM), or CNN (cellular neural networks) [53]. The initial and boundary conditions are set up as inputs according to a given scenario. Context information (free flow speed, jam density, etc.) are also provided in order to fit with specific scenarios for a given road segment. The result obtained expresses the spatiotemporal evolution of the three fundamental parameters of traffic flow (i.e., density $k(x, t)$, speed $u(x, t)$, and flow $q(x, t)$). After solving, the models are further processed in order to fit with the input of the calibration module. Block (b) in Fig. 5 is the calibration module based on a neural network. The multilayer perceptron (MLP) architecture is used to build this module.

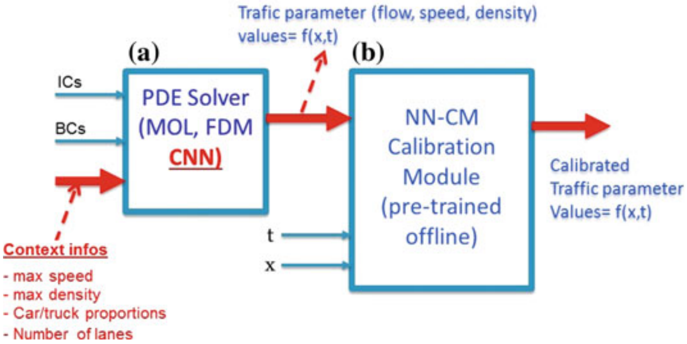

Trafic parameter (flow, speed, density)
values= f(x,t)
(a)
(b)
ICs
BCs
PDE Solver (MOL, FDM CNN)
NN-CM Calibration Module (pre-trained offline)
Calibrated Traffic parameter Values= f(x,t)
t
x
Context infos
- max speed
- max density
- Car/truck proportions
- Number of lanes

Fig. 5 The NN-PDE concept. **a** The PDE solver (ICs: initial conditions and BCs: boundary conditions). **b** The calibration module

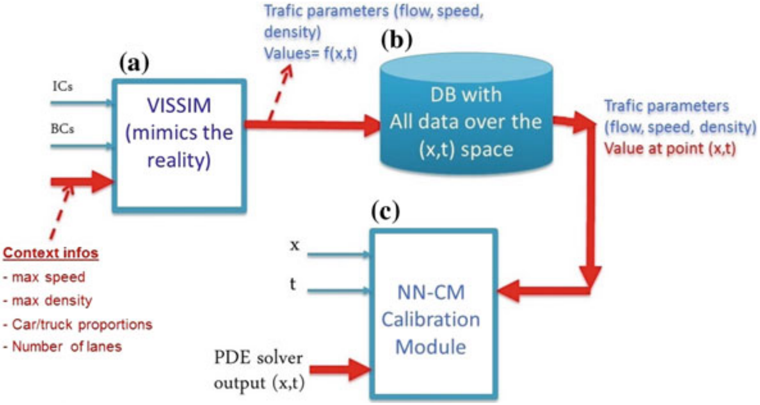


Fig. 6 Offline training of the calibration module

As mentioned before, the training of the calibration module is done offline. The block diagram in Fig. 6 shows the offline training procedure.

Block (a) in Fig. 6 is the traffic simulator tool. We use VISSIM, as mentioned before, to obtain quasireal traffic data. These data are processed in order to fit the calibration module and are stored in the database (see block (b)). The calibration module in block (c) is trained offline by considering as input the data from the PDE solver (taking into account different scenarios) and target (corresponding prestored) data from the database (real traffic data).

5 Simulation Results and Their Commenting

5.1 Numerical Schemes

We consider a scenario consisting of a single-lane model without ramp, as illustrated in Fig. 1, for numerical simulation. This scenario is modeled mathematically by Eq. (7). Applying the finite difference method (e.g., Lax–Friedrichs scheme) to Eq. (7), we obtain the following difference equation:

$$k_i^{n+1} = \left(\frac{k_{i+1}^n + k_{i-1}^n}{2}\right) - \frac{\Delta t}{2\Delta x}\left[\left(u_f - 2\frac{u_f}{k_j}\left(\frac{k_{i+1}^n + k_{i-1}^n}{2}\right)\right)\left(k_{i+1}^n - k_{i-1}^n\right)\right], \tag{14}$$

where the index i represents the road section and the index n denotes the discrete time. The stability condition of the numerical scheme is given by (see [54])

$$u_f\frac{\Delta t}{\Delta x} \leq 1; \quad k_j = p[\max(k_0(x_i))]; \quad p \geq 2. \tag{15}$$

The condition (15) is used during numerical simulation in order to guarantee the stability of solutions. Thus the parameters of Eq. (7) are chosen according to condition (15). During our various numerical simulations, several sets of parameters are envisaged, each of which encompasses the parameters Δt, Δx, u_f, k_j, and k_0 (see cases 1, 2, 3, and 4 in Sect. 5.3). In all four cases, the numerical solutions of Eq. (7) are valid if the condition (15) is satisfied. Otherwise, the solutions are not valid for the scenario in Fig. 1.

5.2 Neural Network Module for Calibration

The neural network module used for calibration encompasses the following input and output variables:

- $k(x,t)$ is the traffic flow obtained as a solution from the PDE model in Eq. (7) in the form of an MxN matrix; M corresponds to the number of iterations in the time domain, while the number of iterations in the space domain corresponds to N; $k(x,t)$ is transformed into a row vector of size $(M * N)$ and is further used as the first input of the neural network (See Fig. 7). The second input corresponds to the vector t of size M, while the size of the third input corresponding to vector x is N. An appropriate use of the neural network in (Fig. 7) is recommended by choosing the inputs 2 and 3 of equal size. This is the fundamental condition for all inputs of the neural network (inputs 1, 2, and 3) to be of the same size.
- $\bar{k}(x,t)$ is the traffic flow data obtained from the VISSIM simulator; $\bar{k}(x,t)$ is transformed into a row vector of size $(M * N)$ and is further used as the target

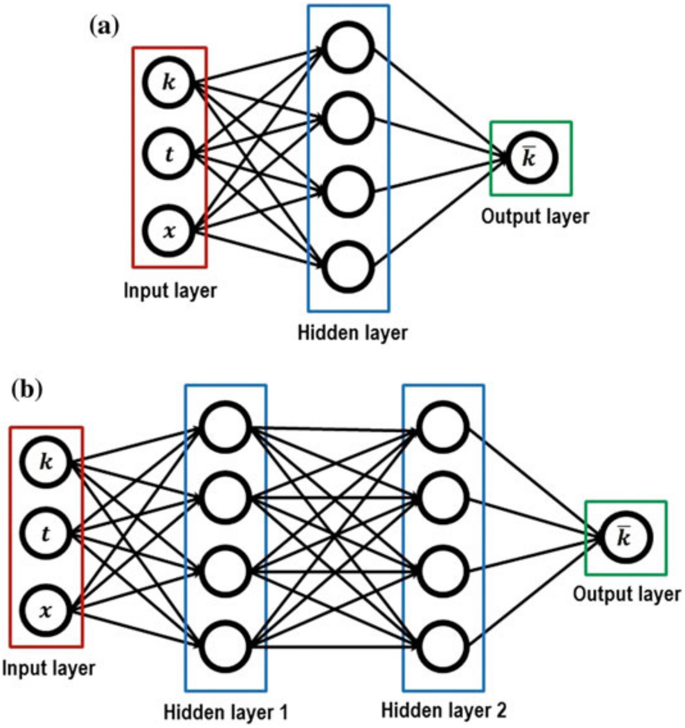


Fig. 7 The one- and two-layer neural networks used for training

solution during the training phase of the neural network. The target solution and the three inputs are of equal size.

- $K(x, t)$ is the result obtained after calibration. This corresponds to the result of the new system developed in this chapter to which we have assigned the acronym of NN-PDE. The data provided by the NN-PDE are in vector form of size $(M * N)$. This vector is of the same size as x, t, k, and $\bar{k}(x, t)$. This is an important condition to ensure that the training procedure of the neural network is effective.

Another important condition for the training procedure of neural networks is concerned with the number of layers and the number of perceptrons per layer. In Fig. 7, the number of layers and/or the number of perceptrons is monitored to increase the accuracy and avoid overfitting as well. During the training process conducted in this work, the choice of both number of layers and perceptrons is based on trial-and-error. It has been observed that increasing the number of perceptrons per layer significantly improves the accuracy of the training process. However, the main problem encountered in increasing the number of perceptrons was the memory consumption and the overfitting observed when the number of perceptrons per layer exceeds 100. It has also been observed that increasing the number of hidden layers leads to an improvement in the accuracy of the training process. However, it has been observed that the accuracy rate remains constant beyond 10 hidden layers. The next section (see Sect. 5.3) is concerned with the presentation of the results obtained using the following methods: (a) numerical solution of the original PDE in Eq. (7) to express the traffic flow dynamics corresponding to the scenario in Fig. 1; (b) solution of the same scenario provided by VISSIM; (c) solution of the same scenario provided by the new concept NN-PDE developed in this chapter. Overall, a comparison between solutions (a) and (b) has led to a significant difference. This difference can be explained by the fact that the theoretical model (Eq. (7)) does not express reality. Thus, the NN-PDE developed is a new mathematical model that is much closer to reality. This statement is justified in Sect. 5.3 through a benchmarking between NN-PDE and VISSIM.

5.3 Results and Comments

5.3.1 Original PDE Versus VISSIM

We apply the finite difference method to the original PDE in Eq. (7) to obtain the discrete form in Eq. (14). This form is further solved using MATLAB to obtain the numerical solution of the original PDE. This solution (see Fig. 8a) expresses the traffic flow dynamics in Fig. 1. We also use VISSIM to simulate the traffic flow dynamics in Fig. 1, and the solution obtained is depicted in Fig. 8b.

Using the set of parameters $\Delta x = 10\,\text{m}$; $\Delta t = 1\,\text{s}$; $u_f = 8.3\,\text{m/s}$; $k_j = 0.160$ Veh/m; and $k_0 = 0.08$ Veh/m, the solution of the original PDE is obtained in Fig. 8a. The solution in Fig. 8b shows the simulation of the same scenario in VISSIM. It can be observed from Fig. 8a, b that there is a significant divergence (see Fig. 8c) between the solutions provided by the two methods (i.e., Original PDE and VISSIM). The divergence in Fig. 8c corresponds to a normalized root mean square error (NRMSE) of 28.02%. This value of the NRMSE justifies the need of calibration in order to reduce the NRMSE. The calibration in this context consists in deriving a new model (to which we have assigned the acronym NN-PDE) that must be able to provide results similar to those obtained using VISSIM. Here VISSIM is considered the target solution, because it expresses the traffic dynamics closer to reality.

5.3.2 NN-PDE Versus VISSIM

The NN-PDE model in Fig. 5 is obtained by combining the original PDE with a neural network. The NN-PDE model is further implemented in MATLAB. Using NN-PDE, several simulation results are obtained for different configurations of neural network. These configurations are obtained by monitoring the number of hidden layers and the perceptrons involved. Further simulation results are obtained for two different values of the free flow speed (i.e., $u_f = 30$ and $u_f = 50\,\text{km/h}$).

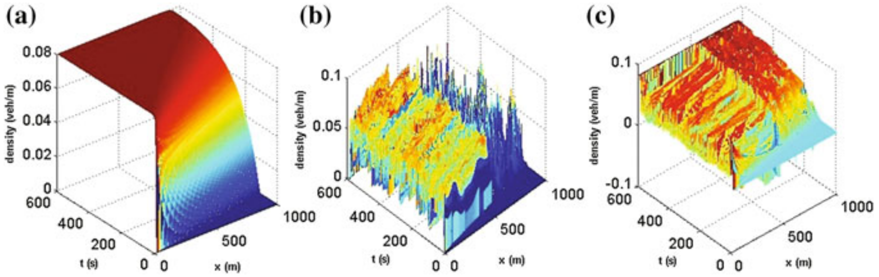


Fig. 8 Simulation results. **a** PDE results. **b** Vissim results. **c** Error (PDE-Vissim). The NRMSE corresponds to 20%

Case-1: We use the following parameters of the NN-PDE: value of the free flow speed ($\mathbf{u_f = 30\,km/h}$), neural network architecture (multilayer perceptron (MLP)), **1** hidden layer involving **50** perceptrons.

The simulation results obtained are illustrated in Fig. 9a–f. The results in Fig. 9a–c correspond to the situation without calibration shown in Fig. 8. In contrast, the results in Fig. 9d–f correspond to the situation with calibration. These results obtained using NN-PDE are compared with the results obtained using VISSIM. The outcome of comparison is depicted in Fig. 9f. Using this figure, we see that the NRMSE calculated corresponds to 6.55%. This value of the NRMSE obtained using NN-PDE shows a significant improvement when compared with the previous value of the NRMSE (28.02%) obtained without calibration (see Fig. 8).

Case-2: We use the following parameters of NN-PDE: value of the free flow speed ($\mathbf{u_f = 30\,km/h}$), neural network architecture (MLP), **10** hidden layers involving **20** perceptrons.

The simulation results obtained are illustrated in Fig. 10a–f. The results in Fig. 10a–c correspond to the situation without calibration shown in Fig. 8. In contrast, the results in Fig. 10d–f correspond to the situation with calibration. These results obtained using NN-PDE are compared with the results obtained using VISSIM. The outcome of the comparison is depicted in Fig. 10f. Using this figure, we see that the NRMSE calculated corresponds to 2.21%. This value of the NRMSE obtained using NN-PDE shows a significant improvement when compared with the previous value of the NRMSE (31.51%) obtained without calibration (see Fig. 8).

Case-3: We use the following parameters of the NN-PDE: value of the free flow speed ($\mathbf{u_f = 50\,km/h}$), neural network architecture (MLP), **1** hidden layer involving **50** perceptrons.

The simulation results obtained are illustrated in Fig. 11a–f. The results in Fig. 11a–c correspond to the situation without calibration shown in Fig. 8. In contrast, the results in Fig. 11d–f correspond to the situation with calibration. These results obtained using NN-PDE are compared with the results obtained using VISSIM. The outcome of the comparison is depicted in Fig. 11f. Using this figure, we see that the NRMSE calculated corresponds to 6.39%. This value of the NRMSE obtained using

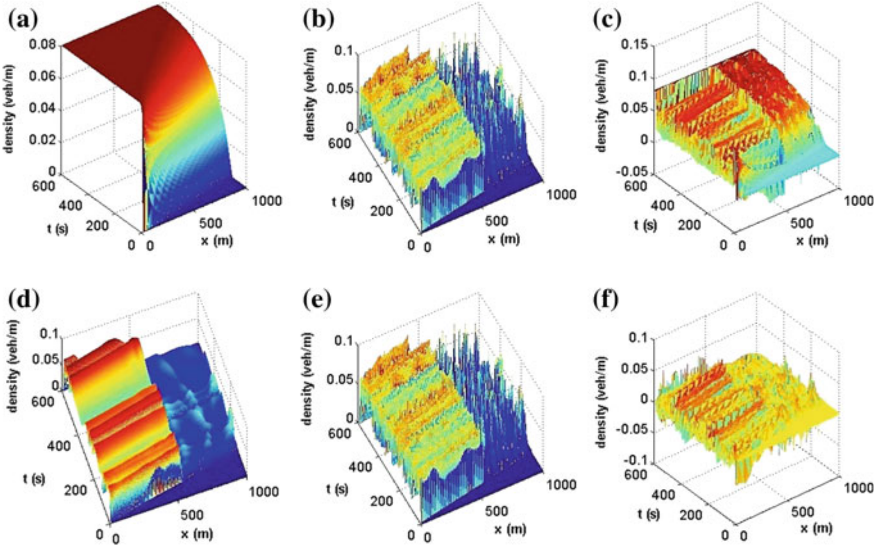


Fig. 9 Simulation results of Configuration-1: MLP, 1 layer, 50 neurons: **a** PDE model (input). **b** Vissim data (targets). **c** Initial error (PDE-Vissim). **d** Calibrated result. **e** Vissim data (targets). **f** Final error (calibrated result-Vissim). The NRMSE corresponds to 6.55%

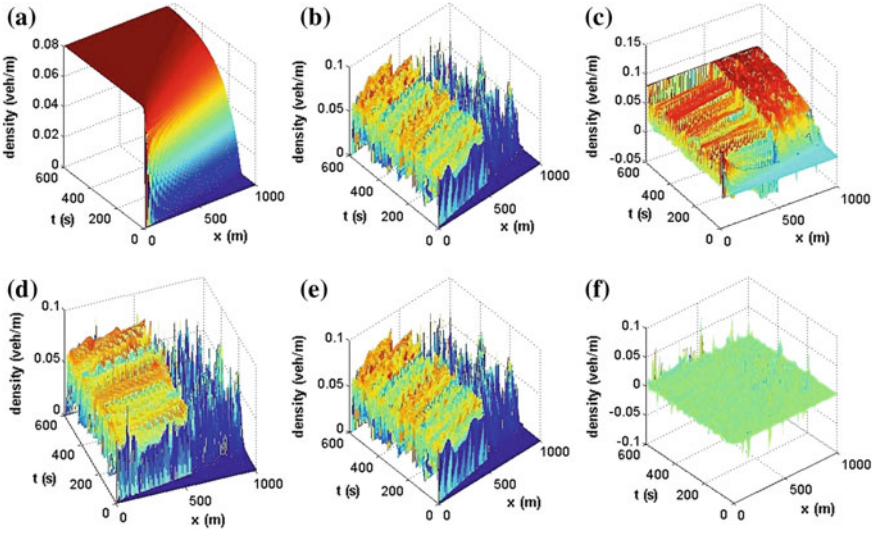


Fig. 10 Simulation results of Configuration-2: MLP, 10 layers, 20 neurons per layer: **a** PDE model (input). **b** Vissim data (targets). **c** Initial error (PDE-Vissim). **d** Calibrated result. **e** Vissim data (targets). **f** Final error (calibrated result-Vissim). The NRMSE corresponds to 2.21%

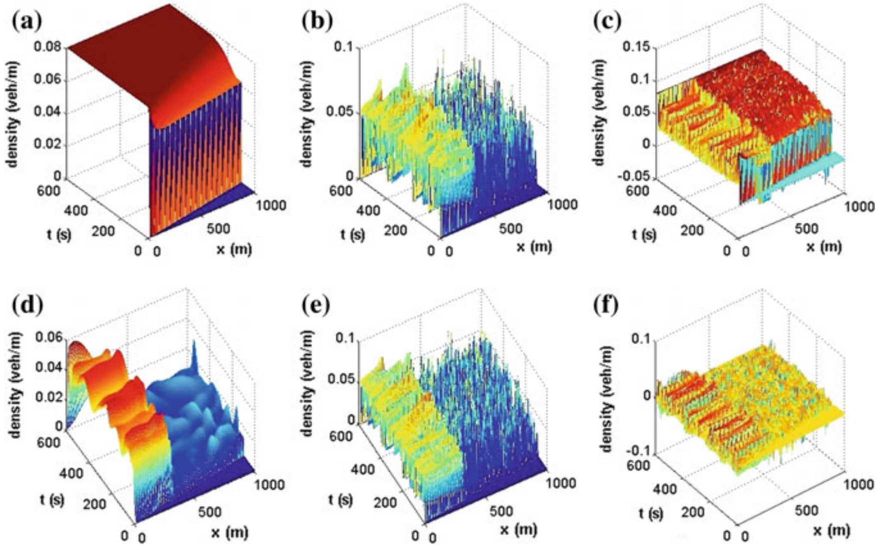


Fig. 11 Simulation results of Configuration-1: MLP, 10 layers, 20 neurons per layer: **a** PDE model (Input). **b** Vissim data (targets). **c** Initial error (PDE-Vissim). **d** Calibrated result. **e** Vissim data (targets). **f** Final error (calibrated result-Vissim). The NRMSE corresponds to 6.39%

NN-PDE shows a significant improvement when compared with the previous value of the NRMSE (28.02%) obtained without calibration (see Fig. 8).

Case-4: We use the following parameters of NN-PDE: value of the free flow speed ($\mathbf{u_f = 50\,km/h}$), neural network architecture (MLP), **10** hidden layers involving **20** perceptrons. The simulation results obtained are illustrated in Fig. 12a–f. The results in Fig. 12a-c correspond to the situation without calibration shown in Fig. 8. In contrast, the results in Fig. 12d–f correspond to the situation with calibration. These results obtained using NN-PDE are compared with the results obtained using VISSIM. The outcome of the comparison is depicted in Fig. 12f. Using this figure, we see that the NRMSE calculated corresponds to 3.5%. This value of the NRMSE obtained using NN-PDE shows a significant improvement when compared with the previous value of the NRMSE (31.51%) obtained without calibration (see Fig. 8).

5.3.3 Comment on Results

The four cases envisaged (see cases 1–4) show that the accuracy of NN-PDE significantly depends on the neural network architecture. This dependence is confirmed by the different values of the NRMSE obtained in each case. Further, it has been observed that the variation of the free flow speed also significantly affects the results obtained using NN-PDE. Our various numerical simulations have consisted in monitoring and varying both the number of perceptrons in the hidden layer and the free

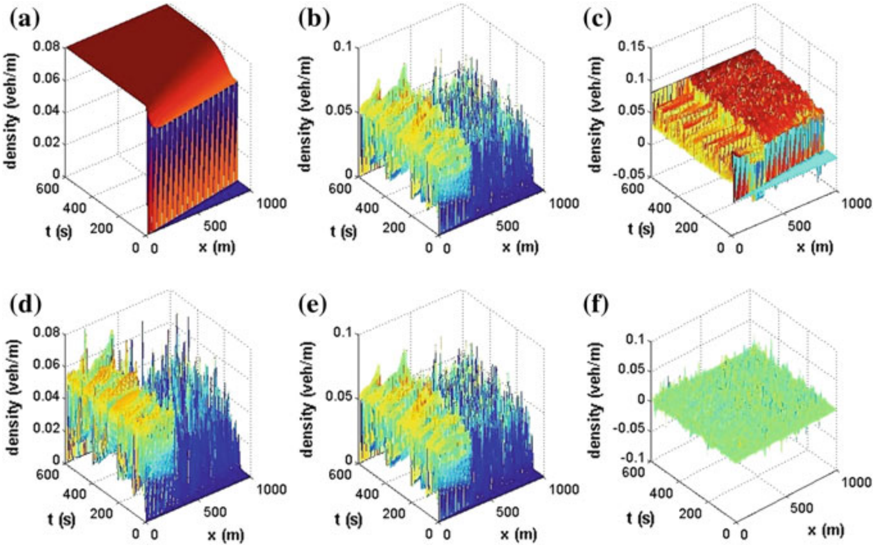


Fig. 12 Simulation results of Configuration-1: MLP, 10 layers, 20 neurons per layer: **a** PDE model (input). **b** Vissim data (targets). **c** Initial error (PDE-Vissim). **d** Calibrated result. **e** Vissim data (targets). **f** Final error (calibrated result-Vissim). The NRMSE corresponds to 3.5%

Table 1 Summary of the results of the eight configurations of NN-PDE

Network configuration	NRMSE in %	
	$u_f = 30\,\text{Km/h}$	**$u_f = 50\,\text{Km/h}$**
No calibration	28.02	31.51
1 layer, 50 perceptrons	6.55	6.39
10 hidden layers, 20 perceptrons per layer	2.21	3.5

flow speed. The results obtained by varying the two cited parameters are summarized in Table 1. This table provide the NRMSE for different combinations (each of which consists of a fix number of perceptrons in the hidden layer and a fixed value of the free flow speed). The results reported in Table 1 show the values of NRMSE for cases without calibration versus the NRMSE for cases with calibration. The outcome of this comparison clearly demonstrates that the new NN-PDE model developed provides results that are closer to those provided by VISSIM. This statement can be used to justify the realistic nature of the NN-PDE model developed in this chapter.

6 Concluding Remarks

In this chapter we have proposed a partial differential equation (PDE) based traffic flow model that is calibrated using a neural network (NN). The model developed is called NN-PDE (i.e., a combination of a PDE model with a calibration module based on NN). The calibration carried out is important, because the traditional mathematical models for traffic flow (proposed by the state of the art of traffic flow modeling) involve many parameters due to the complex dynamics undergone by traffic flow. Further, these models are based on assumptions that are generally nonrealistic, since they are unlikely to express the real traffic dynamics observed on arterial roads. Among the huge number of mathematical models for traffic flow proposed in the literature we have considered the LWR model, which is the seminal model proposed in the literature for modeling and simulating the dynamics of traffic flow at a macroscopic level of detail. The advantage of considering the LWR model is twofold. The LWR model is simple and can reproduce some relevant insights into the dynamics of traffic flow. These insights express specific phenomena (shockwaves, rarefaction waves, stop-and-go waves, etc.) that are generally observed on arterial roads. The calibration of LWR carried out in this chapter has led to a new model, called NN-PDE, which is very accurate and more realistic in describing the dynamics of traffic flow when compared to the basic LWR model. This statement has been proven in the framework of a benchmarking process conducted by considering the numerical simulation of various scenarios of traffic flow on arterial roads. Regarding the numerical simulation and benchmarking carried out in this chapter, the basic LWR model (corresponding to traffic flow in a single lane) has been simulated using the finite difference method (FDM) combined with the related stability conditions derived analytically. Both FDM scheme and stability conditions have been implemented in MATLAB to obtain numerical results. Further, in order to obtain the new NN-PDE model as a calibrated version of the basic LWR model, the MLP architecture has been used to design a neural network (NN) scheme that was trained offline using as target the data provided by VISSIM. The input data used for training the NN scheme are those provided by the LWR model. The outcome of the training has led to the NN-PDE model, which is further used for numerical simulations. Using NN-PDE, the various numerical simulations performed have revealed that the accuracy of NN-PDE can be improved by varying the configuration of the neural network module and also by monitoring the parameters (e.g., free flow speed) of the NN-PDE model. We have noticed that increasing both the number of perceptrons per layer and the number of layers involved in the NN architecture leads to a significant improvement in the accuracy of the NN-PDE model. However, we have detected a maximum number of layers and also a maximum number of perceptrons in layers above which the numerical results obtained do not lead to further significant improvement of the accuracy of NN-PDE. In contrast, a further increase of the number of perceptrons and layers may inherently lead to a waste of resources (memory), convergence issues (e.g., failure to converge), instability (loss of robustness), and low or worse computing performance. A benchmarking has been considered with the aim of comparing results of the numerical simulation of several specific traffic flow scenarios obtained using three different methods, namely the LWR model, the NN-PDE model, and VISSIM. The outcome of this comparison has revealed that the NN-PDE model developed provides results that are closer to results provided by VISSIM for all traffic flow scenarios envisaged in this chapter. For the same scenarios, the LWR model provides results that significantly diverge from those obtained using VISSIM. Hence the NN-PDE model developed shows very good accuracy and thus appears to be more appropriate (than the classical LWR) to model and simulate traffic flow scenarios on arterial roads.

References

1. Banister, D.: Sustainable urban development and transport -a Eurovision 2020. Transp. Rev.: Transnatl. Transdiscipl. J. **20**(1), 113–130 (2000)
2. Dimitrakopoulos, G., Demestichas, P.: Intelligent transportation systems. IEEE Veh. Technol. Mag. **5**(1), 77–84 (2010)
3. Deakin, E.: Frick, karen trapenberg and Skabardonis, Alexander, Intelligent Transport Systems: Linking Technology and transport Policy to help steer the future, University of california Transportation center (2009)
4. Ming-wei, L., Jun, Y.: Calculating the contribution rate of intelligent transportation system for the smooth general characteristic of urban traffic. In: 2014 International Conference on Management Science and Engineering (ICMSE), Helsinki (2014)
5. Papageorgiou, G., Maimaris, A.: Intelligent Transportation Systems. InTech, Rijeka (2012)
6. Lighthill, M.J., Whitham, G.B.: On kinematic waves: II. A theory od traffic flow on long crowed roads, in Proceeding Royal Society, London (1955)
7. Gupta, A.K., Katiyar, V.K.: Phase transition of traffic staes with on-ramp. Phys. A **371**(2), 674–682 (2006)
8. Hm, Z.: A theory of nonequilibrium traffic flow. Transp. Res. Part B **32**(7), 485–498 (1998)
9. Chera, C.M., Balino, J. L., Dauphin-Tanguy, G.: Two-equation traffic flow models framed within the bond graph theory. In: Proceedings of the 2010 Spring Simulation Multiconference in SpringSim '10, San Diego (2010)
10. Baruah, Ak: Traffic control problems using graph connectivity. Intern. J. Comput. Appl. **86**(11), 1–3 (2014)
11. Tolba, C., Lefebre, D., Thomas, P., El Moudni, A.: Continuous and timed petri ntes for the macroscopic and microscopic traffic flow modelling. Simul. Model. Pract. Theory **13**, 407–436 (2005)
12. Zhang, Y., Qiang, W., Yang, Z.: A new traffic signal control method based on hybrid colored petri net in isolated intersections. Intern. J. Intell. Transp. Syst. Res. (2016)
13. Bains, M.S., Ponnu, B., Arkatkar, S.S.: Modeling of traffic flow on Indian expressways using simulation technique in Procedia - social and behavioral sciences, Changsha (2012)
14. Trafficware: Synchro, Trafficware Group Inc. http://www.trafficware.com/synchro-studio.html (2015). Accessed 18 Oct 2016
15. PTV Visum, PTV Group. http://vision-traffic.ptvgroup.com/fr/produits/ptv-visum/ (2016). Accessed 18 Oct 2016
16. Siebel, F., Mauser, W., Moutari, S., Rascle, M.: Balanced vehicular traffic at a bottleneck. Math. Comput. Modell. **49**, 689–702 (2009)
17. Ai, W., Shi, Z., Liu, D.: Phase plane analysis method of nonlinear traffic phenomena. J. Control Sci. Eng. (2015)
18. Li, T.: Nonlinear dynamics of traffic jams. In: Second International Multi-Symposium on IMSCCS (2007)

19. Lo, S.-C., Cho, H.-J.: Chaos and control of discrete dynamic traffic model. J. Franklin Inst. **342**, 839–851 (2005)
20. Bellemans, T., De Schutter, B., De Moor, B.: Models for traffic control. J. A **43**(4), 13–22 (2002)
21. DTALite: a queue-based mesoscopic traffic simulator for fast model evaluation and calibration. Cogent Eng., *1*(1) (2014)
22. Burghout, W., Koutsopoulos, H.N.: A discrete-event mesoscopic traffic simulation model for hybrid traffic simulation. In: IEEE Intelligent Transportation Systems Conference, ITSC '06, Toronto, Ontario (2006)
23. Richards, P.: Shock waves on the highway. Oper. Res. **4**, 42–51 (1956)
24. Treiber, M., Hennecke, A., Helbing, D.: Derivation, properties, and simulation of a gas-kinetic-based, nonlocal traffic model. Phys. Rev. E, **59**(1) (1999)
25. Richards, P.I.: Shock waves on the highway. Oper. Res. **4**, 42–51 (1956)
26. HJ, P.: Models of freeway traffic control. In: Bekey, G.A. (ed.) Mathematical Models of Public System, pp. 51–61. Simulation Councils, inc., La Jolla (1971)
27. Ross, P.: Some properties of macroscopic traffic models. Transp. Res. Rec. **1194**, 129–134 (1988)
28. Del Castillo, J.M., Pintado, P., Benitez, F.G.: A formulation for the reaction time of traffic flow models. In: International Symposium on the Theory of Traffic Flow and Transportation, Berkeley (1993)
29. Jiang, R., Qing-Song, W., Zhu, Z.-J.: A new continuum model for traffic flow and numerical tests. Transp. Res. Part B: Methodol. **36**(5), 405–419 (2002)
30. Spiliopoulou, A., Papamichail, I., Papageorgiou, M., Tyrinopoulos, I., Chrysoulakis, J.: Macroscopic traffic flow model calibration using different optimization. In: 4th International Symposium of Transport Simulation (ISTS'14), Ajaccio (2014)
31. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965)
32. Golberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Professional, Reading (1989)
33. Holland, J.: Adaptation in Natural and Artificial Saystems. MIT Press, Cambridge (1992)
34. Rubienstein, R.Y., Kroese, P.: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning. Springer, New York (2004)
35. De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)
36. Mohan, R., Ramadurai, G.: State-of-the art of macroscopic traffic flow modelling. Int. J. Adv. Eng. Sci. Appl. Math. **5**(2–3), 58–176 (2013)
37. Shlayan, N., Kachroo, P.: Feedback ramp metering using godunov method based hybrid model. J. Dyn. Sys. Meas. Control **135**(15) (2013)
38. Greenshields, B.D., Bibbins, J.R., Channing, W.S., Miller, H.H.: A study of traffic capacity. In: Proceedings of the Fourth Annual Meeting of the Highway Reseach Board, Washinton (1934)
39. Greenberg, H.: An analysis of traffic flow. Oper. Res. **7**(1), 79–85 (1959)
40. Underwood, R.T.: Speed, Volume, and Density Relationships: Quality and Theory of Traffic Flow, Yale Bureau of Highway Traffic. Yale University Press, New Haven (1961)
41. Wang, D., Ma, X., Ma, D., Jin, S.: A novel speed-density relationship model based on the energy conservation concept. IEEE Trans. Intell. Transp. Syst. **99**, 1–11 (2016)
42. Kabir, M.H., Andallah, L.S.: Numerical solution of Multilane traffic flow model. Ganit J. Bangladesh Math. Soc. **33**, 25–32 (2013)
43. Helbing, D., Treiber, M.: Numerical simulation of macroscopic traffic equations. IEEE Comput. Sci. Eng. **1**(5), 89–98 (1999)
44. Pay, A., et al.: Calibration of traffic flow models using a memetic algorithm. Transp. Res. Part C: Emerg. Technol. **55**, 432–443 (2015)
45. Spiliopoulou, A., et al.: Macroscopic traffic flow model calibration using different optimization algorithms. In: 4th International Symposium of Transport Simulation (ISTS'14), Ajaccio, France (2015)

46. Pinter, J.D.: Calibrating Artificial Neural Networks by Global Op (2010)
47. Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Cornell Aeronautical Laboratory, Washington DC (1961)
48. Mandic, D., Chambers, J.: Recurrent Neural Networks for Prediction: Learning Algorithms. Architectures and Stability. Wiley, Chichester (2001)
49. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA **79**(8), 2554–2558 (1982)
50. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Natl. Publ. Group **323**, 533–536 (1986)
51. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Q. Appl. Math. **2**, 164–168 (1944)
52. Marquardt, D.: An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math. **11**(2), 431–441 (1963)
53. Chedjou, J.C., Kyamakya, K.: A universal concept based on cellular neural networks for ultrafast and flexible solving of differential equations. IEEE Trans. Neural Netw. Learn. Syst. **26**(4) (2015)
54. Ahsan, A., Andallah, L.S., Hossain, Z.: Numerical solution of a fluid dynamic traffic flow model associated with a constant rate inflow. Am. J. Comput. Appl. Math. **5**(1) (2015)

Travelers in the Second Modernity: Where Technological and Social Dynamic Complexity Meet Each Other

Oana Mitrea

Abstract Can we speak of a totally normal chaos (Lash, Individualization in a Non-Linear Mode, 2002, [21]) of mobility in the second age of modernity? The current essay draws its inspiration mainly from the findings of systems theory and sociology about nonlinearity and individualization with particular application to mobility systems. Its first part analyzes the societal system of the second modernity as an open system with high nonlinearity. The second part applies reflections from the construction of hypotheses about the usage dynamics of an intelligent concept of social interaction on the move.

1 Introduction

Perhaps one of the best ways of becoming aware of the importance of dynamic complexity is just to look out of the window while taking breakfasting in a big city awakening to a hectic life. In (post)modern daily life, anything can happen, in spite of the expectations and efforts to maintain a rather predictable daily agenda. The passing of people and vehicles, the movements and sounds of the natural environment, the invisible flows of information and communication, the various objects, and human actions and interactions are in constant flux and seem to be closely related to each other. Small modifications of behavior may produce important effects in unexpected distant places. The sociotechnical systems of a technological civilization are characterized by change over time; the ability of components to interact with each other, feedback, nonlinearity, self-organization, adaptability and counterintuitiveness [32].

The characteristics of dynamic complexity are particularly visible in today's mobility systems. The common observation of daily life in Western Europe depicts indeed an explosive circulation of persons and goods at rates never before encountered.

O. Mitrea (✉)
Institute for Smart Systems Technologies, Alpen-Adria-Universitaet Klagenfurt,
Universitaetstrasse 65-67, Klagenfurt, Austria
e-mail: oana.mitrea@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_8

However, mobility statistics reveal some more nuanced or unexpected facts. For instance, the *Statistics Brief 2013. Trends in the Transport Sector of OECD* emphasizes the difference in transport growth between emerging and advanced economies, with emerging economies continuing to outperform developed markets [26]. Important tendencies in transportation are stagnation and drop in air freight, the growth of rail freight in Russia and China, significant differences in road freight between EU and China, the rapid recovery of inland waterways, the increase of air transport in markets linked to emerging economies, the stability of EU rail passenger traffic at precrisis levels. Interestingly, OECD 2013 shows that car use appears stagnant or falling in high-income countries:

> Over the past 10 to 15 years, the growth of passenger vehicle travel volumes has decelerated in several high-income economies and in some growth has stopped or turned negative. Figure 11 shows an index of passenger-km volumes in a selection of high-income economies from 1990 through 2011. The slowdown in growth is clear in Germany. In France, car use is virtually unchanged since 2003. In Japan, car use has been declining since 1999. In the United Kingdom growth is negative since 2007 and it had slowed down considerably since 2003. The United States displays a decline since around 2005 or even earlier. (OECD 2014, p. 2)

These developments demonstrate that the transportation of persons is often particularly counterintuitive. Public opinion is far from being aware of a decline in automobility—quite the contrary. For comprehension of future developments in this field it is necessary to examine more closely the dynamic complexity of sociotechnical systems.

The mobility of persons is a challenging field for analysis, particularly due to its strong social flavor. Manifold emotions, power, and social interactions are involved in the decisions and actions of individual actors and institutions. These only-human particularities are revealed in the intercity commuting and intracity travel between home and office; mobility of children; leisure mobility during the week or on weekends; paying of social visits, trips to cultural and social events involving family members; transport trips for picking up or bringing children and other family members; supply movements (for fetching goods, shopping), other types of daily movements; as well as in the communication and information necessary to coordinate all such movements. On the other hand, technology has recently become more and more complex and intelligent (that is, active and interactive), aiming to resemble human characteristics. In technology designers' visions, technology should come closer to humans behavior and intentionality and become a true companion in everyone's mobility.

The starting challenge is to be able to identify how and where the social and technological complexities best meet and beneficially complement each other. A key phenomenon of current social complexity is individualization, with various consequences in social interaction, family, education, employment, transportation. The influence of individualization of social relationships and transportation is difficult to grasp without considering the system dynamics perspective of tipping points, feedback loops, and delays between causes and effects. On the other hand, mobility technologies do not influence social systems in a straightforward manner (pushing a button for more during the journey social contacts and obtaining them on the spot).

On the contrary, between technology design and actual use there are multiple dimensions and variables involved, among which nonlinear reinforcing or balancing relationships exist. The current chapter takes some first steps toward this explanatory goal: In the first part, the societal system of the second modernity is analyzed as an open system with high nonlinearity. The second part applies the reflections from the first part to the design of a particular concept for intelligent mobility: journey as purpose [25].

2 Manifestations of Dynamic Complexity

2.1 About Dynamic Complexity

The dynamic complexity of societies is revealed mainly in the processes of social change. Let us only think at what European societies looked like two centuries ago and what tremendous changes have occurred in the way that humans communicate, move, think, learn, work, entertain, live, and die. At the same time, changes in society feed on themselves, in the sense that they produce further positive and negative modifications of the relevant states of society on which changes had previously operated. For instance, the widespread presence of mobile information and communication technologies has modified the nature of work toward more flexibility of time organization and more mobility. The new type of work requires contextual real-time information necessary for speedy mobilization, a fact that challenges the way that information and communication are gathered, processed, and presented, and puts pressure on the development of real-time (dynamic) information systems.

The feedback view of the world in the vision of [29, 32] implies that our decisions alter the environment, leading to new decisions, but also triggering side effects, delayed reactions, changes in goals, and interventions by others. It is important also that these feedbacks lead to unanticipated results and ineffective policies [32], p. 9. From my perspective, one of the most interesting aspects from the list of complexity features synthesized by [32] is nonlinearity. Nonlinearity has received a variety of definitions, depending on the perspective of analysis adopted. For [32] this means that effect is rarely proportional to cause, and what happens locally in a system (near the current operating point) often does not apply in distant regions (other states of the system). Another interesting aspect is that complex systems are counterintuitive: In Forrester's words, people misjudge the behavior of social systems [12]. Many policies fail because policymakers do not recognize the feedback-loop dynamics underlying obvious facts such as human suffering in cities. The classical example given by [12] shows that an increase in housing as a measure to alleviate this suffering would only turn more and people into captives of depressed social systems and would therefore have many negative unexpected consequences. The comprehension of dynamic complex systems benefits from a perspective of modeling, where the role of feedback, nonlinearity, and self-organization is widely recognized. It is true that the comprehension and modeling of the whole social system in which the problem is integrated may represent an overwhelming task. By combining qualitative analysis with quantitative analysis and system synthesis reasoning to describe these undefined behavioral characteristics, system dynamics represents the appropriate method for dealing with nonlinear, high-order, complex time-varying systems [24]. The system dynamics simulation models have concentrated so far on macroscopic tasks such as urban passenger transport, energy consumption, and CO_2 emissions [24], the diffusion of and competition among multiple types of alternative vehicles, along with the evolution of the ICE fleet [33] and intercity transportation [22].

2.2 Travel and Mobility

According to Merriam Webster's Dictionary, the verb "to travel" means "to move or undergo transmission from one place to another." This quite general definition will be adopted in the current chapter. It complies with the majority of sociological positions that integrate the movement of humans with that of things, ideas, imagination (see new mobilities paradigm, [31, 35]), or propose various perspectives on modern transportation systems. Travel as an analysis category comprises bodily enacted movement (such as walking, cycling), technology-enabled transportation by various means (car, train, bus, plane), and the use of information and communication technologies (ICTs) for various purposes, whether or not travel-related. The thematic field of travel includes several possible categories of analysis such as the journey to work (travel to the office and back, short- and long-range commuting, business trips), coordination of daily life (such as shopping, procurement of goods, daily errands, dropping off and picking up children from school), regular leisure travel (such as relaxation walks after work, morning/evening walking, travel to sports locations), infrequent touristic travel (holiday travel, excursions), and socially driven travel (for instance to visit an aged relative, friends), etc.

The sociology of mobilities [35] has long recognized the importance of dynamic complexity. For instance, automobility is defined by Urry as a self-organizing autopoietic, nonlinear system that spreads worldwide, and includes cars, car drivers, roads, petroleum supplies, and many novel objects, technologies, and signs. The hallmarks of its dynamic complexity are the interconnection of many technical and nontechnical elements and the more recent reliance on specialized and arcane forms of expertise [35] (p. 14). The more recent concept of automobilities derives from "the new mobilities paradigm" [31, 35] and proposes a systemic integration of the movements of individuals, goods, ideas, representations, made possible mainly by advances in information and communication technologies.

The new smart technologies for information and communication and transportation enhance the mobilities of some, while reinforcing the immobilities of others. The growing availability of personal information over communication and traffic networks also bears new risks, such as accidents, diseases, terrorism, surveillance, and environmental damage [35] (pp. 11–12). Modern mobility includes "interspaces," places of intermittent movement where groups come together, use mobile phones, laptops, SMS, make arrangements on the move (idem). Such interspaces add value to the time spent traveling. That is, it is “not always a wasted dead time that people wish to minimize” (ibid.). Intracity travel, commuting, intercity and international travel cannot be imagined now without the all-pervading information and communication technologies (ICTs) for traveler information, driving, entertainment, and travel safety. A concept of smart mobility/smart travel has been coined for these forms of human movement in which information about events on the way, destinations, schedules, time, weather, entertainment, even one’s own psychical and physical state and possibilities for social interaction on the way should be provided reliably and in real time to travelers and the involved institutions.

The consideration of (mobile) ICTs within a dynamic complexity perspective about mobilities is important, particularly because these unify travel and social life, particularly in the making and remaking of connections on the move and planning of meetings. The possibilities for social relations and interactions, as well as for human mobility, are more and more dependent on the technological networks operating in the background [19]. In the perspective of mobilities, it is difficult to focus only on the mobility of persons and to ignore other types. Travelers of the second modernity experience a variety of mobile contextual configurations on the move, which integrate pieces of communication and information, portability of goods, imagination, and reveries on the way. The multiplicity of involved variables and the variety of feedback loops that characterize their relationships make the prediction of the behavior of today’s travelers particularly difficult. Small perturbations caused by events communicated via ICTs may have important unintended consequences on the further course of daily mobility: delays, side effects, cancellation of intentions, or new actions and decisions.

2.3 Individualization and Social Relationships

Contemporary sociology and social theory use various concepts to characterize our epoch: the second modernity [8], liquid modernity [5], reflexive modernization [7], technological civilization [3], risk society [6]. In the context of the research on individualization and the manifold changes, risks, and destabilizations that occur in the present highly nonlinear social systems [8, 21, 30], there has been a strong interest in the future evolution of social interactions. Expected is a passage from living for others to a life of one’s own, from the community of needs to elective affinities in a postfamilial family, new division of labor, declining birthrates, and desire for children [8].

While the concept of “first modernity” refers to the quite predictable and linear societal system of industrial society, characterized by rigid social structures and traditional relationships to institutions and between individuals, the “second modernity” represents an open system characterized by high nonlinearity [8]. Among sociologists there is still no consensus about the proper usage of terms: for instance, Giddens [13] argues that rather then entering a period of postmodernity, we are moving into one in which the consequences of modernity are becoming more radicalized and universalized than before [13].

The key driver of system destabilization in the second modernity is individualization [21]. Societal individualization means that central institutions of modern society, including basic civil, political, and social rights, but also paid employment and the training and mobility necessary for this, are geared to the individual and not to the group. Individualization is becoming the social structure of the second modern society [8] (p. xxii). Individualization is not equal to the individualism, but represents a structural sociological transformation of social institutions and the relationship of the individual to society [8] (p. 202). Its mechanisms (at the same time consequences) are the transformation of work, the decline of public authority, increasing personal isolation, a greater emphasis on individuality and self-reliance, the changing balance of power between men and women, a redefinition of relationships between men and women, the emergence of a culture of intimacy, informality, and self-expression [8].

Individualization represents a societal problem because it seems to destroy the established foundations of social existence: family and group relations, social structures on which individuals usually rely [8]. The disembedding from such traditional structures without reembedding confronts individuals with a precarious biographical freedom constructed like a do-it-yourself biography, constantly living with risk and uncertainty and being solely to blame for their eventual failures [8]. It seems that individuals of the second modernity do not have the time and reflective distance to themselves to construct linear and narrative life and work biographies (idem). They describe themselves as flexible and thus represent passage points for the unintended consequences that lead to system disequilibrium [21]. Flexibilization as a consequence of individualization is responsible for a vicious circle of deterioration of social (including work) relationships. The demand of being always flexible leads to uncertainty, a decrease of trust and commitment, and a superficiality of teamwork [30]. A short-term mentality replaces long-term ideas about love, family, and career, and this finally leads to a corrosion of the human character [30]. Individuals socialized in this way reproduce such nonvalues. Uncertainty, superficiality, poor trust between people, feed on themselves until a certain critical point whose consequences are difficult to foresee.

At the same time, individualization has also an important side effect for the activation of social relationships. Currently, the mobile Internet with its world of social apps sets new conditions for the display of one’s sociability and social relationships. Individuals have now the possibility of being virtually surrounded by more persons than their imagination can grasp. A cohort of virtual fellows can accompany each traveler in his or her trip and mingle with the reality of his or her physical travel. Supported by mobile ICTs, the second modernity’s individuals act as amalgamators who put together networks, construct alliances, and make deals in an atmosphere of risk in which knowledge and life changes are precarious [21]. The increasing number of networks in which they are involved brings new possibilities for action and interaction.

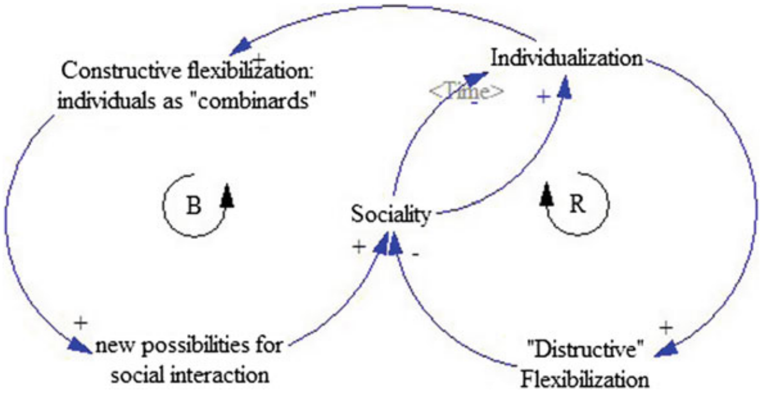

Constructive flexibilization: individuals as "combinards"
Individualization
<Time>
+
-
+
B
R
Sociality
+
-
+
new possibilities for social interaction
+
"Distructive" Flexibilization
+

Fig. 1 Mental model of the duality of individualization

In Fig. 1 these relationships among variables are described by two main causal loops:

1. Individualization induces flexibilization with negative side effects on the quality of social relationships (mutual trust, loyalty, commitment). A poor quality of social relationships reinforces the part of individualization that has generated "destructive" flexibilization.

2. On the other hand, individualization can also induce a "good" flexibilization, understood as the ability to create social connections and to creatively combine sequences of life trajectories. This can bring new possibilities for social interaction and can potentially increase the quality of social relationships (if contacts lead to accomplishments), balancing some of the negative aspects of individualization.

From the considerations above it can be concluded that individualization represents a complex issue. This can negatively influence the quality of social relationships in time but can also open spaces for creativity and positive societal change.

In the following it will be shown how these positive and negative factors intermingle with mobility. The sociologists of mobility have emphasized how mobility nowadays becomes more and more individual and therefore personalized through its enrichment information and communication. However, little has been researched about the way that the various types of human movement (physical, virtual, imaginary) and larger social phenomena influence each other and merge into a complex societal problem: individualization of mobility.

Some open research questions are as follows: Is societal individualization beneficial for mobility in the long term? Is the long resistance to change of the system of automobility [34] a side effect of the overall societal individualization? Is the personalization of mobility enabled by numerous "apps" a problem or a way (a solution) to achieve better (more socially responsive and environmentally friendly) mobility? The answers to these questions pose numerous difficulties that also derive from the nonlinear relationships between variables.

2.4 Social Relationships and Mobilities

In general, there is consensus among researchers about the existence of a certain relationship between the forms and amount of mobility and the nature and intensity of social relationships [18, 19, 35]. There is also agreement about the role that (mobile) ICTs play in travel and social life. The possibility for social relations and interactions, as well as for human mobility, is indeed more and more dependent on the technological networks operating behind the scenes [19] (pp. 92–93). However, the reciprocal relationship among the three areas has not yet been clarified. The exact nature of this relationship (referring to the exact specification of the model, linearity/nonlinearity involving circularity and feedback, time dynamics) represents quite an unexplored territory. There is significant research about how mobile ICTs and sociality (sociability and association) interact.

The mobile phone could be regarded as a new stage on which mobile society is acted out [28]. The rather weak and superficial mobile interactions turn mobile telephony contexts into places: where one can go and chat about anything. With a mobile phone, one can feel that one is in a place where emotional arguments and friendly laughter, for example, are appropriate. It is a stage that resembles a virtual cafeteria or marketplace where people meet each other [28] (p. 27). Mobile phones also play a significant role in the managing of distant interactions, with a focus on trust and negotiated local context together with construction of bonds and commitments [23] (pp. 94–109). Some of the most studied subjects in the area of relationship travel sociality has been the influence of modern transportation enriched with mobile communication and information on the traditional community [4, 30, 34], the impact of business mobility on family coordination [19], the interactions between the spatial structure of social networks and the travel pattern of the members of these networks [2], friendship and mobility in location-based social networks [10].

Negative as well as positive aspects have been highlighted. Bauman's analysis of the changes in local communities highlights the way in which the increasingly technicized transportation of interpersonal communications weakens the substance of the traditional community. The social relationships in traditional communities were based on proximity, vicinity, high density of face-to-face communication, and intensive daily interaction [4] (p. 37). The borders of communities were defined by the volume and speed of possible transportation. Such proximity-based intensive direct contacts become difficult in the age of globalization and heavy travel. With the use of information and communication technologies, even space came to matter less (idem). Flows instead of places and the idea of the death of distance seem to be justified by the possibility of interaction at the speed of light through IT networks. The physical (noncyber) space is degraded to a site for delivery, absorption, and recycling of the essentially extraterritorial cyberspace information [4] (p. 38). "Currently, some persons have no other choice than to stay in place due to the lack of mobility resources." Therefore, immobility becomes a measure of social deprivation and the lack of freedom (idem) [4].

Modern work travel is criticized mainly in what concerns its recent tendency to turn into a domination practice [19]. The business mobility regime [19] pursues the disciplining of individuals' movements, in particular the normalization of work travel, the rationalization of mobility management, and the self-optimization of one's mobility behavior [19] (pp. 87–89). Mobile workers, as entrepreneurs of their own mobility skills, should accept the necessity to travel and renounce their free time and self-determination [19] (p. 89). The fragmentation and flexibilization of time, space, and life experiences in business mobility regimes also contaminates the private life domain. The increasing distances between communicators challenge the coordination with the members of the trusted social space such as family and close friends. Multiple processes of negotiation and renegotiation over (mobile) ICTs are necessary [11] in [19] (p. 92). A reembedding in the family context still represents the priority of business travelers: as interviews with frequent business travelers and travel managers in Sweden showed, although individual travelers often appreciate having good working conditions while traveling, they want to minimize time spent away from home and family rather than to make productive use of their travel time [16].

The flexibilization of the new capitalism [30] is criticized for negative consequences such as the uncertainties of flexibility, the absence of deeply rooted trust and commitment, the superficiality of teamwork, most of all the specter of failing to make something of oneself in the world, to get a life through one's work [30] (p. 138), the short-term mentality that replaces the long-term one: family, love and family, career, and ultimately the corrosion of human character. The short-term perspective represents a principle that corrodes trust, loyalty, and mutual commitment. Such communities are not empty of sociability or neighborliness, but no one in them becomes a long-term witness to another person's life [4] (p. 38). The fleeting forms of association become more powerful than the long-term connections (idem). Individuals adapt to the present circumstances by trying to invest sense and purpose and to identify positive aspects in them [4] (p. 23). This leads to the high valuation of mobility in all senses. Commenting on the results of a study of singles highly valuing mobility in their professions, Beck and Beck-Gernsheim conclude that the idea to having to practice a lifelong profession is considered a burden rather then something desirable, while change, in work as well as in relationships, is considered natural and desirable by many [8] (p. 162). Because the establishment of a short-term mentality is considered a powerful mechanism for the dissolution of community relationships [4, 30], it is also important to analyze how this is involved with sociability and mobilities.

On the positive side, real-time and mobile communication technologies seem to reinforce the initiation of new on-trip connections, sometimes by means of mobile social networking. Even if these interactions cannot compete with face-to-face contacts in what concerns long-term investment in trust [4] (p. 38), they do extend the circle of social possibilities. The access to travel information and communication with family, friends, and colleagues while on the road also opens possibilities for the concomitant planning of mobility and sociality, one as a function of the other: when to meet somebody and whom to meet, the reformatting of social actions and relationships under mobile conditions. A mobile contextual configuration emerges

[1], characterized by intensive information exchange in mobile phone calls about location, transportation, current activity, etc. In a recent article, Urry emphasizes the fact that mobile ICTs make and remake connections across time and over space, which are essential for the extension and/or consolidation of social networks [36]. It is worth mentioning here the ad hoc nature of the spontaneous network-type mobility that follows the connective logic of integration in meaningful contexts and socially coupled connections and interactions: one no longer meets not in a specific place but in the city or city quarter, the exact coordinates of which are communicated by means of the mobile phone [19] (p. 93).

3 Toward Sociability-Friendly Mobility Systems

Sociability represents a psychosociological characteristic of individuals that have a feeling for, and a satisfaction in, the very fact that one is associated with others and that the solitariness of the individual is resolved into togetherness, a union with others [17] (p. 255). In this view, sociability represents a play-form of association, completely oriented around personalities that should bring joy and imagination (idem). A slightly different perspective is that of a fund of sociability [37]. According to this idea, individuals require a certain amount of interaction with others and would experience stress only if the total amount of relating to others was too little or too great [37]. One must also distinguish between the intrinsicality of sociability (stating that sociability is dependent on internal factors such as social class and age, instead of the individual situation) and the interactivity deriving from a socialization behavior that assesses that individuals aim to fulfill their social interaction needs and to build social ties until these needs are met or by satisfying these needs through meeting already known acquaintances [9] (p. 2). To understand sociability as interactivity, it is important to focus on the characteristics of the context (mobile/not mobile, ICT enriched or not) in which individuals manifest various needs of interacting with others. The context-dependent variation of these needs is the element that distinguishes interactivity from individual-specific intrinsic sociability. This perspective on sociability as interactivity will be adopted in the present research.

Another important differentiation is between the sociability oriented toward direct (face-to-face) interaction and the one expressed in virtual social networks. A person who is constantly on social networks in trains and pays no attention to the other travelers is sociable in a different way than an individual who starts chatting with the fellow travelers as soon as he or she takes a seat. An examination of the combination of these two forms of sociability in traveling can represent a captivating issue for research. The analysis of the interactive sociability (or interactivity) should take into account the social interaction needs of an individual as a function of the context, the sociability values (such as joy, relief, vivacity), the manifestation of forms of sociability in various mobile contexts (direct/online, forms of presentation of self), other variables such as age, gender, family status, number and age of children, number of friends, perceived intensity of contacts with family/friends, the number of groups in which a person is integrated, the dimensions of groups, the number of new contacts per month, the number of exchanged (responded) messages per week, the level of trust between participants, participant satisfaction with social interactions, number of persons who still interact with the person after a longer time (adapted from [27]).

The disposition toward being sociable while traveling may represent an important condition for actually engaging in social interactions in the mobile space. Existing studies are not very encouraging about this aspect, particularly those analyzing public transportation in Western societies. Their starting observation is that people who are traveling by train or bus are actively trying to avoid eye contact and initiation of conversation with strangers. Studies from symbolic interactionism [14] tried to understand why some people actively disengage from others over the course of the ride [20] and identified the need of travelers to design and carefully coordinate interaction rituals by avoiding people nearby and slipping into a personal space of the self [20] (p. 16). This avoidance behavior varies in terms of the cultural context and reference time periods. As emphasized in [25], active social disengagement did not apply to the lengthy shop queues in the period of Romanian Communism, where citizens actively sought interaction with fellow sufferers during the long waiting times for basic necessities. Also, train journeys usually provided a venue for very lively interactions and problem-solving sessions where even lifetime relationships could be initiated. The key difference was that a feeling of spontaneous community accompanied the common ride. The perceived community cohesion could represent a powerful mechanism toward interaction and cooperation in the mobile public space.

A mobile contextual configuration [1] refers to the reciprocal shaping of travel, communication and social interaction by means of mobile ICT devices for communication, travel information, and location-based information and services. Meeting places and events are communicated on mobile phones (and soon by smart cars interconnected with homes, offices, and leisure targets). People who have previously been unknown to each other are unified in common missions, for instance in ride-sharing. Their trajectories can change rapidly as a function of events that have occurred and other persons encountered on the way. Small perturbations of routes as a result of interpersonal communication, new tasks, new interests, make the movements of today's travelers extremely difficult to understand and predict. Mobility should be thus conceived in a nonlinear manner (feedback loop) as a continuous adaptation of humans, society, and technical systems to changes in purposes and tasks that appear along the way. The simultaneous immersion and switching between multiple offline and online worlds, quite often observed in trains and buses, can also represent a chance for a richer interaction and communication that combines the benefits of pretrip information with direct communication [25]. Due to the new possibilities opened by both smart mobile communication and ITS, new approaches in the field of intelligent mobility systems are called for in which not only time savings and the minimization of travel variability are important, but also the intrinsic valuation of travel time in terms of sociability and action possibilities along the way.

A conceptual implementation of this idea in the field of intelligent public transportation has been advanced in [25]. The concept of journey as purpose aims at turning public transportation into a space of social interaction and multipurpose action (transient mobile social space). This vision is consonant also with the considerations of Urry (2007) about the productive value of travel and the potential of meetings on the move [35] (pp. 250–251).

A journey as purpose story:

> Imagine a busy young mother who works as a freelancer in advertising. She usually travels a lot to meet her contractors or colleagues. Today it takes her around 45 min to reach an office in the neighboring city. She has just received bad news from her contractor there by mail, why she now feels totally confused and upset and would gladly cancel the contract, only that she doesn't know very much about legal conditions and obligations. She is then sitting in the commuting train for almost one hour. What if there would be some opportunity inside the train to get at least some general orientation and advice about her problem? What if somebody with this knowledge would travel on the same route by the same means? In such a situation she would gladly take even a later train to meet this person there in order to be able to explain him/her the problem and to receive feedback on the spot.

How are decisions and actions of this user in this scenario linked and how do they flow in time?

Figure 2 depicts several main hypotheses about the possible dynamics of the use of the Journey as Purpose concept (abbreviation JaP). The nodes refer to variables as well as actions and events. The structure of the figure contains nodes, their relationships, and several feedback loops that account for the behavior of the system (Fig. 2).

Explanation of the dynamic relationships in Fig. 2:

- At the beginning of actions and decisions lies a problem generated by a given situation. This generates a need to search for a solution in a reasonable amount

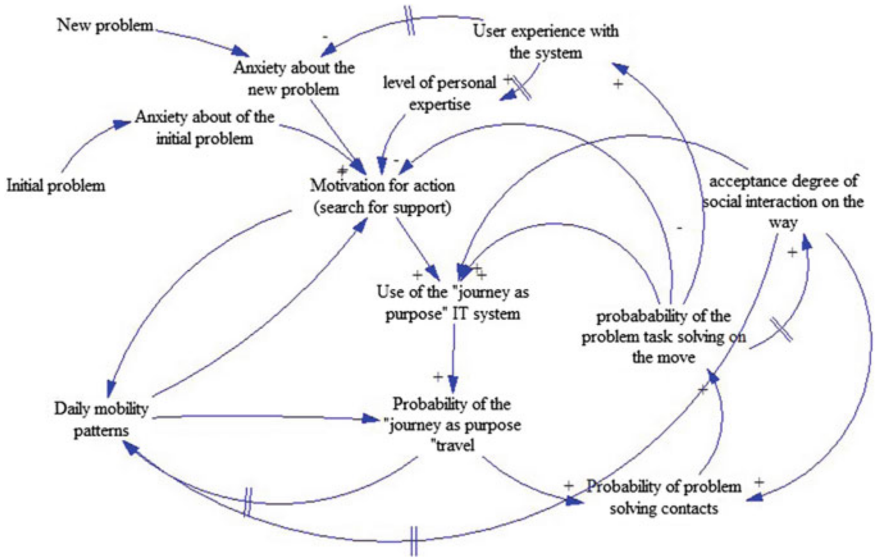

New problem
User experience with the system
Anxiety about the new problem
level of personal expertise
Anxiety about of the initial problem
Initial problem
Motivation for action (search for support)
acceptance degree of social interaction on the way
Use of the "journey as purpose" IT system
probabability of the problem task solving on the move
Daily mobility patterns
Probability of the "journey as purpose "travel
Probability of problem solving contacts

Fig. 2 Hypotheses about the usage dynamics of the journey as purpose (JaP)

of time, without great effort. Mobility is already there, since this person usually commutes. The idea of the concept is to combine mobility with problem-solving via an IT system that enables various interactions and actions during travel.

- Several feedback loops and delays can be observed in this simplified model. The initial problem generates the motivation to look for support, which can lead to the use of the IT JaP solution and subsequent travel by a given mode in order to come into contact with the problem-solving person. A successful solution of the problem by the IT system JaP reduces the motivation for further action for this particular problem (balancing).
- A new problem appears. At this time, some users have already integrated the knowledge about journey as purpose into their way of approaching problems. As users get advice from expert contacts, their level of personal expertise may increase. This fact can lower their motivation to discover solutions trough JaP (balancing), with negative impact on the probability of the travel triggered by JaP and the number of social contacts on the move (similar to saturation in adoption).
- On the other hand, successful problem-solving can enhance the acceptance of social contacts on the move in general, a fact that can be beneficial for use of the system for various cooperative actions and initiatives by the supported persons to assist others.

4 Conclusion and Future Work

To sum up, the present chapter has aimed at illustrating how small spontaneous choices of individualized travelers about their movement, communication, and information have the power to create (sometimes unintended) consequences that can set new conditions for future social interactions on the move.

The temporal evolution of both mobility and social interaction behaviors of a particular traveler is also influenced by some variables described below that are not included in this model, but will represent key elements of future work: First of all, background (static) information about the present life situation of the users can be collected: gender, age, personality type, family situation, employment, lifestyles and values, level of sociability, attitudes toward employment, mobility, satisfaction with life, etc. They can be determined in a one-way surveying step and summarized in the form of personal scores. In a second step it is necessary to concentrate on obtaining dynamic information about the following:

- Internal state variables: stress, enjoyment, health state, emotions, tiredness, etc.
- Context variables of a social or technical nature: a concrete demand for something, various tasks/goals, mobile information, communications on the move (face to face or over ICTs).
- Mobility variables: travel mode, distance, time to destination, delays, stops, mode changes and pauses, traffic variables, degree of comfort. The importance of the mode of travel for social relations is highlighted by interesting research in the field

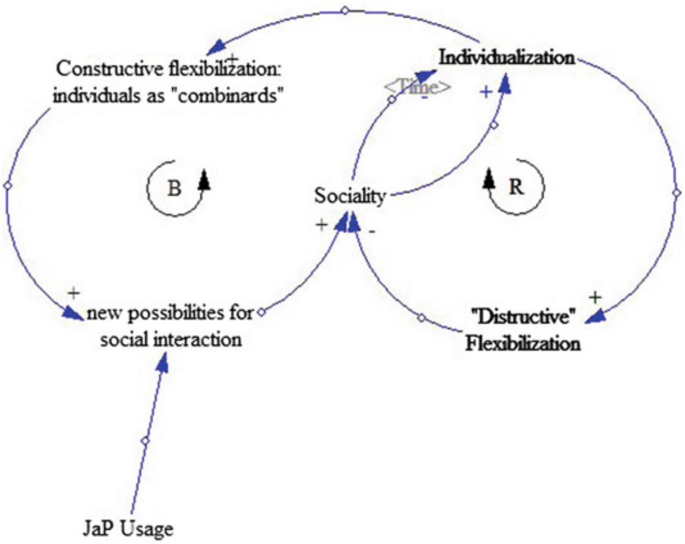


Fig. 3 The intervention of JaP in the relationships between individualization and social interaction

of travel behavior and society. It is shown that a mixed environment that allows much walking increases the social capital in the community and the probability of making stable connections in time [15].

- Variables characterizing candidates for contacts: age, gender, visual attributes of class, occupation, income.

In Fig. 3, the quality of social relationships can be positively influenced by the opportunities for interaction and harmed by the destructive forces of individualization (flexibilization, risky biographies, and uncertainty, leading to the deterioration of social relationships). Two main feedback loops are visible there: (a) the right-hand circle describing the reinforcing of individualization via flexibilization and the deterioration of social relationships and (b) the left-hand balancing circle via better opportunities for social interaction along the way. In the left-hand circle, the use of a system that enriches the mobile space with new possibilities for action and social interaction has the potential of increasing the power of the constructive circle. The positive side of individualization might gain in importance and become more widespread.

References

1. Arminen, I., Weilenmann, A.: Mobile presence and intimacy: reshaping social actions in mobile contextual configuration. J. Pragmat. **41**(10), 1905–1923 (2009)
2. Axhausen, K.W.: Social networks and travel: some hypotheses, Arbeitsbericht Verkehrs- und Raumplanung 197. IVT and ETH (2003)

3. Bamme, A.: Technologische zivilisation. In: Grossmann, R. (ed.) Iff Texte, Band 3. IFF (2002)
4. Bauman, Z.: The Individualized Society. Wiley, New York (2001)
5. Bauman, Z.: Liquid Modernity. Wiley, New York (2013)
6. Beck, U.: The terrorist threat: world risk society revisited. Theory Cult. Soc. **19**(4), 39–55 (2002)
7. Beck, U., Bonss, W., Lau, C.: The theory of reflexive modernization: problematic, hypotheses and research programme. Theory, Culture and Society **20**(2), 1–33 (2003)
8. Beck, U., Beck-Gernsheim, E.: Individualization: Institutionalized Individualism and its Social and Political Consequences. Published in Association with Theory, Culture & Society. SAGE Publications, California (2002)
9. Borrel, V., Legendre, F., de Amorim, M.D., Fdida, S.: Simps: using sociology for personal mobility (2006). CoRR, abs/cs/0612045
10. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’11, New York, NY, USA, pp. 1082–1090. ACM (2011)
11. Elliott, A., Urry, J.: Mobile Lives. International Library of Sociology. Taylor & Francis, London (2010)
12. Forrester, J.W.: Counterintuitive Behavior of Social Systems. Massachusetts Institute of Technology, Sloan School of Management (1970)
13. Giddens, A.: The Consequences of Modernity. Raymond Fred West Memorial Lectures. Stanford University Press, Stanford (1990)
14. Goffman, E.: The Presentation of Self in Everyday Life 1959. Garden City, New York (2002)
15. Grannis, R.: From the Ground Up: Translating Geography into Community through Neighbor Networks. Princeton University Press, Princeton (2009)
16. Gustafson, P.: Travel time and working time: What business travellers do when they travel, and why. Time Soc. **21**, 203–222 (2012)
17. Hughes, E.C., Simmel, G.: The sociology of sociability. Am. J. Soc. **55**(3), 254–261 (1949)
18. Kakihara, M., Sorensen, C.: Mobility: an extended perspective. In: HICSS Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002, pp. 1756–1766 (2002)
19. Kesselring, S.: Betriebliche mobilitátsregime. zur sozio-geografischen strukturierung mobiler arbeit corporate mobility regimes. socio-geographical structurations of mobile work. Zeitschrift fúr Soziologie **41**(2), 83–100 (2012)
20. Kim, E.C.: Nonsocial transient behavior: social disengagement on the greyhound bus. Symb. Interact. **35**(3), 267–283 (2012)
21. Lash, S.: Individualization in a Non-Linear Mode. In: Beck,U., Beck-Gernsheim, E. (eds.): Individualization: Institutionalized Individualism and its Social and Political Consequences. Published in Association with Theory, Culture & Society. pp. vii-xiii. SAGE Publications. California (2002)
22. Lewe, J.-H., Hivin, L.F., Mavris, D.N.: A multi-paradigm approach to system dynamics modeling of intercity transportation. Transp. Res. Part E Logist. Transp. Rev. **71**(C), 188–202 (2014)
23. Licoppe, C., Heurtin, J.-P.: France: preserving the image. In: Katz, J.E., Aakhus, M. (eds.) Perpetual Contact, pp. 94–109. Cambridge University Press, Cambridge (2002). Cambridge Books Online
24. Liu, X., Ma, S., Tian, J., Jia, N., Li, G.: A system dynamics approach to scenario analysis for urban passenger transport energy consumption and CO_2 emissions: a case study of Beijing. Energy Policy **85**(C), 253–270 (2015)
25. Mitrea, O.S., Kyamakya, K.: The journey is the purpose: a concept for public transportation as social transient space. In: Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013), October, pp. 493–498 (2013)
26. OECD. Statistics Brief 2013. Trends in the Transport Sector. OECD (2013)
27. Preece, J.: Sociability and usability in online communities: determining and measuring success. Behav. Inf. Technol. **20**, 347–356 (2001)

28. Puro, J.-P.: Finland: a mobile culture. In: Katz, J.E., Aakhus, M. (eds.) Perpetual Contact, pp. 19–29. Cambridge University Press, Cambridge (2002). Cambridge Books Online
29. Richardson, G.P.: System Dynamics. Kluwer Academic Publishers, Dordrecht (1999/2011)
30. Sennett, R.: The Corrosion of Character: The Personal Consequences of Work in the New Capitalism. Social Science. W.W Norton, New York (1998)
31. Sheller, M., Urry, J.: The new mobilities paradigm. Environ. Plan. A **38**(2), 207–226 (2006)
32. Sterman, J.: System dynamics: System thinking and modeling for a complex world, Working Paper Series ESD-WP-2003-01.13-ESD Internal Symposium, MIT (2002)
33. Struben, J., Sterman, J.D.: Transition challenges for alternative fuel vehicle and transportation systems. Environ. Plan. B Plan. Design **35**(6), 1070–1097 (2008)
34. Urry, J.: The system of automobility. Theory Cult. Soc. **21**(4–5), 25–39 (2004)
35. Urry, J.: Mobilities. Wiley, New York (2007)
36. Urry, J.: Social networks, mobile lives and social inequalities. J. Transp. Geogr. **21**, 24–30 (2012). Social Impacts and Equity Issues in Transport
37. Weiss, R.S.: The fund of sociability. Trans-action **6**(9), 36–43

COMPRAM Assessment and System Dynamics Modeling and Simulation of Car-Following Model for Degraded Roads

A.K. Kayisu, M.K. Joseph and K. Kyamakya

Abstract The presence of potholes in roads is a complex societal reality in developing countries that leads to situations such as congestion, chaotic driving, and acceleration of road degradation. More and more money is spent on the maintenance of the same segment of road and on the repair of cars due to potholes. This complex phenomenon "traffic congestion in Kinshasa linked to degraded roads" is analyzed with the COMPRAM methodology. It is shown that more policy intervention is needed via improvement of legislation, road maintenance, and road monitoring. In this paper we elaborate on traffic flow models, system dynamics (SD), and COMPRAM. We briefly discuss the relationship between the "car-following" model and the "microscopic/macroscopic" traffic model. For measuring the pothole effect on road users such as cars, a simulation of a car-following model was done with system dynamics (SD). We considered two scenarios for simulation: a scenario with a single pothole on a one-lane road and a scenario with two potholes separated by a distance of 590 meters on a one-lane road. The results of the simulations demonstrate that in the presence of the pothole at the microscopic level, speed and travel time are negatively affected, impacting road capacity at the macroscopic level.

Keywords Road pothole · Simulation of a car-following model · COMPRAM methodology · System dynamics · Traffic congestion in Kinshasa

A.K. Kayisu (✉) · M.K. Joseph
University of Johannesburg, Johannesburg, South Africa
e-mail: antoinekayisu045@yahoo.fr

M.K. Joseph
e-mail: meeraj@uj.ac.za

K. Kyamakya
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_9

1 Introduction

Car-following models are important in seeing/showing the microscopic behavior of the driver–vehicle system during follow-the-leader scenarios, as well as in modeling microscopic traffic propagation and evolution [1]. At the macroscopic level, it is necessary to see the effect of microsimulation on road traffic in general [1, 2].

At the microscopic level, embedded devices in vehicles assist drivers to maintain safe and comfortable driving conditions for collision avoidance or for intelligent speed adaptation [2, 3].

In the developed world, road users, traffic engineers, and traffic managers use advanced traveller information systems (ATIS) and advanced traffic management systems (ATMS) to get information such as travel time estimation, prediction, and congestion state. Advanced sensor systems placed either at roadsides or embedded in cars can thereby collect the necessary basic information or raw data [1, 4, 5].

The authors of [6] propose a rethinking of intelligent transportation system (ITS) design and traffic management, particularly in India and generally in developing countries because of considerations such as the absence of traffic lights and road markings as well as chaotic (i.e., undisciplined) driver behavior. These situations in developing regions led us to consider traffic management and specifically a car-following model on degraded roads as a complex system that is a reality in developing countries and areas (cars, drivers' reactions to potholes, people around a degraded road, weather, road traffic management, and road maintenance policies, etc.).

The complex problem-handling methodology (COMPRAM) is used to analyze this phenomenon globally. COMPRAM allows us to figure out a way to handle complex societal problems while involving a system dynamics (SD) simulation option [7–9].

In this work, we first briefly present and discuss the traffic flow models, system dynamics (SD), and COMPRAM methodology. These models, tools, and methodology are applied to the development and simulation of a car-following model on a degraded road specifically in the presence of road potholes for a traffic congestion management related initial investigation.

The main objective of this work is to analyze road potholes as a complex societal problem and to propose a sustainable solution. Secondly, we focus on systems dynamics [10] based modeling to assess the effect of potholes on a car-following model in order to depict the related impact on microscopic variables, and finally, we discuss the negative macroscopic incidence of congestion through a significant road capacity reduction.

2 Traffic Flow Modeling and a Brief Introduction to System Dynamics (SD)

2.1 Macroscopic Traffic Flow Model

Macroscopic models [11, 12] involve the motion of groups of vehicles moving in a system. Traffic flow appears to conform to the theory of fluid dynamics, and first- as well as second-order partial differential equation (PDE) models are mainly used in macroscopic models. One of them, the Lighthill–Whitham–Richards (LWR) model, is based on three variables: speed, flow, and density [11, 13]. The system is completely determined by a system of three equations:

$$\begin{cases} q(x,t) = k(x,t) \times u(x,t), \\ \frac{\partial q(x,t)}{dx} + \frac{\partial q(x,t)}{dt} = 0, \\ u(x,t) = u_e(k(x,t)), \end{cases} \tag{1}$$

where q represents the traffic flow (vehicles/hour), k is the traffic density in a road segment (vehicles/km), u and u_e stand respectively for the speed (km/hour or m/sec) and equilibrium speed in steady-state condition, x (km or m) and t (sec, min, or $hour$) are respectively spatial and temporal dimensions.

The first equation is the fundamental diagram, the second expresses traffic conservation (the number of vehicles on a road segment is constant, i.e., there are no additional vehicles coming from a ramp or going to a ramp), and the third equation is the definition of speed. Here, the assumption is that the traffic is always in an equilibrium state and that it evolves from one equilibrium state to another.

For analytical resolution purposes, the Eq. (1) can be reduced to a single hyperbolic PDE:

$$\frac{\partial k(x,t)}{\partial t} + q_e'(k(x,t))\frac{\partial k(x,t)}{\partial x} = 0, \tag{2}$$

$$q_e(x,t) = k(x,t) \times u_e(x,t). \tag{3}$$

Equation (3) is the fundamental diagram in density, and $q_e'(k(x,t))$ in Eq. (2) is a slope that is equal to the tangent to the fundamental diagram in flow at the point k:

$$\begin{cases} q(x,t) = k(x,t) \times u(x,t), \\ \frac{\partial q(x,t)}{\partial x} + \frac{\partial q(x,t)}{\partial t} = 0, \\ u(x,t) = u_f[1 - (\frac{k(x,t)}{k_j})^l]^p, \end{cases} \tag{4}$$

where k_j is the maximum car concentration or the jam density occurring when the road section is full of vehicles, and in that extreme situation, flow rate and velocity are zero, and u_f is the desired speed (it is the maximum speed limit indicated by traffic signs at the roadside), which occurs when the number of vehicles is low enough on the road section and there is no interaction between cars, so every car can move freely, that is, traffic can flow with free speed; l and p are constant values, which are generally determined empirically.

2.2 Microscopic Traffic Models and Car-Following Model

Microscopic traffic models [12, 14] attempt to describe the motion of individual vehicles within a system. The following components are contained in a microscopic traffic model: a car-following model, a lane-change model, a route choice model, etc. [15]. The microscopic traffic flow model analyzes very small changes in the traffic stream over time and space.

A car-following model can enable us to understand traffic flow, common driving behavior and forms the bridge between the individual car behavior and the macroscopic model.

Car-following theories describe a model's behavior dictated by a lead vehicle in an uninterrupted flow. Each vehicle's behavior is described by its own ordinary differential equation. A driver, who reacts based on the stimulus–response mechanism, follows another vehicle by judging distance, speed difference, reaction time, and road conditions (especially the existence of potholes), as described in this study. The car-following process models microscopic behavior of the driver–vehicle system during follow-the-leader scenarios mainly to maintain safe and comfortable driving conditions (i.e., distance between a given car and the car ahead of it) to avoid collisions.

The global traffic flow is determined by tracking multiple individual vehicles. Three functions—position, velocity, acceleration—and driving states (one of three states: free traffic state, following state, braking state) are involved.

Various car-following models exist: microscopic car-following, speed–distance, and stimulus–response models, etc. [14, 16]. The first model is a simple linear relationship between driving speed and distance headway with the leader and follower vehicles or between speed and time headway [14]. Another series of car-following models is a stimulus–response system introduced by researchers at General Motors (GM), as shown in Eq. (5).

In this car-following system, the stimulus is represented by the relative speed between a leader ($(n-1)$th vehicle) and a follower (nth vehicle), and the response is represented by the acceleration of the follower, as illustrated in Eq. (11).

The acceleration or deceleration from a driver may be regarded as the response to the stimulus received from the interaction with other particles in the system, here specifically in the presence of either a vehicle ahead or an obstacle such as a pothole. Also, the stimulus variable is a function of speed and relative speed. The proportionality factor represents the follower driver's sensitivity to the stimulus:

$$Response = \lambda \times Stimulus. \tag{5}$$

Equation (5) is based on the principle of action (stimulus from the leader vehicle or the discovery of an obstacle such as a pothole) and reaction (response from the follower vehicle to the vehicle ahead or/and to the obstacle in front such as a pothole). The response is proportional to the intensity of the stimulus, but it also depends on the driver's sensitivity to this stimulus.

In this process of action and reaction, the follower has to avoid collisions, which means he must maintain a safe distance to the vehicle ahead. The gap Δ required for the nth vehicle is given by

$$\Delta x_n(t) = \Delta x_{safe} + \tau \dot{x}_n(t) \tag{6}$$

$$x_{n-1}(t) - x_n(t) - \Delta x_{safe} + \tau \dot{x}_n(t) \tag{7}$$

where the nth and $(n-1)$th vehicles are follower and leader vehicles respectively, Δx_{safe} is the safe distance, $\dot{x}_n(t)$ and $\dot{x}_{n-1}(t)$ are the velocities, τ is a sensitivity coefficient.

Differentiating equation (6) with respect to time, we obtain

$$\dot{x}_{n-1}(t) - \dot{x}_n(t) = \tau \ddot{x}_n(t) \tag{8}$$

$$\ddot{x}_n(t) = 1/\tau[\dot{x}_{n-1}(t) - \dot{x}_n(t)] \tag{9}$$

From Eq. (8), five generations of models for the sensitivity coefficient term exist, and the GM model has the form

$$\ddot{x}_n(t) = \left[\frac{\alpha_{l,m}[x_{n-1}(t)]^m}{[x_{n-1}(t) - x_n(t)]^l}\right][\dot{x}_{n-1}(t) - \dot{x}_n(t)] \tag{10}$$

Therefore, the equations of a car-following model can be developed as below, where T_n represents the reaction time of the follower, and t is the updating time. Then the equations can be written as

$$\ddot{x}_n(t + T_n) = \lambda[\dot{x}_{n-1}(t) - \dot{x}_n(t)] \tag{11}$$

$$\ddot{x}_n(t + T_n) = \left[\frac{\alpha_{l,m}[\dot{x}_{n-1}(t + T_n)]^m}{[x_{n-1}(t) - x_n(t)]^l}\right][\dot{x}_{n-1}(t) - \dot{x}_n(t)] \tag{12}$$

2.3 Car-Following and Macroscopic Model "Bridging" Relation

The car-following model has to be connected, or "bridged," to suitable macroscopic models in order to determine the road capacity reduction on a degraded road environment, due to potholes especially. The assumption is to consider a leader–follower scenario and after that a platoon of vehicles [2]. Fundamental diagrams are models that describe the relationship between speed and density (see Fig. 1), flow and density (see Fig. 2), and speed and flow. Fundamental diagrams express macroscopic traffic flow models coming mainly from observed road traffic data. These macroscopic traffic stream models, developed historically by Greenberg [17], Underwood [18], and Drake et al. [19], show a nonlinear relationship between density and speed. The GM model is selected due to its capability of representing both microscopic and macroscopic traffic flow conditions and its characteristics of representing several various models with different parameters. The "bridging" relation could be developed as explained below.

From Eq. (12) and with $m = 0$ and $l = 1$ (m and l are parameters of the Greenberg model), we get the GM3 model:

$$\ddot{x_n}(t + T_n) = \alpha \left[\frac{x_{n-1}(t) - x_n(t)}{x_{n-1}(t) - x_n(t)} \right] \tag{13}$$

And by integrating Eq. (13) on both sides with respect to time t, we obtain the following equation:

$$\dot{x}(t + T_n) = \alpha ln|x_{\mathrm{n}-1}(t) - x_{\mathrm{n}}(t)| + C_n = \alpha ln|S_n(t)| + C_n \tag{14}$$

where C_n is an integration constant for the vehicle under an equilibrium or steady-state condition, i.e., there is no difference among the drivers' identities n, all cars are traveling at the same speed, and there is the same space (expressed in the form of time headway) between them. Also, the speed doesn't depend on time and $\dot{x_n}(t + T_n)$ is the same as the traffic speed v. In this situation, spacing S becomes the reciprocal of traffic density k, i.e., $S \longrightarrow \frac{1}{k}$. Thus, Eq. (14) can be reduced to

$$v = \alpha ln \frac{1}{k} + C. \tag{15}$$

An interesting interpretation of Eq. (15) is the prevailing traffic condition, here density k; thus drivers are required to adjust their speed, i.e., $v = V(k)$. The relationship between traffic speed and density can be roughly represented as the red curve, which is opposite to the dashed-line blue curve of the Greenberg model (see Fig. 1).

The speed–density curve possesses two known points: $(v_f, k_c 0)$ and $(0, k_j)$. Putting the two points into Eq. (15), we get the following speed expression:

$$v = \frac{v_f}{ln \frac{k_j}{k}} ln \frac{k_j}{k}. \tag{16}$$

Equation (16) represents the traffic speed in the Greenberg model as derived from the GM model. This is the "bridge" relation, which appropriately expresses the traffic speed.

We present the speed–density relation with two different curves. The red curve is an approximation with empirical data, and the dashed-line blue curve is a fine-tuned representation of a Greenberg model. It is also known that traffic density and traffic speed vary respectively between $0 \le k \le k_j$ and $0 \le v \le v_j$. k_c is the critical density; it allows drivers to maintain their desired speed, and hence the free-flow speed v_f can be sustained up to a k_c density level.

Thus, Eq. (15) became Eq. (17), and in the range $0 \le k \le k_j$, $\alpha = \frac{v_f}{ln(k_j/k_c)} = v_m$,

$$\begin{cases} v_m = v_f \quad & when \quad 0 \le k \le k_e, \\ \frac{v_f}{ln\frac{k_j}{k}} = v_m \quad & when \quad k_c \le k \le k_j. \end{cases} \tag{17}$$

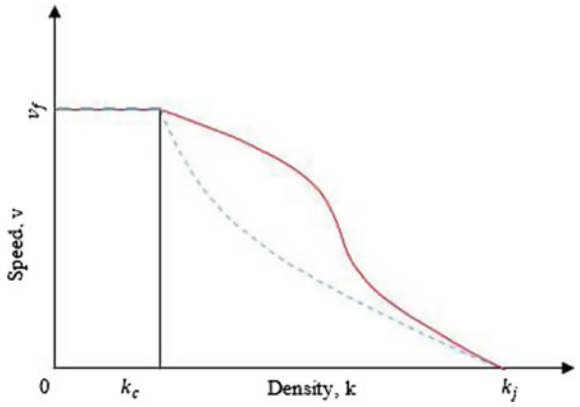
Figure 1 illustrates the variation of speed with respect to density, from a "free flow" speed state to a congested state (congested state means at jam density, where the speed of each car is near zero). The critical density appears when in the car-following process, a follower driver is not driving or is unable to drive in "free flow" mode. The corresponding illustration of Fig. 1 in a fundamental flow–density diagram is described in Fig. 2 (see Eq. (18)):

$$\begin{cases} q_m = v_f k \quad & when \quad 0 \le k \le k_e, \\ q_m = \frac{v_f}{ln\frac{k_j}{k}} \quad & when \quad k_c \le k \le k_j. \end{cases} \tag{18}$$

From Eq. (1), where $q = kv$, and Eq. (17), we get Eq. (18).

Flow is a linear function in the interval $0 \le k \le k_c$. After traffic density grows beyond k_c and flow continues to increase until it reaches the maximum flow (q_m), then q is decreasing until the density reaches its maximum (k_j).

Fig. 1 Speed–density relationship diagram



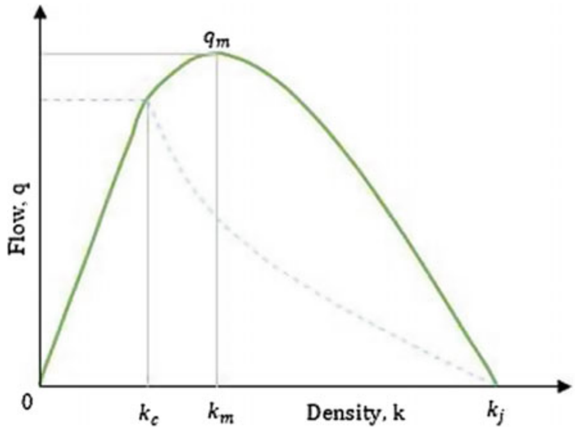


Fig. 2 Flow–density relationship diagram

To find the capacity (q_m) and conditions for its existence, we derive as follows:

$$\frac{dq}{dk} = 0 = \frac{v_f}{ln\frac{k_j}{k_c}}(ln\frac{k_j}{k} - 1),$$

$$ln\frac{k_j}{k} - 1 = 0 \Longrightarrow k = k_m = \frac{k_j}{e},$$

where k_m is the density at capacity q_m, and $k_m = \frac{k_j}{e}$,

$$q_m = \frac{v_f}{ln\frac{k_j}{k_c}} ln\frac{k_j}{k_m} k_m = \frac{v_f}{ln\frac{k_j}{k_c}}\frac{k_j}{e};$$

q_m exists if $\frac{d^2q}{dk^2} < 0$,

$$\frac{d^2q}{dk^2} = \frac{v_f}{ln\frac{k_j}{k_c}} ln\frac{k}{k_j} = \frac{v_f}{ln\frac{k_j}{k_c}}\frac{1}{k},$$

where $k_c \leq k_m = \frac{k_j}{e} \leq k_j$; thus $0 \leq ln(e) \leq \frac{k_j}{k_c}$.

The flow–density relationship can be plotted as the green curved illustrated in Fig. 2.

The relation provided above by both Greenberg and GM3 shows that it might be possible to relate some of the existing equilibrium traffic flow models to a GM generalized model by aggregating or integrating the model with varying speed and spacing exponents. Table 1 summarizes the link between GM (see Eq. (12)) for the car-following model and the macroscopic relation for different speed (m) and spacing (l, n) exponents for various models (e.g., Pipes–Munjal, l=n + 1; Drew, l=n + 1.5).

Table 1 Car-following model and macroscopic traffic flow model bridge [Source: [2]]

Model	l;m	**Traffic speed**
Greenberg	$l = 1; m = 0$	$v = \frac{v_f}{ln\frac{k_j}{k_c}} ln\frac{k_j}{k}$
Greenshields	$l = 2; m = 0$	$v = v_f - \frac{v_f}{k_j}k$
Underwood	$l = 2; m = 1$	$v = v_f e^{\frac{-k}{k_m}}$
Drake	$l = 3; m = 1$	$v = v_f e^{\frac{-1}{2}(\frac{k}{k_m})^2}$
Pipes-Munjal	$l = n + 1; m = 0$	$v = v_f([(1 - (\frac{k}{k_j})^n])$
Drew	$l = n + 1.5; m = 0$	$v = v_f([(1 - (\frac{k}{k_j})^{n+\frac{1}{2}}])$

The “bridge” between microscopic and macroscopic models under equilibrium conditions supposes that all vehicles are equally separated in space and traveling at the same speed. With this consideration, the driver identity n or $n - 1$ is dropped. The GM model (Eq. (12)) expresses the generalized car-following model with model parameters l, m at the microscopic level. The choice of different values of the parameters in the Greenberg, Greenshields, Underwood model, etc., is at the macroscopic level. This model constitutes a unifying factor for connecting microscopic to macroscopic models in equilibrium state.

The choice of $l = 1$, $m = 0$, as illustrated in Eqs. (13)–(16), determines traffic speed, as shown in Fig. 1, representing the Greenberg model and the corresponding traffic flow shown in Fig. 2. The same procedure will lead to other traffic speed formulas.

2.4 System Dynamics: A Brief Introduction

Grounded in the theory of nonlinear dynamics and feedback control developed in mathematics, physics, and engineering, system dynamics (SD) models are built to solve complex problems and to understand the nonlinear behavior of complex systems over time. Thus, in SD models, human behavior and physical and technical systems are (can be) simultaneously considered as displaying an interdisciplinary characteristic. Components such as stocks, flows, converters, internal feedback loops, and time delays are used for system modeling and simulation [10].

In SD, a stock represents a part of a system whose value at any given instant in time depends on the system’s past behavior. The value of the stocks at a particular instant in time cannot simply be determined by measuring the value of the other parts of the system at that instant in time; the only way you can calculate it is by measuring how it changes at every instant and adding up all those changes. Thus, flows represent the rate at which the stock is changing at any given instant, they either flow into a stock or flow out of a stock. Converters either represent parts at the boundary of the system or parts of a system whose value can be derived from other parts of the system at any time through some computational procedure [10].

In this work, the Stella SD software tool (it is one SD tool among many others) has been used to perform the SD modeling of the case study at hand, namely the one addressing the shortcomings in the comprehension of the impact of potholes on traffic microscopic scenario and to reflect space, speed, and deceleration/acceleration in the presence of potholes within a microscopic traffic simulation scenario.

3 COMPRAM Methodology for Handling Complex Societal Problems

Handling complex societal problems, a branch of operations research, generally focuses on all phases of the problem-handling process from awareness of the problem to evaluation of potential interventions [20]. The process of handling a problem is divided into two subcycles: (a) defining the problem and (b) changing the problem (see Table 2). Defining the problem is to define a conceptual model by describing the problem, defining relevant domain(s), and finding relevant concepts of the problem, the main phenomena as well as the relations between phenomena, and theoretical ideas based on the description of the problem. The second subcycle, with the conceptual model as the entry point, begins with the construction of the empirical model of the problem and ends with the evaluation of the interventions.

The COMPRAM methodology, which is mainly based on societal complexity theory, offers a more structured and transparent way of optimally handling real-life and real-world complex societal problems. According to DeTombe, the inventor of the COMPRAM methodology, handling *"means to find out what is going on, finding the causes, indicating possible interventions, implementing interventions and evaluating the process and the outcome of the problem handling process."* Handling complex societal problems involves many phenomena and actors with different views of the problem, different interests, and different "solutions" in mind. Thus, the methodology is multidisciplinary, multilevel, as well as multiactor.

The creator of the methodology argued also that "problem handling" differs significantly from "problem solving" because solving refers to a certain desired goal, since the desired goal potentially (and probably often) differs from actor to actor [7, 8, 20]. Since most societal problems concern various actors and stakeholders, the best "problem handling" should take those various respective perspectives into consideration. Such a nuancing is not provided by a standard "problem solving" approach and involves a much higher complexity.

COMPRAM, which improves the handling of complex societal problems, provides guidelines starting from the conceptual model and six steps of the methodology.

The first outcome of handling a complex societal problem using the COMPRAM methodology is to "make a conceptual model" of the problem, i.e., a clear definition of the problem at stake, and thereby to avoid a hazardous handling of the problem.

Table 2 Following model and macroscopic traffic flow model bridge

The first subcycle of the problem handling process **defining the problem**	
Phase 1.1	Becoming aware of the problem and forming a (vague) mental idea of the problem
Phase 1.2	Extending the mental idea by hearing, thinking, reading, talking, and asking questions about the problem
Phase 1.3	Extending the mental idea by hearing, thinking, reading, talking, and asking questions about the problem
Phase 1.4	Forming a problem-handling team and starting to analyze the problem
Phase 1.5	Gathering data, exchanging knowledge, and forming hypotheses about the problem
Phase 1.6	Formulating a conceptual model of the problem
The second subcycle: **changing the problem**	
Phase 2.1	Constructing the empirical model and the desired goal
Phase 2.3	Defining the handling space
Phase 2.4	Suggesting interventions
Phase 2.5	Implementing interventions
Phase 2.6	Evaluating interventions

The "conceptual model" comprehensively structures the problem from the process of defining it up to the simulation of the obtained problem-related model.

The conceptual model has been structured (in COMPRAM) as a seven-layer model (Fig. 3). It begins by describing the problem in text form as the first layer. Retrieved concepts and phenomena from the text constitute the second layer. A reflection is made on the knowledge status based on hypotheses, theories, experience, intuition, or assumption through verbal description; that constitutes the third layer. A further step explains the influence of the concepts on the phenomena or vice versa, and a graphical representation of the knowledge; this is performed in the fourth layer. In layer five, a semantic model represents graphically the relations between the concepts and the phenomena. And in layer six, a causal model is provided, which is the graphical representation of the causal relations from layer five. In layer seven, the system dynamical simulation of the problem-related developed system model is performed through an SD computer software tool such as Stella/Ithink/PowerSim.

The COMPRAM methodology includes six steps: knowledge, power, negotiation, societal reaction, implementation, and evaluation (see Table 3).

Knowledge, power, and emotion are core elements of the COMPRAM methodology in the management of complex societal problems. "Knowledge" is about having information of different aspects of the problem by providing an integrated simulation model (see Fig. 4). Indeed, each expert has knowledge of a part of the problem. We need "power" to reach an agreement on the definition of the problem and to select the interventions. Influence of emotions can stimulate or obstruct cooperation between people and between groups [21].

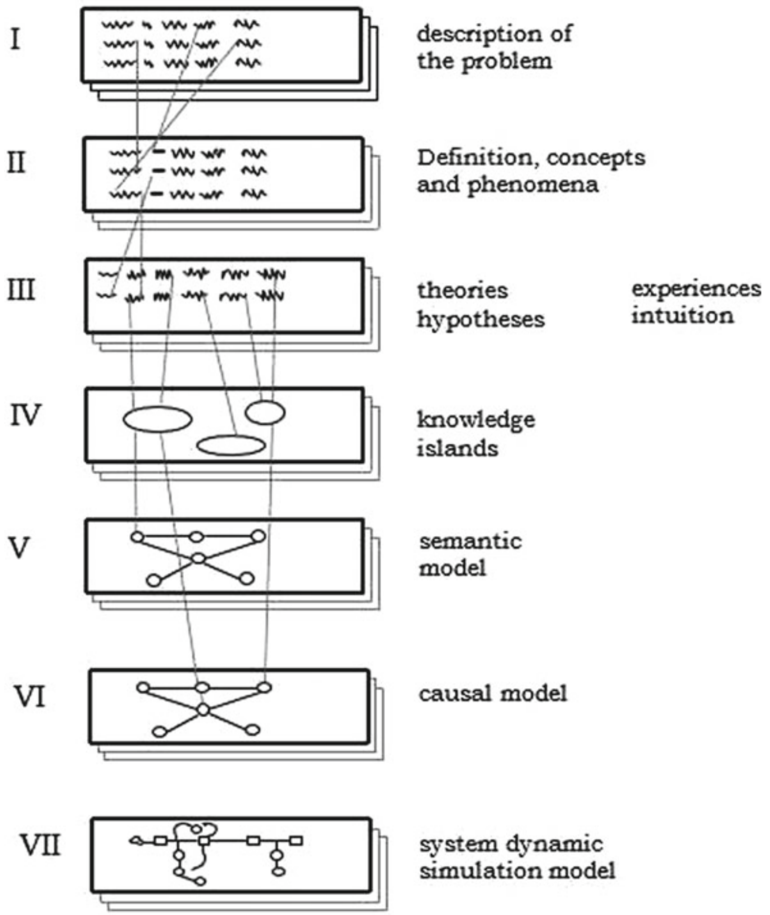


Fig. 3 The seven-layer communication model of the COMPRAM method [Source: [21]]

Table 3 The six steps of the COMPRAM methodology [Source [21]]

Step 1	Analysis and description of the problem by a team of neutral content experts (knowledge)
Step 2	Analysis and description of the problem by different teams of actors (power)
Step 3	Identification and negotiation of interventions by experts and actors
Step 4	Anticipation of the societal reactions
Step 5	Implementation of the interventions
Step 6	Evaluation of the changes

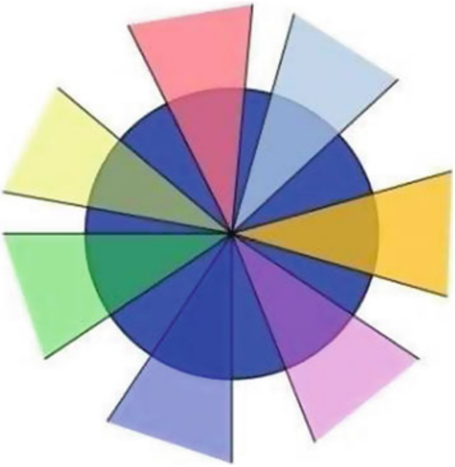


Fig. 4 Each expert sees only part of the problem from his respective perspective [Source [7]]

It is known that the COMPRAM methodology is an iterative process. In this study, for the first COMPRAM-based assessment regarding degraded roads with potholes, we focus on the initial assessment by describing and/or defining the "**road potholes problem**" while targeting the effects of potholes, which could lead to or contribute to "**road congestion as a complex societal problem**" (one should note that potholes are not the only cause of congestion, one of the other major causes being increasing traffic beyond available road capacity).

This study was conducted to obtain initial knowledge (i.e., preliminary information) on the issue. As a complex societal problem, other aspects must be evaluated in considering the multidisciplinary and interdisciplinary flavors of the problem, all actors involved, and so on. Finally, in dealing with such a complex societal problem, the COMPRAM methodology has to be used for defining and improving the process by which road potholes are handled, namely, what actions are to be taken to achieve specific goals from the perspectives of different actors and stakeholders.

4 COMPRAM in the Framework of Applicability to Understanding a Complex Societal Problem: "Traffic Congestion in Kinshasa Linked to Degraded Roads"

4.1 Origin, Formation and Phenomena in Producing Road Potholes

A pothole is a sharp-edged hole in a road, usually caused by water (which is a major cause of other weather-related damage) or related to poor construction. Regarding its civil engineering construction, a road is divided into three major layers: soil, subbase,

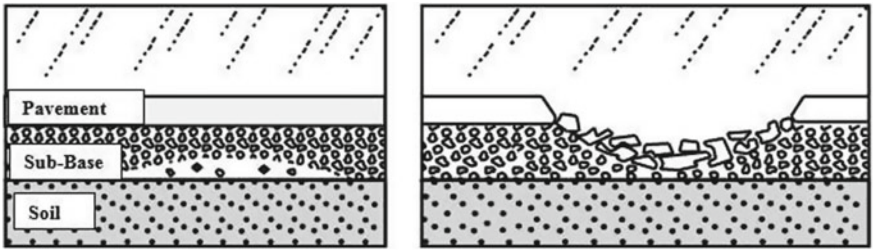


Fig. 5 Road layers and pothole formation due to water penetration [22]

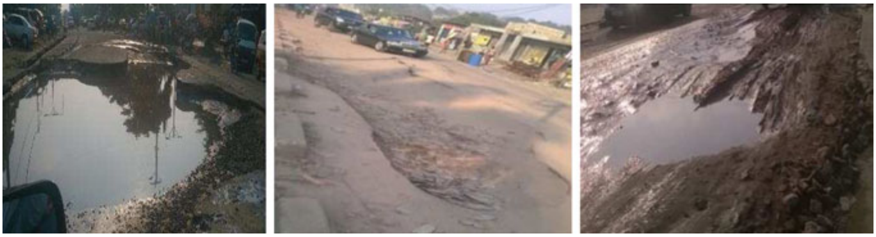


Fig. 6 Road potholes in Kinshasa (Photograph taken by A.K. Kayisu on 13 December 2015)

and pavement. A pothole arises when water penetrates the pavement, reaches its base, and creates a small cavity. The unsupported pavement then falls away (due to the small cavity), creating a pothole (see Fig. 5).

The potholes considered here are classified as *carriageway potholes*, with dimensions of more than 300 mm across and 40 mm of depth, while a *footway pothole* must be deeper than 20 mm and more than 150 mm across. The pothole could be rocky, muddy, full of rainwater, or very deep, causing a time delay for a vehicle trying drive through it (see Fig. 6).

The formation of potholes differs with different kind of pavements, i.e., surfacing types, asphalt or thin bituminous seals. In the asphalt case, the structure of pavement and the material used for road construction or repairs have an impact on the formation of potholes. Another source of potholes is poor road design over certain subgrades (i.e., expansive, collapsible, and dispersive), resulting in a negative influence of heavy traffic, with overloaded trucks and heavy vehicles causing the road to weaken. The repetition of this type of heavy traffic over small potholes causes cracking of asphalt as a result of fatigue of road pavement due to poor support.

As cracking increases, more water gets into the structure, the material shear strength drops, failure occurs rapidly, and water penetration into the asphalt causes separation of the asphalt from the underlying layer. This explains why more potholes are created in rainy seasons compared to dry seasons, for example in the city of Kinshasa. Environmental cracking of asphalt is also caused by ultraviolet light from the sun, heat, and oxidation, all resulting in shrinkage of the asphalt over time. In the case of thin bituminous seals, cracking of the seal allows water into the road, and potholes propagate downward. Some causes of this are bitumen age, fatigue under the load of heavy vehicles, and lack of adhesion between thin seals and the base course.

4.2 State of Traffic Congestion and Degraded Road Problem Due to the Potholes in Kinshasa

The mobility of persons and goods constitutes a basic economic and social need for every country worldwide. Intelligent transportation systems (ITS) were proposed to assist and optimize the technological solution of this need in a given region while avoiding or reducing traffic congestion through suitable and optimized road traffic management. Traffic congestion is a phenomenon that occurs either when the demand is greater than the supply in terms of road capacity or when an unusual event occurs such as an accident or a truck breakdown on the road, resulting in a significant "temporal" capacity reduction. Other contributors to congestion are roads littered with potholes as well as suboptimal driver behavior (such as a misunderstanding between drivers at a crossroads, for example).

From field observations it is well known that road traffic is a complex phenomenon, especially in crowded large cities like Kinshasa (i.e., in a country like the Democratic Republic of Congo (DRC), particularly in the capital city Kinshasa). Traffic in Kinshasa suffers from a lack of appropriate strategic traffic management and the nonexistence of real-time traffic control. Similarly to what has been noted in India [5, 6], the city of Kinshasa experiences daily recurrent traffic-related problems: congestion, degraded roads, lack of road maintenance, inappropriate driver behavior, etc.

Congestion in Kinshasa is due mainly to very limited road capacity, road surface conditions, and chaotic (i.e., undisciplined) driving behavior. At least one or a combination of these reasons can lead to severe congestion states. The road network in Kinshasa consists of roads that are mostly in an advanced state of degradation, roads with neither street signs nor traffic lights, roads that are mostly more or less degraded. Further, there is no parking system worth mentioning that meets even minimal international standards. One can daily observe that traffic congestion also occurs when roads are interspersed with potholes, with vehicles, mostly in a car-following process or platoon-following process, significantly perturbed by the presence of potholes.

The poor construction quality of the roads, the poor and partial road monitoring and maintenance strategy, and the poor water drainage system, when combined with other aspects such as weak transportation legislation, the relatively long rainy seasons, all contribute to the road potholes problem, and the consequent "effective road reduction" and congestion phenomena. Also, most car drivers have never attended a driving school (where they could have learned about best practices in driving a car in road traffic). The result is chaotic and undisciplined driver behavior. The situation becomes even worse whenever chaotic driving behavior occurs in the presence of potholes: much greater congestion is the consequence of such situations.

On the other hand, administration officials in charge of road maintenance and traffic management are never in a hurry to solve problems of degraded roads within a reasonable amount of time according to the standards related to a given type of "road maintenance." Sometimes, they wait until a road has suffered advanced degradation to justify the large sums of money that will be needed for repairs.

It is known that water is one of the major causes of the formation of potholes, and this formation differs according to the types of pavement, the method of construction, and materials used. Other causes are poor road design, chemical spillage on the roads, effects of heavy traffic, and cracking of asphalt.

Traffic congestion in Kinshasa due to potholes is a complex societal problem with interdisciplinary aspects, and the use of methods such as COMPRAM can assist in a comprehensive handling of this societal problem, which should benefit the quality of life of those using the roads in Kinshasa. According to the founder of COMPRAM, the shared features of a complex interdisciplinary societal problem are *uncertainty about the starting point, development, and end of the problem; incomplete or not directly available knowledge and data about the problem; a large number of relevant actors (people, institutes, countries) with diverse/varying interests* [21].

The COMPRAM methodology improves the problem-handling process and increases the quality of interventions and therefore the quality of life [7]. The COMPRAM methodology has been involved in multidisciplinary studies at the micro, meso, and macro levels using selected methods, models, and tools from the field of the methodology of societal complexity [7].

The COMPRAM methodology has been successfully used in fields such as healthcare, economics, climate change, terrorism, large city problems (such as traffic, water supply problems), and large technological projects [7–9, 20]. In 2006, the Organization for Economic Cooperation and Development advised the use of COMPRAM to handle "global safety." The International Society on Methodology for Handling Complex Societal Problems is one of the communities that uses this methodology, and the Dutch Federation of Methodology in the Social Sciences has been involved in this enterprise since 1994, particularly the research groups on Complex Societal Problems and Simulation [7].

The real-life problem addressed in this study, namely traffic congestion due to road potholes, cannot be resolved by one field alone, since for sustainable change, interdisciplinary work is needed. Different scientific fields have to be combined for the comprehension (integrated knowledge) and handling of the problem (definition, analysis, description, and evaluation of intervention, etc.). Verbal language, visual language, mathematical and simulation models, all are tools and methods used among facilitators, experts, and actors from the above-mentioned fields to avoid miscommunication and misinterpretation for problem definition and handling.

4.3 "Community" Concerned by Road Potholes and Traffic Congestion

For an initial description of traffic congestion and road potholes, there is no doubt about the need to define all aspects of the problem. This is required for accumulating integrated knowledge from different perspectives and for coming up with adequate interventions. The combination of methods and tools from scientific and political fields such as engineering sciences (civil, traffic, and transportation engineering), social sciences, chaos theory, and legal science are needed to comprehensively handle the "potholes on the road" and the related traffic congestion problem for better road traffic management in Kinshasa.

The selection of appropriate disciplines and domains is based on traffic, civil engineering, sociology, law, and policy. The community composed of facilitators, experts, and actors involved in the maintenance of roads (i.e., of relevance to the pothole problem) and in the traffic congestion problem in Kinshasa is diverse and includes academic researchers, people living around pothole-plagued roads, road users, companies and agencies in charge of road construction, maintenance, and transportation, as well as various state organizations. The knowledge of traffic congestion induced by the phenomenon of road potholes is held by experts from a number of different professions.

In different domains, a list of professions has to be carefully selected to form the community in charge of road potholes and traffic congestion management (see Table 4). In the traffic domain, we select the profession of "traffic engineer." In the discipline of civil engineering, we select the profession of "road infrastructure engineer." In the social sciences, we select the profession of "phenomenology analyst." In legal science, we select the profession of "traffic and transportation legal expert." Facilitators should be persons with a sufficient background in management, social sciences, and road traffic fundamentals.

The community comprises people and companies or national agencies. Here we identify components that should help in both problem definition and problem handling. A deeper description of actors and experts will ensure an improvement in understanding the road pothole problem and the related congestion problem in the Kinshasa city case study.

For comparison purposes, two different sections of a road are considered, one with potholes and another without. Vehicles take more time on roads with potholes, and speeds are affected. Degraded roads caused by the pothole phenomenon leading to traffic congestion need an overall study in order to identify impacts on other components involved in traffic systems. Apart from civil engineering works for filling and repairing potholes in due time, an adequate policy intervention is needed for sustainable change.

Figure 7 shows the negative impact of potholes in the reduction of traffic streams' speed and consequently on road capacity. Potholes on roads reduce the speed of cars both in microscopic scenarios in which cars are in a close car-following process and in macroscopic scenarios in which cars are even in nearly free-flow processes,

Table 4 Community involved in defining and handling the "road potholes and traffic congestion" problem in Kinshasa/DRC

Experts and actors	Mission or background
Traffic engineer	Strong knowledge of traffic systems's science, specifically of road capacity and congestion problems in general and specifically in Kinshasa
Civil engineer	Strong knowledge of Kinshasa's road infrastructure and related problems related mainly to congestion and infrastructure. The expert must have a strong knowledge of the historical development of the road infrastructure in Kinshasa
Social science	A university researcher with a good knowledge and understanding of how people living in Kinshasa react to diverse phenomena of societal relevance
Lawyer	This expert has a profound knowledge of special legislation related to roads, traffic, and transportation in DRC and especially in Kinshasa
Road users	This group includes car and truck drivers, motorcycle drivers, and even pedestrians as well as persons who daily experience recurring congestion
CNPR* (Commission Nationale de Prévention Routière, the DRC Road Safety Agency)	An Engineer or a road safety expert with solid background in road traffic engineering, road geometry, and road markings, safety prevention measures, etc.
Office des routes, the DR Congolese Agency for intercity road maintenance	A manager from the administration in charge of collecting funds from road tolls and other similar taxes destined for financing various road maintenance and construction projects (tax collector and road project manager)
OVD*, the DR Congolese agency in charge of intracity road maintenance	An engineer with profound knowledge of the road water drainage system of Kinshasa city, and of roads maintenance policies and standards
ACGT* (Agence Congolaise des Grands Travaux, the DR Congolese agency in charge of managing large road and transportation-related infrastructure projects)	An engineer with broad experience in managing construction and modernization projects of road infrastructures
PCR* (Police de Circulation Routière, the traffic police in charge of traffic related legal enforcement)	A senior Traffic Police officer for road traffic legal enforcement
National ministries in charge of transportation and finances	A senior officer from the DR Congolese Ministry in Charge of Transportation and Road Management; a second senior officer from the DR Congolese Ministry in Charge of Finances

* Companies or national agencies

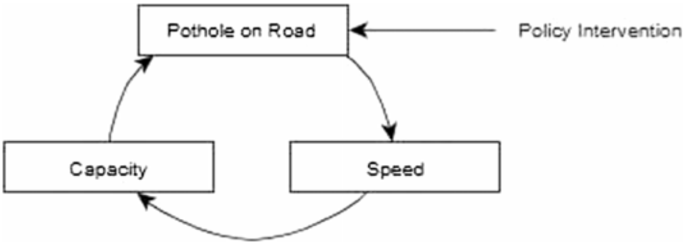


Fig. 7 Potholes on roads and consequent negative impact on road capacity and speed

because in both cases, whenever a driver encounters a pothole, the tendency is to reduce speed, since otherwise, the car will be damaged on hitting the pothole. At the macroscopic level, the road capacity is directly affected, leading to an effect on the traffic's macroscopic variables: flow, density, and average speed. Thus, a policy intervention is an appropriate mechanism for pothole elimination.

Potholes also have an effect on variables such as travel time (increase) and on further components such as vehicles' health (depreciation, higher probability of car defects), increased rate of accidents, and chaotic/undisciplined driver behavior (see Fig. 8). This process demonstrates also the effect of potholes on roads and related consequences on capacity, speed, and further on related car operation costs (i.e., gas consumption, lost time, etc.), leading to congestion whenever the resulting car density comes near to or goes beyond $\frac{1}{2} \times k_j$ (note: congestion starts when the density becomes greater than half of maximum density k_j) on a given road section.

Some of the possible interventions are improvement of road maintenance, permanent road state monitoring, and appropriate legislation to enforce good road maintenance policies and standards. Further, bribery avoidance in the framework of (mainly outsourced) road construction and maintenance projects can significantly improve the quality of road construction. Figure 8 integrates all the above-described causal relationships into a global causal effect diagram and policy intervention. This causal loop diagram illustrates interdependencies and feedback processes responsible for the effects of the road pothole problem and some considerations that could contribute to its resolution [10]. The resolution takes into account the implication of stakeholders such as different actors, policy-makers, road executives, drivers, and driving instructors.

The effect of potholes on all the above identified variables (i.e., speed, capacity, lost time, accidents, driver behavior, vehicle depreciation, accident and other costs) confirms the complex multidisciplinary nature of the problem.

From Fig. 8, it is seen that the causal loop diagram is needed to describe the structure of the road pothole system. An analysis of this causal loop diagram shows that the presence of potholes, which is an independent variable in relationship to other independent variables, produces various effects (positive (+) or negative (−)) on them. The diagram shows also that the presence of potholes reduce capacity and

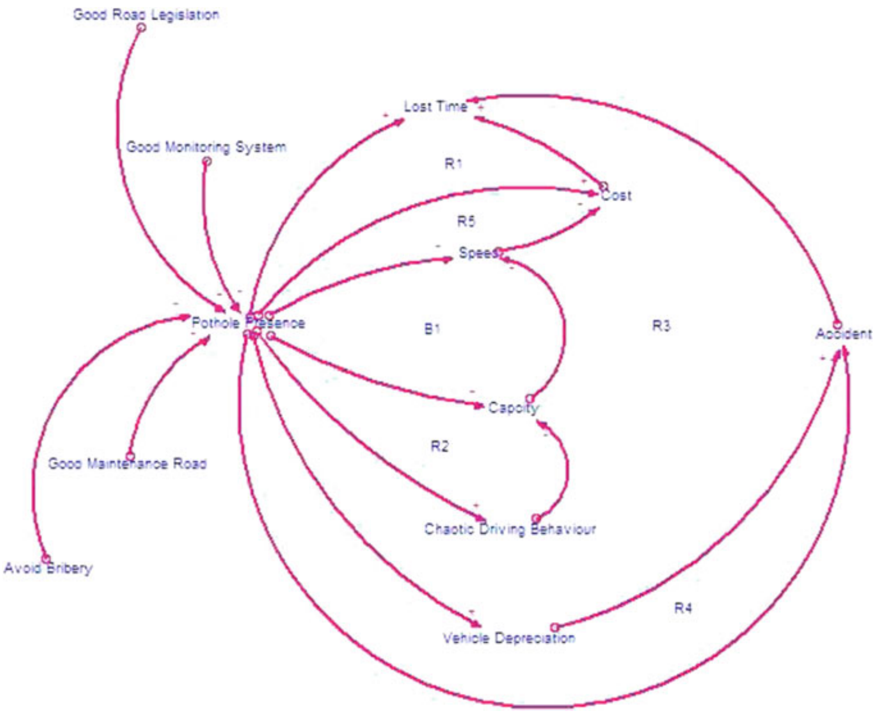


Fig. 8 Causal loop diagram of both effects of pothole presence and policy intervention

the speed of vehicles. Another piece of information is that more potholes reduce both traffic flow and average travel speed. That is why the loop polarity of "loop B1" is negative.

Another issue is lost time and the cost of an increase in fuel consumption. By observing the positive signs of different links, these effects are amplified with an increase in the number of potholes. Chaotic driving behavior is also caused by the presence of potholes, since drivers have a tendency to avoid potholes, and this raises the probability of vehicular collisions. Potholes have an amplifying effects on accident variables, lost time, and deterioration of cars. That is why the polarities of the loops R1, R2, R3, R4, R5 are positive.

The situation described above leads to congestion. Interventions such as good road legislation, good road monitoring systems, good road maintenance, and the avoidance or prohibition of bribery should decrease the incidence of potholes, decrease lost time and costs, and also increase traffic flow (speed and capacity) as well as the average travel speed.

Simulation of potholes' effect microscopic traffic parameters within a car-following process reinforces the core hypothesis that potholes on roads effectively and significantly reduce traffic flow and the effective road capacity and thereby lead to congestion.

5 System-Dynamics-Based Modeling and Simulation of a Car-Following Process on a Pothole-Degraded Road

In this modeling case, the car-following model considers the initial context of a single-lane road having respectively one and then two potholes, and three vehicles driving on that road. No lane change is authorized during this driving observation, which means that the vehicles remain in the same lane. The first vehicle goes through a pothole as the lead vehicle, and a second lead vehicle follows. Then comes the third and last vehicle as a follower vehicle. The second vehicle is a follower of the first one and is the lead vehicle for the third. A driver follows another vehicle by judging distance, speed difference. Each driver has a particular reaction time. The road condition, which may be good or bad, as is the case here, where it is marked with potholes, is another influencing factor in the stimulus–response equation.

The basic difference between this model and existing car-following models is that existing car-following models consider each vehicle pair separately, whereas in this model, several vehicles are considered simultaneously. Furthermore, this car-following model exploits the concept of a "desired safe spacing or operating speed" that depends on pavement conditions in general and on the presence of potholes in particular. This last-named concept lies at the center of our formulation.

A pothole can be rocky, muddy, full of rainwater, very deep, etc., each such feature impacting the time-delay of a vehicle attempting to through it. Regarding the pothole impact, more travel time will be required to pass it compared to a road in good condition, and this will influence the macroscopic traffic stream in general. The abrupt speed reduction resulting from the presence of a pothole generates a shock wave, which can produce a relatively long queue (i.e., congestion) of several meters within a very short time.

One of the mathematical equations of relevance to describe the situation is Eq. (5). But here, in the presence of an obstacle, which is the pothole, the car-following model's algorithm on a degraded road should be redefined accordingly, mainly taking the pothole's characteristics and the related impact on the driver into account. Thus, a comprehensive modeling and simulation involving SD tools will be used to capture all related system-behavioral aspects.

The novel SD-based context modeling and computer simulation of the pothole-related driving situation should take the following facts into account:

1. The driver's reaction to change in speed of the front car (vehicle ahead or lead vehicle) will occur after a reaction time, which is a time gap during which the follower perceives the change in speed and then begins to react to it.
2. The vehicle's motion parameters (i.e., position, speed, and acceleration) will be updated with a given frequency depending on the accuracy required (whereby a higher updating frequency leads to greater accuracy).
3. The vehicle's position and speed are governed by Newton's laws of motion, but the vehicle's acceleration is governed by the car-following model.

Therefore, the governing equations of traffic flow can be derived as explained below. We consider Δt to be the reaction time, and Δt to be the updating time interval of the vehicle's motion parameters.

The governing equations can be written as follows:

$$v_L^t = v_L^{t-\Delta t} + a^{t-\Delta t} \times \Delta t, \tag{19}$$

$$x_L^t = x_L^{t-\Delta t} + v_L^{t-\Delta t} \times \Delta t + \frac{1}{2} a_L^{t-\Delta t} \times \Delta t^2, \tag{20}$$

$$a_F^t = \frac{[\alpha_{l,m}(v_F^t)^m]}{(x_L^{t-\Delta t} - x_F^{t-\Delta t})}[v_L^{t-\Delta t} - v_F^{t-\Delta t}], \tag{21}$$

where v_t^L, x_t^L stand respectively for speed and distance traveled by the leader vehicle during the simulation interval Δt; a_F^t is a response (acceleration or deceleration of a follower vehicle to a leader vehicle; it is a function of sensitivity and stimuli. Stimuli are represented by the relative velocity, and sensitivity is represented by the driver's sensitivity. A sensitivity coefficient reflects the characteristics of a driver (driver sensitivity). The driver does not respond directly to relative speed. The follower driver reacts to the lead vehicle mainly to the probability of avoiding a rear-end collision by accelerating or decelerating. The driver takes into account both relative speed and the estimated risk of collision.

Equations (19) and (20) are a simulated version of Newton's law of motion for the leader vehicle regarding respectively speed and spacing (location). The acceleration of the follower vehicle (see Eq. (21) depends on the following values: the relative velocity between leader and follower vehicles, a sensitivity coefficient, and the (spatial) gap between the vehicles.

5.1 Proposed SD Car-Following Model Considering Pothole Influences

We propose a car-following model that replicates the behavior of the follower vehicles' drivers. This proposed model considers endogenously the speed, the acceleration/deceleration of the follower vehicles, and the respective spacing between them and their respective lead vehicles. Road geometry and pavement conditions have an impact on the speed profiles. The proposed car-following model determines internally the acceleration/deceleration rates, speeds, and the spacing profiles of the follower vehicles on the basis of the behavior of downstream lead vehicles as these are influenced by the potholes.

Concerning underlying assumptions, it is assumed that the acceleration/ deceleration rate of a follower vehicle's driver depends on the current speed, the control speed, and the perception–reaction time of the lead vehicles' drivers. And the acceleration/deceleration rate of the first lead vehicle depends on the current speed and the influence of the pothole on the current speed. A decrease of speed begins when the driver of the first lead vehicle sees the pothole. And due to the resulting shock wave, the speed also decreases for the second lead vehicle, similarly for the further following vehicle. The control speed is defined as the maximum speed at which the following vehicle's driver might travel given the spacing and rate of change in spacing with respect to the lead vehicle immediately ahead of it.

5.2 Stock-Flow Diagram and Simulation of a Car-Following Model on a Road with Potholes

The stock–flow diagram responds to the limitations of the causal-loop diagram, which is rather just a mental model, in that it captures stock and flow structures of the system's behavior. The terms "stock," "flow," and "feedback" characterize the state of the system with information generated for decision or action. Stocks express accumulations, which can be altered by inflow or outflow and can represent, for example, in general, an account balance, number of people in a company, or a vehicle's speed [10].

The proposed car-following model consists of four sectors: first lead vehicle, second lead vehicle, following vehicle, and spacing. Functions in each sector interact with functions in the other sectors through feedback links. The first and second lead vehicles' speed profiles are specified externally and prescribe a lead vehicle's conditions for input into the following vehicle's speed and spacing sectors.

The simulation assumptions are the following:

- Vehicles drive with a rectilinear motion uniformly accelerated (RMUA) before their drivers discover the pothole.
- Vehicles drive with a rectilinear motion uniformly decelerated (RMUD) when their drivers see the pothole. RMUA or RMUD mean that the trajectory of the vehicle is a straight line and acceleration is constant.
- Speed reduction depends on the characteristics of potholes (namely their respective length, depth, and width).
- Vehicles enter into the pothole with an RMUD.
- Vehicles exit the pothole with an RMUA.

5.3 Stock-Flow Diagram of a Car-Following Model of a Road with a Single Pothole

The simulation experiment is performed using the software tool Stella, in which the SD model for this first scenario with one pothole on the roadway has been implemented. All variables are implemented to reflect equilibrium in all stocks (See Fig. 9).

Table 5 Aggregated data on the road with one pothole; data related to the simulation presented in Figs. 10 and 11

Simulation time	Motion mode	Distance covered by first lead (m)	First lead vehicle speed (m/s)	Percentage of speed variation of first lead (%)
$t = 0\,\text{s}$ to $t = 20\,\text{s}$	MRU	205	10	0
$t = 20.25\,\text{s}$ to $t = 36\,\text{s}$	MRUD	102.7425	9.82 − 4.38	1.8− 56.2
$t = 36.25\,\text{s}$ to $t = 40\,\text{s}$ (within pothole)	MRUD	16.4975	4.35 − 3.92	56.5 − 60.8
$t = 40.25\,\text{s}$ to t = 80 s	MRUA	341.7975	4.08− 9.89	59.2 − 1.1

Initially, we assume that the speeds of the first lead, second lead, and follower vehicles are the same and approximately 10 m/s (36 km/h) as a realistic speed on a one-lane road in Kinshasa. From $t = 0\,\text{s}$ to $t = 20\,\text{s}$, cars are moving in MRU mode and acceleration is zero. In this 20-second period, the distances between the first lead and second lead as well as between the second lead and follower vehicle are 15 m each, which means that the spacing between the first lead and the follower vehicle is 30 m. The total road length is 672 m.

Furthermore, the first lead vehicle sees the pothole at $t = 20\,\text{s}$. It begins to decrease its speed until a certain value is reached (i.e., from 10 m/s to 4.38 m/s) in order to enter the 16 m pothole slowly at $t = 36\,\text{s}$. In the pothole, the speed is between 4.35 m/s and 3.92 m/s ($t = 36\,\text{s}$ to $t = 40\,\text{s}$). After $t = 40\,\text{s}$, the vehicle leaves the pothole, and its speed returns to a final value of about 10 m/s, as shown in Fig. 10.

The empirical data obtained from the simulation show that the presence of one pothole reduces the first lead vehicle's speed by **61**% (from 10 m/s to 3.92 m/s). Such a situation in the presence of more cars on the road, for example at peak hours, will lead to congestion.

If we consider the initial speed of 10 m/s and in MRU motion of the first lead vehicle, the entire distance would be covered in 67 seconds. Due to the pothole, the same distance is covered in 80 s. This assertion shows that the pothole adds 13 more seconds of travel time, and the average speed falls from 10 m/s to 8.4 m/s. It represents 16% more travel time and average speed depreciation. If the first lead vehicle is moving at the road's maximum speed of 13.9 m/s (50 km/h), the entire distance is covered in 48 seconds. It will be 40% more and **almost three times the average speed decrease**. The reduction in speed and more time taken are mainly due to the deceleration mechanism when a driver perceives the pothole and at the same time collects information about the pothole's characteristics. This information obtained by real-time observation will guide the driver to adapt his speed while approaching the pothole and after entering it. In Table 5, we summarize information in regard to the simulation process of a road of length 16 meters with one pothole, the distance covered by the first lead vehicle related to speed, and percentage of speed variation compared to the initial value of the speed, 10 m/s.

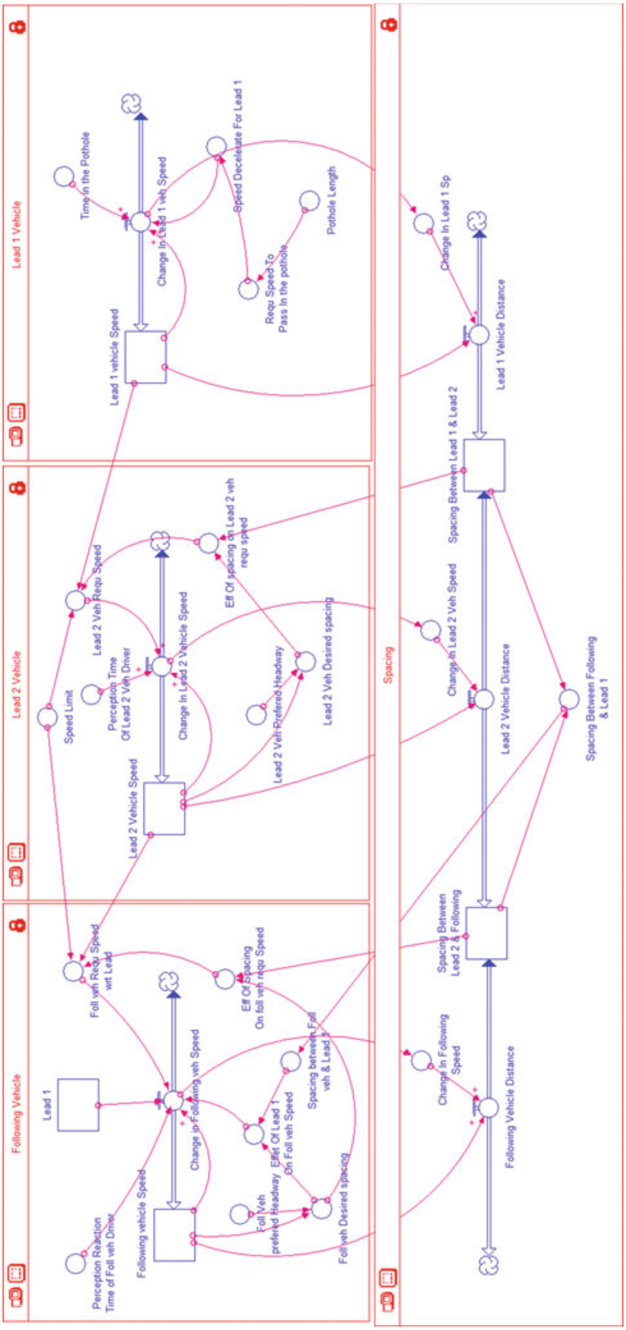

Lead 1 Vehicle
Time in the Pothole
Change In Lead 1 veh Speed
Speed Decelerate For Lead 1
Pothole Length
Requ Speed To Pass In the pothole
Lead 1 vehicle Speed
Lead 2 Vehicle
Lead 2 Veh Requ Speed
Eff Of spacing on Lead 2 veh requ speed
Speed Limit
Perception Time Of Lead 2 Veh Driver
Change In Lead 2 Vehicle Speed
Lead 2 Veh Prefered Headway
Lead 2 Veh Desired spacing
Lead 2 Vehicle Speed
Following Vehicle
Lead 1
Foll veh Requ Speed wrt Lead
Eff Of Spacing On foll veh requ Speed
Change in Following veh Speed
Spacing between Foll veh & Lead 1
Effet Of Lead 1 On Foll veh Speed
Foll Veh prefered Headway
Foll veh Desired spacing
Perception Reaction Time of Foll veh Driver
Following vehicle Speed
Spacing
Change In Lead 1 Sp
Lead 1 Vehicle Distance
Spacing Between Lead 1 & Lead 2
Change In Lead 2 Veh Speed
Lead 2 Vehicle Distance
Spacing Between Following & Lead 1
Spacing Between Lead 2 & Following
Change In Following Speed
Following Vehicle Distance

Fig. 9 Stock-flow SD model diagram of the scenario with one single pothole on the roadway

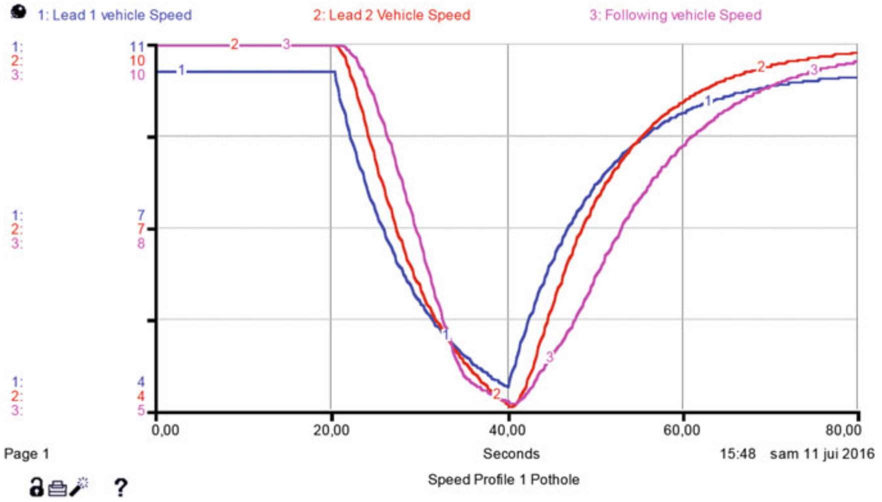


Fig. 10 Speed profile of the first lead, second lead and following vehicles, as obtained from the SD simulation

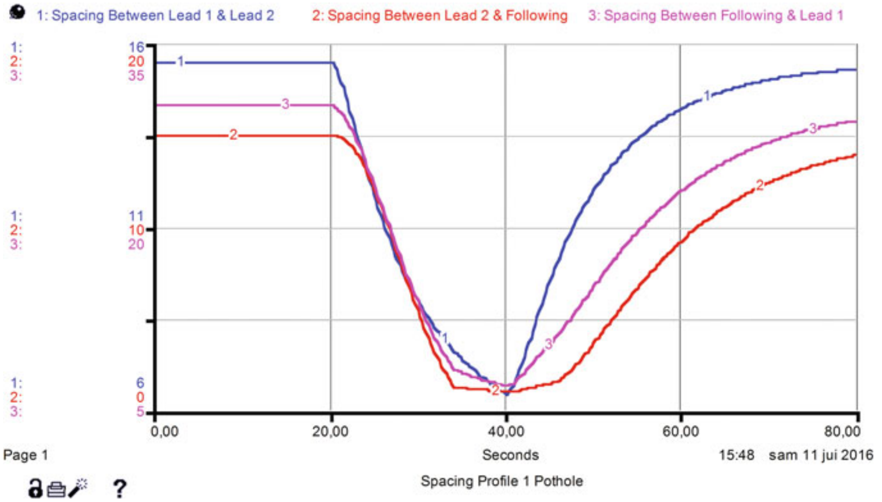


Fig. 11 Spacing profile obtained using proposed model with one pothole on the roadway, as obtained from the SD simulation

Figure 10 illustrates the speed profiles of all three vehicles. The simulation covers a time interval of observation of 80 s (real-world time); the motions of all three vehicles (i.e., speed and spacing) are presented in Figs. 10 and 11. As shown in Fig. 11, the speed reduction of the first lead vehicle results in a corresponding speed reduction of both second lead and follower vehicles.

In Fig. 11, the second lead vehicle's driver perceives the speed reduction of the first lead vehicle; he then reacts by slowing down, and this induces a spacing reduction between them due to the pothole in front of the first lead vehicle. The driver of the following vehicle also reacts in a similar fashion and reduces his or her speed to match that of the second lead vehicle while keeping an appropriate spacing to avoid a collision.

5.4 Stock-Flow Diagram of a Car-Following Model on a Road with Two Potholes

Here we assume that the first lead vehicle travels on a roadway with two potholes with different characteristics: the first pothole has a length of 18.5 m and the second is of length 14 m. The distance between the two potholes is 590 m. The entire road length is 1003 m.

From $t = 0$ s to $t = 20$ s, cars are moving in MRU mode, and the acceleration is thus zero. In this 20 s period, the spacing between the first lead and the second lead vehicle is 16 m. Also, the distance between the second lead and the follower vehicle is 16 m. This means that the spacing between the first lead and the follower vehicle is 32 m. The speed of the three vehicles is approximately 10.5 m/s (38 km/h). At $t = 21$ s, the first lead vehicle sees the pothole and begins to decrease its speed down to a certain value (from 10.5 m/s to 3.82 m/s) in order to enter the first pothole slowly (with length 18.5 m) at $t = 35$ s. In the first pothole, the speed is between 3.8 m/s and 3.62 m/s ($t = 35$ s to $t = 40$ s). After $t = 40$ s, the vehicle leaves the first pothole with MRUA with a speed between 3.96 m/s and 10.49 m/s ($t = 40.25$ s to $t = 75$ s). After $t = 75$ s, the first lead car moves in MRU mode ($t = 75$ s to 80 s), MURD ($t = 80$ s to 97 s), and in the second pothole, ($t = 97$ s to $t = 100$ s) as well as in MRUA ($t = 100$ s to $t = 120$ s) (See Table 6).

The empirical data obtained from the simulation show that the presence of two potholes reduces the first lead vehicle's speed by **66%** within the first pothole and by **57%** in the second pothole.

If we consider the initial speed of 10.5 m/s and MRU motion of the first lead vehicle, the entire distance would be covered in 96 s. In the presence of two potholes separated by 590 m, the same distance is covered in 120 s. This assertion shows that 24 s of travel time is added due to potholes. It represents a 20% loss of time

In Figs. 13 and 14, the speed reduction of the first lead vehicle results in a corresponding speed reduction of the following vehicles. The second lead vehicle driver perceives the speed reduction of the first lead vehicle through seeing both the rear brake lights and also the spacing reduction between them as caused by the first pothole ahead of the first lead vehicle. The driver of the third following vehicle also reacts in a similar fashion and consequently reduces his or her speed to match that of the second lead vehicle and to avoid a possible collision.

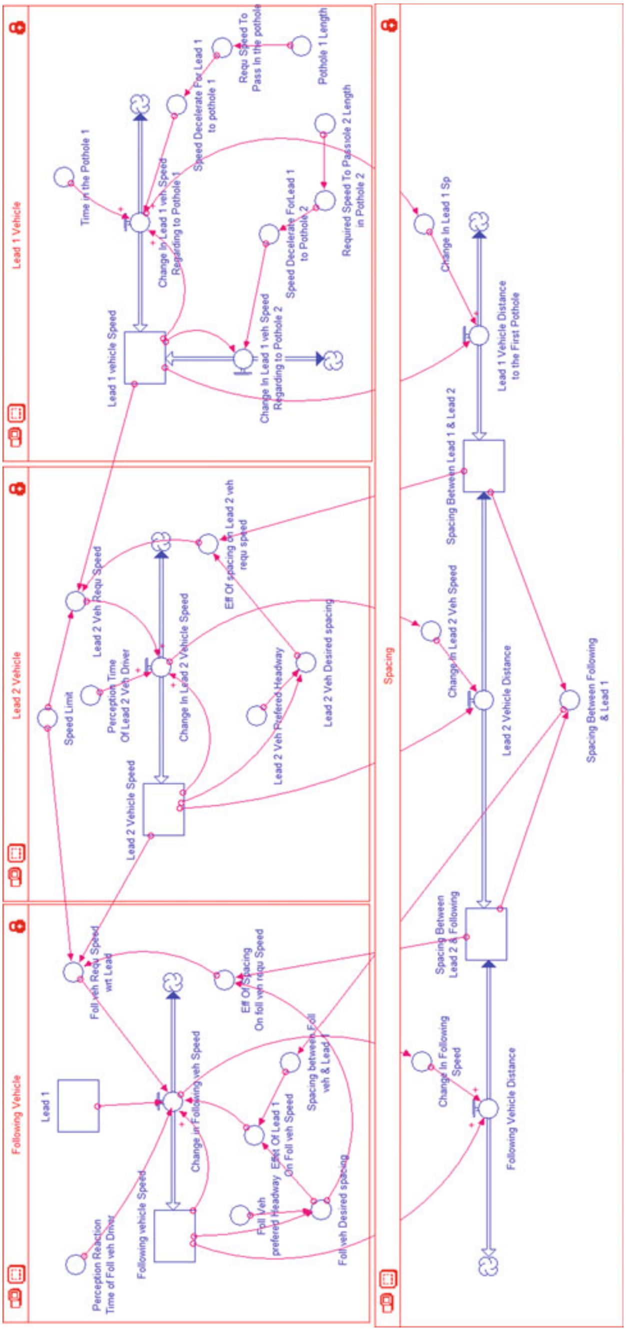

Lead 1 Vehicle
Time in the Pothole 1
Change In Lead 1 veh Speed Regarding to Pothole 1
Speed Decelerate For Lead 1 to pothole 1
Requ Speed To Pass In the pothole
Pothole 1 Length
Required Speed To Pass In Pothole 2
Pothole 2 Length
Speed Decelerate ForLead 1 to Pothole 2
Change In Lead 1 veh Speed Regarding to Pothole 2
Lead 1 vehicle Speed
Lead 2 Vehicle
Speed Limit
Perception Time Of Lead 2 Veh Driver
Lead 2 Veh Requ Speed
Change In Lead 2 Vehicle Speed
Eff Of spacing on Lead 2 veh requ speed
Lead 2 Veh Prefered Headway
Lead 2 Veh Desired spacing
Lead 2 Vehicle Speed
Following Vehicle
Lead 1
Perception Reaction Time of Foll veh Driver
Following vehicle Speed
Change in Following veh Speed
Foll veh Requ Speed wrt Lead
Eff Of Spacing On foll veh requ Speed
Spacing between Foll veh & Lead 1
Effe Of Lead 1 On Foll veh Speed
Foll Veh prefered Headway
Foll veh Desired spacing
Spacing
Change In Lead 1 Sp
Lead 1 Vehicle Distance to the First Pothole
Spacing Between Lead 1 & Lead 2
Change In Lead 2 Veh Speed
Lead 2 Vehicle Distance
Spacing Between Following & Lead 1
Spacing Between Lead 2 & Following
Change In Following Speed
Following Vehicle Distance
+

Fig. 12 Stock-flow diagram of the proposed SD model for a scenario with two potholes on the road

Table 6 Aggregated data collected on road with two potholes; data related to the simulation presented in Figs. 13 and 14

Simulation time	Motion mode	Distance covered by first lead (m)	First lead vehicle speed (m/s)	Percentage of speed variation of first lead (%)
$t = 0\,\text{s}$ to $t = 20\,\text{s}$	MRU	212.625	10.5	0
$t = 20.25\,\text{s}$ to $t = 35\,\text{s}$	MRUD	84.2125	10.15 – 3.82	3.3 – 63.62
$t = 35.25\,\text{s}$ to $t = 40\,\text{s}$ (within first pothole)	MRUD	18.4825	3.81 – 3.62	63.71 – 65.52
$t = 40.25\,\text{s}$ to $t = 75\,\text{s}$	MRUA	334.82	3.96 – 10.49	62.3 – 0.09
$t = 75.25\,\text{s}$ to $t = 80\,\text{s}$	MRU	52.5	10.5	0
$t = 80.25\,\text{s}$ to $t = 97\,\text{s}$	MRUD	104.125	10.2 – 4.68	2.86 – 55.43
$t = 97.25\,\text{s}$ to $t = 100\,\text{s}$ (within second pothole)	MRU	13.9	4.67 – 4.6	55.52 – 56.19
$t = 100.25\,\text{s}$ to $t = 120\,\text{s}$	MRUA	182.43	4.89 – 10.4	53.43 – 0.95

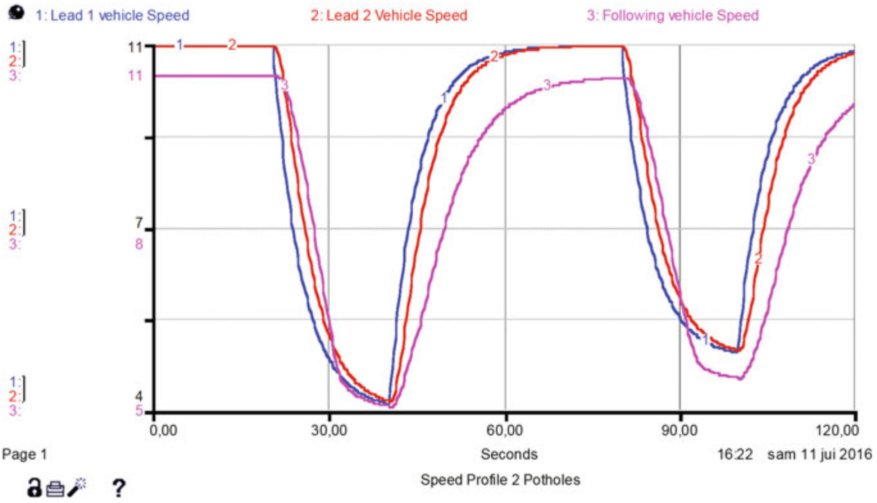


Fig. 13 Speed profiles of the first lead vehicle, of the second lead vehicle, and of the third following vehicle obtained from the SD simulation of Fig. 12

The tendency of the first lead vehicle after exiting the first pothole is to increase its speed, but as soon the second pothole (separated by 590 m from the first one) is discovered, the driver gradually decreases speed again. The situation remains the same for the following vehicles. We remark that through simulation it can be shown that on the same road segment, travel time increases significantly if a road has more potholes.

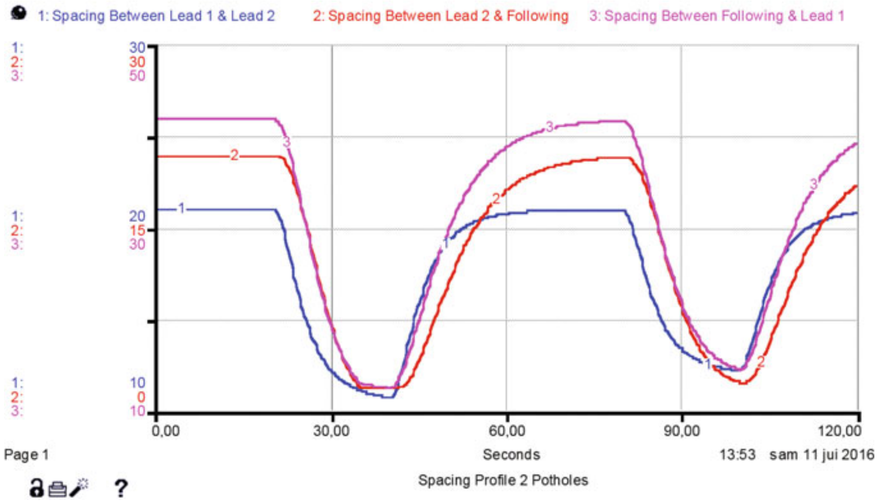


Fig. 14 Spacing profile obtained using the proposed model with two potholes on the roadway, as obtained from the SD simulation of Fig. 12

6 Road Capacity Reduction Due to a Degraded Road with Potholes

According to the United States Highway Capacity Manual (US HCM) [23], there are some basic factors affecting road capacity under basic conditions, which assume good weather, good pavement conditions, and users' familiarity with the roadway and no obstacles to traffic flow.

It is known that road capacity (RC) represents the maximum traffic flow obtainable on a given roadway using all available lanes; it is usually expressed in vehicles per hour or vehicles per day (see Eq. (22)). Capacity is considered in terms of a relation between speed, time, and flow:

$$RC = max[q(k)] = max[k(x,t) \times u(x,t)]. \tag{22}$$

US HCM methods are applicable in both developing and underdeveloped countries with adapted considerations. In Kinshasa, traffic and driving conditions exist that differ from those in developed regions. We experience a deterioration of roads mainly expressed through potholes. This situation causes a slowdown (speed deceleration) of vehicles in order to pass through potholes; this situation is worse during the rainy seasons. The pothole characteristics (i.e., depth, composition (rocky, muddy, full of water, etc.) have an influence on car speed profile, steering, and on time required to pass the pothole.

Figures 10, 11, 13 and 14 have sufficiently illustrated that the presence of pothole(s) along the road has an impact on speed profiles and travel times. Figure 10 (one pothole) and Fig. 13 (two potholes) convey a message explaining how far the pothole(s) reduce road capacity. The **lowest speed value observed in these two figures constitutes an effective bottleneck for road capacity**. The road cross section where the lowest speed value occurs constitutes the road portion with the lowest effective capacity due to the pothole(s). This lowest speed value in the speed profiles of Figs. 10 and 13 is the maximum speed within the pothole cross section. As a car emerges from the pothole, one also observes the birth of a shockwave leading eventually to congestion if the traffic density on the road segments to the rear of the pothole is near or greater than half of the maximum traffic density k_j.

For the two pothole scenarios illustrated in Fig. 13 with the same assumptions, the congestion is not more severe than in the one pothole case, because the minimum speeds with both potholes are nearly the same. Since a pothole is an effective capacity bottleneck, two consecutive bottlenecks at two or more consecutive road cross sections have the same effect as a single pothole. This statement is crucial for road maintenance: it means that a responsible administration must take even a singe pothole very seriously and not wait until the road is full of dozens of potholes. The capacity reduction begins immediately when the first pothole appears on the road.

A way to determine road capacity (RC) is explained in the highway capacity manual and in the road traffic literature:

$$RoadCapacity = ([max.speed] \times [maximumdensity]/4).$$

If we consider a one-lane standard urban road with a maximum speed of 50 km/h, in applying the road capacity relationship with a maximum density of $k_j = 167$ vehicles/km, one obtains a road capacity pf

$$RC(for\ 1 - laneroadwithoutpothole) = (50 \times 166/4)veh/h = 2087\mathit{vehicles/hour}.$$

Further, according to Figs. 10 and 13 the maximum speed within the pothole is approximately 4 m/s, which corresponds to 14 km/h. In general, we can state that depending on the pothole constitution (material, full of water or dry, etc.) and dimensions (size, width, and depth), the maximum speed within the pothole ranges between 10 km/h and 20 km/h.

Thus, a maximum speed of 10 km/h appears to be a good indicative maximum speed within potholes. Therefore, the road capacity in the cross section with a pothole is given according to the road capacity relation by

$$RC(for\ 1 - laneroadwithpothole(s)) = (10 \times 167/4)veh/h = \mathbf{417vehicles/hour}.$$

To assess the effective capacity reduction due to one or more potholes on a one-lane road, one just compares the two last values. The road capacity decreases from an initial value of 2087 vehicles/hour to a value of 417 vehicles/hour. This corresponds to reduction by a factor of 5 in the effective road capacity. This should raise a huge alarm in all those who are in charge of road maintenance and traffic management in an urban environment, especially for Kinshasa. A single pothole has an extremely huge negative impact on road capacity. A road capacity reduction by a factor of 5 has the negative consequence that peak traffic (in the busiest hour) that usually traversed a road in 1 hour will now take a minimum of 5 hours to traverse the same road segment. Therefore, for most road users, the usual travel time (experienced when the road pavement is safe) on a specific road segment having one or more potholes will be multiplied by at least 5.

Important remark: One should notice that the above-mentioned "capacity reduction" estimation due to a road pothole is **in the context of a single-lane road with no possibility to change lanes to avoid the pothole**. In the case (not considered in this document) of a road with a single lane in each direction, and with the possibility of using the lane of the opposite direction to avoid the pothole in some cases or for overtaking, if the current traffic allows such a maneuver, then the capacity reduction factor will be a bit different. This case and the cases of multiple lanes in each road direction will be addressed and simulated in a future work.

7 Conclusion and Outlook

Severe road traffic congestion is an obstacle to mobility in general. Here, it is stressed that the traffic congestion problem specifically linked to road potholes must be carefully addressed.

This work has essentially assessed how far road potholes and traffic congestion are interlinked and together constitute a real-life complex societal problem, especially in the city of Kinshasa, DR Congo. This societal problem can be effectively defined and handled through the COMPRAM methodology because of its multidisciplinary, multiactor, and interdisciplinary flavors.

This research indicates that SD simulation (in this case using the software tool Stella) can reliably simulate the negative impact of road potholes on microscopic traffic variables of a car-following model. A simulation has been proposed regarding the presence of one or more potholes on the roadway in which all variables are implemented to reflect equilibrium in all stocks of the SD system model, which implements a car-following model on a degraded road. After the SD model was developed, two illustrative scenarios were successfully simulated: (a) a scenario with a single pothole on a 1-lane road, and (b) a scenario with two potholes separated by a given distance of 590 meters on a 1-lane road.

In this research we first presented a brief consideration of COMPRAM as applied to "road potholes" and subsequent "traffic congestion." A series of societal considerations is required for a sustainable problem handling. The hypothesis that more policy intervention is needed via improvement in road legislation, road maintenance, and road monitoring during the construction period and afterward will be validated and consolidated by a more detailed and thorough implementation of all seven layers of the COMPRAM model as shown in Fig. 3.

As already mentioned above, further research on this topic will concentrate on the comprehensive problem of the coupled societal problems “road potholes” and “traffic congestion” in the special case of the city of Kinshasa. Thereby all seven layers of the COMPRAM methodological platform will be thoroughly implemented with contextual adaptions to the case study under consideration.

Furthermore, the complex societal problem setting discussed above also links to other societal problems. One of these is related to the fact that the evolution of potholes over time (for example increasing number, increasing special coverage of potholes on ever more roads of the global road network, etc.) has a significant negative impact on the following elements: the reliability of the vehicles of the city will decrease (more defects and higher defect frequency due to potholes), there will be an acceleration of vehicles depreciation, pothole-induced accidents will impact global road safety indicators, driver behavior will be negatively influenced, there will be global macro- or microeconomic consequences (for the target city of Kinshasa) of both potholes and traffic, and there will be a significant impact on short-term predictions of road traffic. These factors are of relevance for a crucially needed optimized traffic management, which is also viewed as an immediate technical way to alleviate traffic congestion.

References

1. Naja, R.: Advanced Motion Control and Sensing for Intelligent Vehicles. Springer, Berlin (2013)
2. Tanaka, M., Uno, N.: Analysis of correlation between car-following platoon size and macroscopic traffic stream models. Proc. East. Asia Soc. Transp. Stud. **9**, 2–7 (2013)
3. Li, L., Wang, F.Y.: Wireless Vehicular Networks for Car Collision Avoidance. Springer, Berlin (2007)
4. De Fabritiis, C., Ragona, R., Valenti, G.: Traffic estimation and prediction based on real time floating car data. In: 11th International IEEE, Intelligent Transportation Systems, pp. 197–203 (2008)
5. Singh, K.S.: Review of urban transportation in India. J. Public Transp. **8**(1), 79–97 (2005)
6. Sen, R., Sevani, V., Sharma, P., Koradia, Z., Raman, B.: Challenges in communication assisted road transportation systems for developing regions. In: ACM Workshop on Networked Systems for Developing Regions (2009)
7. DeTombe, D.: Handling Societal Complexity: A Study of the Theory of the Methodology of Societal Complexity and the COMPRAM Methodology. Springer, Berlin (2015)
8. DeTombe, D.: Handling Complex Societal Problems (Applied on the Aids/HIV Problem). Edward Elgar Publishers, Ann Arbor (2003)
9. DeTombe, D.: Climate change: a complex societal process; analysing a problem according to the COMPRAM methodology. J. Transform. Soc. Change **5**(3), 235–266 (2008)
10. Sterman, J.D.: Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, New York (2000)
11. Lighthill, M.J., Whitham, G.B.: On kinematic waves II: a theory of traffic flow on long crowded roads. Proc. Royal Soc. London **229**, 317–345 (1955)

12. Gazis, D.C., Herman, R., Rothery, R.W.: Car-following theory of steady state flow. Oper. Res. **7**(4), 499–505 (1959)
13. Richards, P.I.: Shock waves on the highway. Oper. Res. **4**, 42–51 (1956)
14. Pipes, L.A.: An operational analysis of traffic dynamics. J. Appl. Phys. **24**, 296–315 (1953)
15. Ahmed, K.I.: Modelling drivers acceleration and lane changing behavior. Doctor of science in transportation systems and decision sciences, Massachusetts Institute of Technology (1999)
16. Forbes, T.W., Zagorsk, H.J., Holshouser, E.L., Deterline, W.A.: Measurement of driver reactions to tunnel conditions. Proc. Highw. Res. Board **37**, 345–357 (1958)
17. Drake, J.S., Schofer, J.L., May, A.D. Jr.: A statistical analysis of speed density hypotheses. In: Third International Symposium on the Theory of Traffic Flow Proceedings, Elsevier North Holland, New York (1967)
18. Greenberg, H.: An analysis of traffic flow. Oper. Res. **7**, 78–85 (1959)
19. Underwood, R.T.: Speed, Volume and Density Relationships. Yale Bureau of Highway Traffic, New Haven (1961)
20. DeTombe, D.: Complex societal problems in operational research. Eur. J. Oper. Res. **140**, 232–240 (2002)
21. DeTombe, D.: Defining Complex Interdisciplinary Societal Problems, A Theoretical Study for Constructing a Co-operative Problem Analyzing Method: The Method COMPRAM. Thesis publishers Amsterdam, Holland (1994)
22. The Ontario Hot Mix Producers Association, The ABCs of potholes. http://www.ohmpa.org (2009)
23. Highway capacity manual, transportation research board (2010)

Part V
Computational Intelligence

Design of a Chaotic Pulse-Position Modulation Circuit

Junying Niu, Zhong Li, Yuhong Song and Wolfgang A. Halang

Abstract An analog chaotic modulation circuit based on Chua's circuit is proposed to generate a chaotic pulse-position signal. The circuit is designed with standard electronic components, and the parameters of the generated signals, including the pulse period, the modulation range of the pulse-position, even the probability distribution of the pulse-position, can be adjusted flexibly.

1 Introduction

The highly unpredictable and random-like features of chaotic signals are very useful and desired for engineering applications [2, 5, 6, 12]. One of the successful applications is a chaotic modulation technique, which has been utilized in the electronic communication field [2, 7, 10]. There, chaotic pulse-position modulation(CPPM), which is to chaotically dither the high-level duration, namely pulse position, of a periodic pulse, has attracted extensive interest, and it has been applied in secure communications [8, 11], EMI reduction of switching mode power supply [9], and so on [5].

Until now, CPPM has been realized digitally [4], while analog CPPM has its unique advantages in low cost and rapid reaction. A digital CPPM signal is generated

J. Niu (✉) · Y. Song
Department of Electronic and Information Engineering, Shunde Polytechnic,
Foshan, China
e-mail: niujunying@foxmail.com

Y. Song
e-mail: syhscut@163.com

Z. Li · W.A. Halang
Faculty of Mathematics and Computer Science, FernUniversität in Hagen,
Hagen, Germany
e-mail: Zhong.Li@fernuni-hagen.de

W.A. Halang
e-mail: Wolfgang.Halang@fernuni-hagen.de

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_10

by a digital device, such as FPGA, digital signal processor (DSP), and high-speed processor. The advantage of the digital CPPM circuit is that the pulse period and the pulse position can be easily adjusted without changing any hardware. However, its disadvantages involves the high cost and the unique character of the signal discretization. In contrast, the analog CPPM circuit is realized by separate electronic components. Firstly, the cost of the analog chaotic carrier is much lower. Secondly, and more importantly, there is no discretization problem in the analog system, so that it satisfies the requirements of high precision and fast reaction.

To facilitate the application of chaos control, an analog CPPM circuit composed of standard electronic components will be designed and simulated. Not only chaotic pulse-position modulation but also chaotic impulse-position modulation are implemented. Meanwhile, the timing parameters, including the pulse period and the modulation range of pulse position, can be set conveniently. More importantly, the probability distribution of the pulse-position can be adjusted by regulating a resistor.

2 The Chaotic Pulse-Position Modulation Circuit

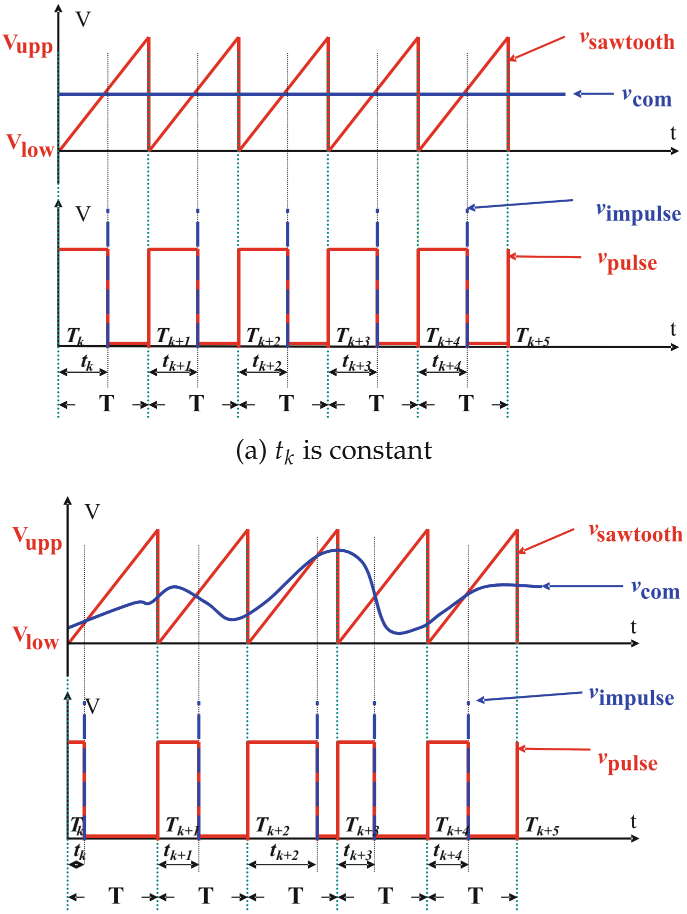
2.1 Working Principle

The working principle of the circuit is shown in Fig. 1. To obtain periodic pulse and impulse (see Fig. 1a), a sawtooth ($v_{\text{sawtooth}} \in [\text{V}_{\text{low}}, \text{V}_{\text{upp}}]$) with period T is used to be compared with a certain voltage (v_{com}). At the beginning of the period of v_{sawtooth}, v_{pulse} is set to a high level. Once v_{sawtooth} is larger than v_{com}, v_{pulse} is locked to a low level until the end of the period. The differential coefficient of v_{pulse}'s falling edge is an impulse (v_{impulse}).

Thus, the pulse position t_k is constant, since v_{com} is constant. Conversely, t_k is modulated chaotically, since v_{com} is chaotic, as shown in Fig. 1b. Hence, the chaotic modulations of pulse-position and impulse-position are implemented.

2.2 Design of the CPPM Circuit

A chaotic pulse-position modulation circuit was designed as in Fig. 2. The circuit is composed of four parts, among which the chaos circuit is used to generate the chaotic voltage v_{com}. In the oscillator circuit, the timing capacitor C_1 is charged and discharged between V_{low} and V_{upp} periodically, and its voltage v_{sawtooth} takes on the waveform of a periodic sawtooth. Then, in the trigger circuit, v_{pulse} is obtained by comparing v_{sawtooth} and v_{com}. Finally, via the differential circuit, v_{pulse} is differentiated to v_{impulse}. A schematic of the oscillator circuit, the trigger circuit, and the differential circuit is given in Fig. 3.



(a) t_k is constant

(b) t_k is chaotic changed

Fig. 1 Waveforms of v_{sawtooth}, v_{pulse}, and v_{impulse}

The charging and discharging circuit of C_1 is a first-order circuit [1], so that the period of the sawtooth, which is also that of the pulse sequence and the impulse sequence, is

$$\mathrm{T} = \frac{1}{\mathrm{F}} = R_6 C_1 ln \frac{\mathrm{V}_{\text{low}} - \mathrm{V}_{\text{REF}}}{\mathrm{V}_{\text{upp}} - \mathrm{V}_{\text{REF}}}, \tag{1}$$

where, as shown in Fig. 3, R_1–R_6 are the resistor parameters, V_{REF} is the fixed reference voltage, and $\mathrm{V}_{\text{low}} = \mathrm{V}_{\text{REF}} \frac{R_4}{R_3 + R_4}$, $\mathrm{V}_{\text{upp}} = \mathrm{V}_{\text{REF}} \frac{R_2}{R_1 + R_2}$. Since V_{low} and V_{upp} can be changed by adjusting R_1–R_4, the period T can be set accordingly.

During a period, the position of the v_{pulse}'s falling edge, t_k, can be calculated as

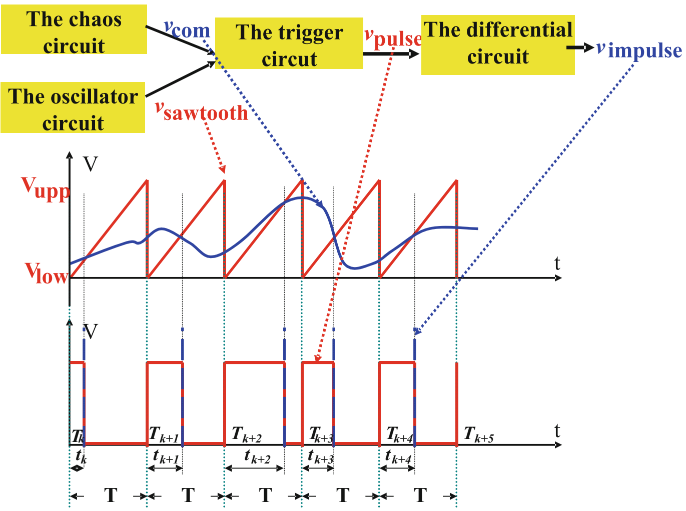


Fig. 2 Diagram of the chaotic pulse-position and chaotic impulse-position modulation circuit

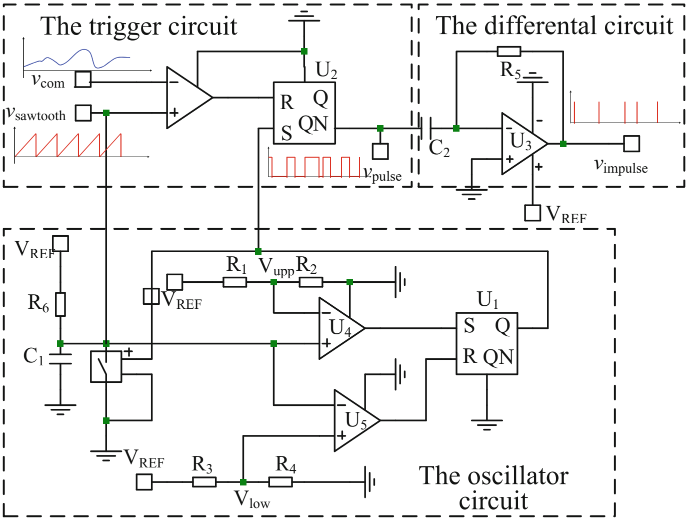


Fig. 3 A schematic of the oscillator circuit, the trigger circuit, and the differential circuit

$$t_k = R_6 C_1 \ln \frac{\mathrm{V_{low}} - \mathrm{V_{REF}}}{v_{\mathrm{com}} - \mathrm{V_{REF}}}. \tag{2}$$

To implement CPPM, v_{com} should be chaotic.

2.3 Chaos Circuit

In this design, Chua's circuit [3] is used to generate the chaotic voltage. Figure 4 illustrates Chua's circuit, where $\mathrm{N_R}$ is Chua's diode. As shown in Fig. 5, $i_{\mathrm{N_R}}$ and v_1 satisfy the relationship

$$i_{\mathrm{N_R}} = f(v_1) = G_b v_1 + \frac{1}{2}(G_a - G_b)(|v_1 + E| - |v_1 - E|). \tag{3}$$

Hence, Chua's oscillator is expressed as

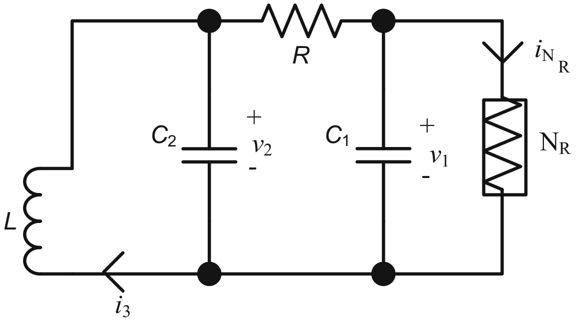
$$\begin{cases} \dfrac{dv_1}{dt} = \dfrac{1}{C_1}[(v_2 - v_1)G - f(v_1)], \\ \dfrac{dv_2}{dt} = \dfrac{1}{C_2}[(v_1 - v_2)G + i_3], \\ \dfrac{di_3}{dt} = -\dfrac{1}{L}v_2, \end{cases} \tag{4}$$

where $G = \frac{1}{R}$.

As R varies, v_1 and v_2 exhibit different waveforms, which may be constant, periodic, and chaotic.

The chaos circuit is composed of Chua's circuit and an amplitude-limiting circuit. Chua's circuit is employed to produce chaotic voltage $v_2(v_2 \in [v_{2_{\min}}, v_{2_{\max}}])$, and then through the amplitude-limiting circuit, v_2 is linearly transformed to v_{com} within a specified range. In reality, Chua's diode, N_R, is designed as the circuit shown in the shadow in Fig. 6.

Fig. 4 Chua's oscillator circuit



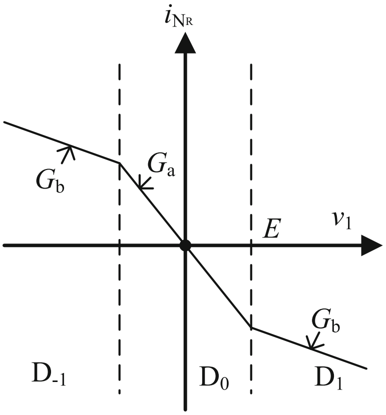


Fig. 5 Characteristic plot of Chua's diode

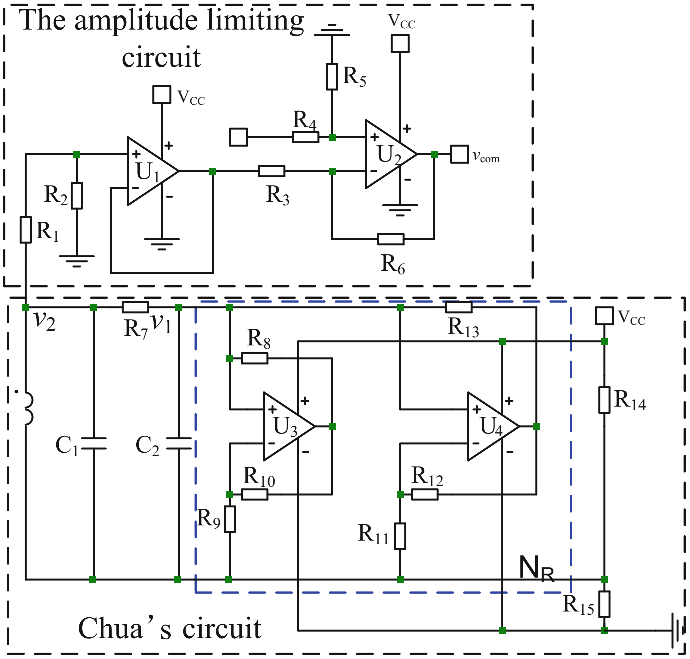


Fig. 6 Schematic of the chaos circuit

As shown in Fig. 6, R_7–R_{12} are the resistor parameters, Vcc is the power voltage; then v_{com} can be expressed as

$$V_{\text{com}} = \frac{1}{R_9}\left[\frac{\text{Vcc}R_{11}(R_9 + R_{10})}{R_{11} + R_{12}} - \frac{v_2 R_8 R_{10}}{R_7 + R_8}\right]. \tag{5}$$

3 Parameter Calculation

3.1 The Timing Parameter

Substituting (5) into (2) leads to

$$t_k = R_6 C_1 \ln \frac{\text{V}_{\text{low}} - \text{V}_{\text{REF}}}{\frac{\text{Vcc}R_{11}(R_9+R_{10})}{R_9(R_{11}+R_{12})} - \frac{v_2 R_8 R_{10}}{R_9(R_7+R_8)} - \text{V}_{\text{REF}}}, \tag{6}$$

where v_2 is a chaotic voltage varying between $v_{2_{\min}}$ and $v_{2_{\max}}$, t_k is modulated chaotically and varied between $t_{k_{\min}}$ and $t_{k_{\max}}$. Therefore, in order for t_k to vary in a specified range, R_7–R_{12} should satisfy the following equations deduced from (6):

$$\begin{aligned} x_{\min} &= e^{-\frac{t_{k_{\min}}}{R_6 C_1}}(\text{V}_{\text{low}} - \text{V}_{\text{REF}}) + \text{V}_{\text{REF}} \\ &= \frac{\text{Vcc}R_{11}(R_9 + R_{10})}{R_9(R_{11} + R_{12})} - \frac{v_{2_{\min}} R_8 R_{10}}{R_9(R_7 + R_8)}, \end{aligned} \tag{7}$$

$$\begin{aligned} x_{\max} &= e^{-\frac{t_{k_{\max}}}{R_6 C_1}}(\text{V}_{\text{low}} - \text{V}_{\text{REF}}) + \text{V}_{\text{REF}} \\ &= \frac{\text{Vcc}R_{11}(R_9 + R_{10})}{R_9(R_{11} + R_{12})} - \frac{v_{2_{\max}} R_8 R_{10}}{R_9(R_7 + R_8)}, \end{aligned} \tag{8}$$

where $x_{\min}$ and $x_{\max}$ are defined to simplify calculations. Then

$$\frac{\text{Vcc}R_{11}(R_9 + R_{10})}{R_9(R_{11} + R_{12})} = \frac{v_{2_{\max}} x_{\min} - v_{2_{\min}} x_{\max}}{v_{2_{\max}} - v_{2_{\min}}}, \tag{9}$$

$$\frac{R_8 R_{10}}{R_9(R_7 + R_8)} = \frac{x_{\max} - x_{\min}}{v_{2_{\max}} - v_{2_{\min}}}. \tag{10}$$

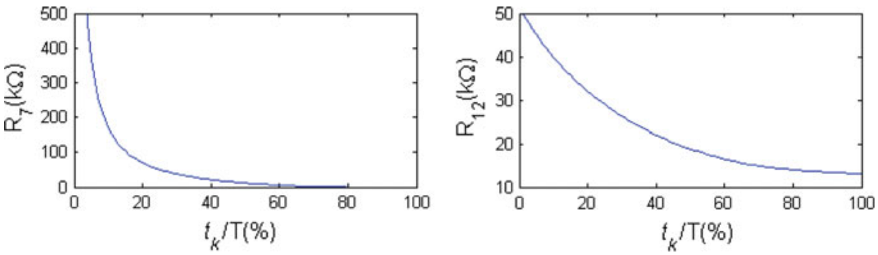
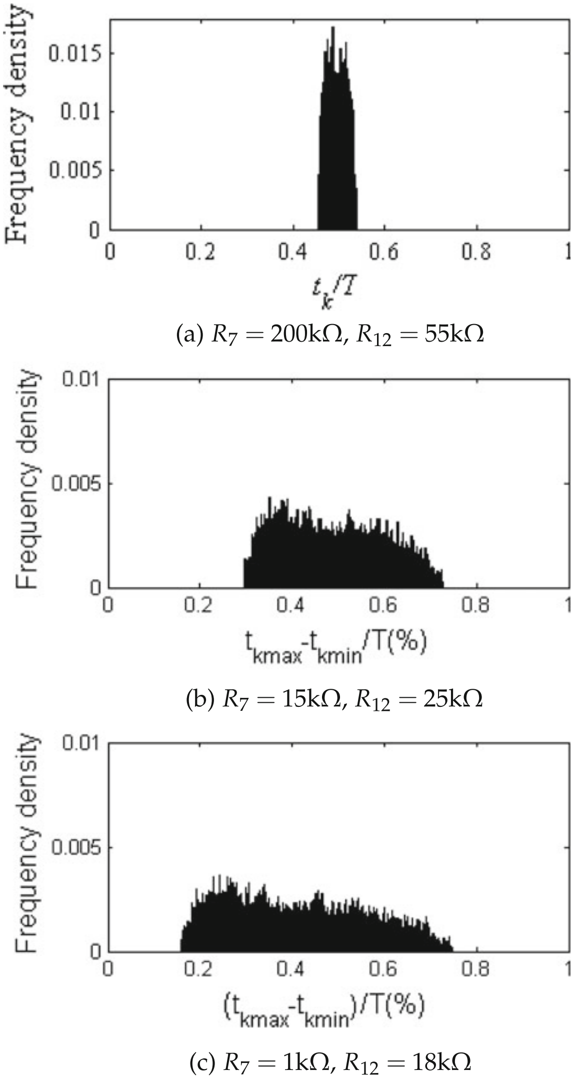


Fig. 7 R_7 and R_{12} versus $\frac{t_{k\max} - t_{k\min}}{T}$

Fig. 8 The histogram of $\frac{t_{k\max} - t_{k\min}}{T}$



(a) $R_7 = 200\text{k}\Omega$, $R_{12} = 55\text{k}\Omega$

(b) $R_7 = 15\text{k}\Omega$, $R_{12} = 25\text{k}\Omega$

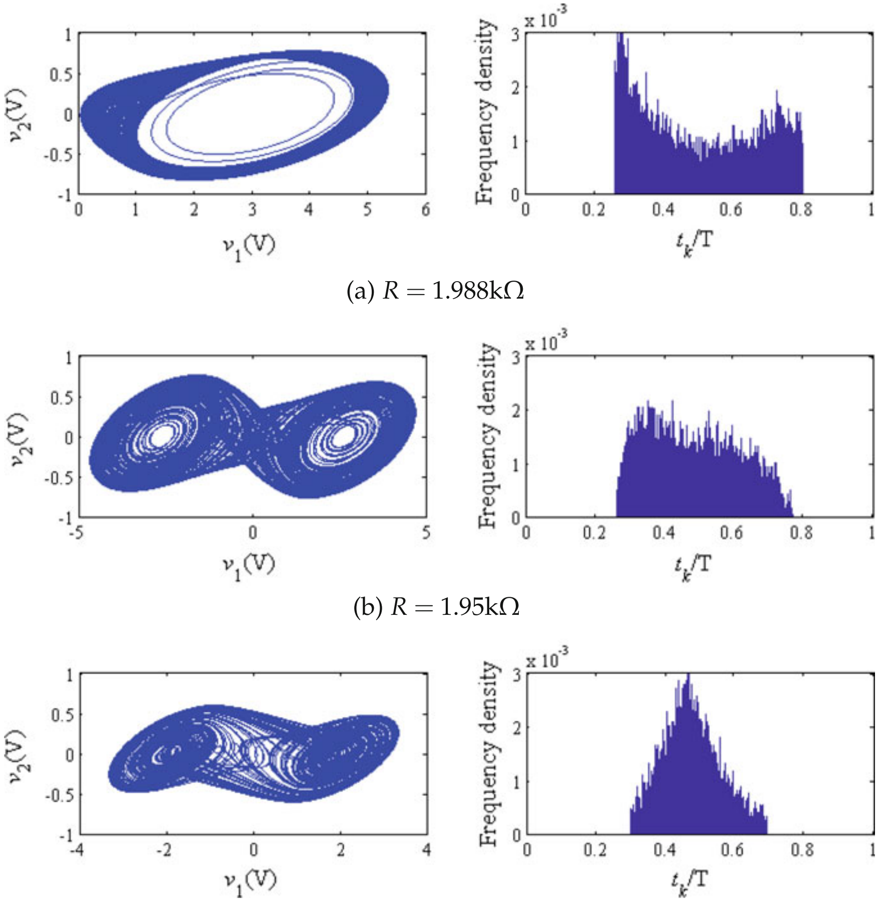
(c) $R_7 = 1\text{k}\Omega$, $R_{12} = 18\text{k}\Omega$

We remark that R_7 and R_{12} can be used to set $t_{k_{\min}}$ and $t_{k_{\max}}$, since the other parameters are fixed. Furthermore, to set $t_k \in [t_{\mathrm{k_{min}}}, t_{\mathrm{k_{max}}}]$, R_7 and R_{12} can be calculated as

$$R_7 = \frac{R_8 R_{10}(v_{2_{\max}} - v_{2_{\min}})}{R_9(x_{\max} - x_{\min})} - R_8, \tag{11}$$

$$R_{12} = \frac{\mathrm{Vcc} R_{11}(R_9 + R_{10})(v_{2_{\max}} - v_{2_{\min}})}{R_9 x_{\max} - x_{\min}} - R_{11}. \tag{12}$$

For example, R_7 versus $\frac{t_{\mathrm{kmax}} - t_{\mathrm{k_{min}}}}{\mathrm{T}}$, and R_{12} vs $\frac{t_k}{\mathrm{T}}$ is given in Fig. 7, as Vcc = 12V, $R_8 = R_9 = R_{10} = 30k\Omega$, $R_{11} = 10k\Omega$, $v_{2_{\max}} = 7\mathrm{V}$, $v_{2_{\min}} = 5\mathrm{V}$, $t_{k_{\max}} + t_{k_{\min}} = \mathrm{T}$.



(a) $R = 1.988\mathrm{k}\Omega$

(b) $R = 1.95\mathrm{k}\Omega$

(c) $R = 1.8\mathrm{k}\Omega$

Fig. 9 *Left* phase portraits of Chua's circuit, *right* the corresponding histogram of $\frac{t_{\mathrm{kmax}} - t_{\mathrm{k_{min}}}}{\mathrm{T}}$

According to Fig. 7, the statistical property of $\frac{t_k}{\mathrm{T}}$ with different values of R_7 and R_{12} is given by Fig. 8.

3.2 Probability Distribution of Chaotic Pulse Position

In addition to the variation range, the probability distribution of the chaotic pulse position can also be set. Figure 9 describes the statistical property of $\frac{t_k}{T}$ with different values of R, which determine the state of Chua's circuit. We remark that the distribution of the values is transferred from the left side to the mean value with a decrease in R.

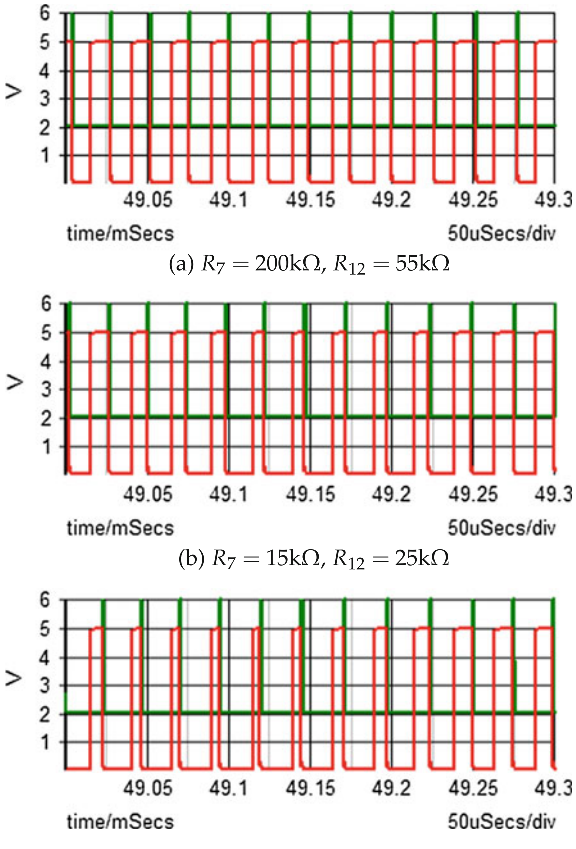


Fig. 10 Waveforms of a chaotically modulated pulse (*red*) and impulse (*green*)

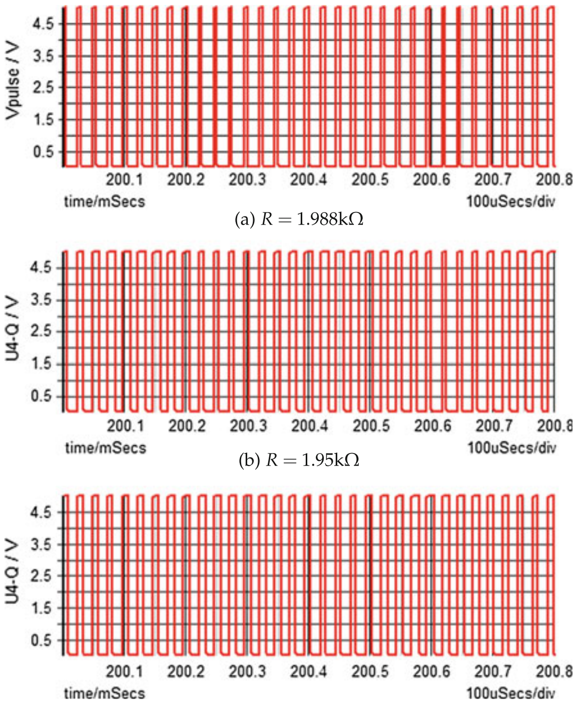
(a) $R_7 = 200\text{k}\Omega$, $R_{12} = 55\text{k}\Omega$

(b) $R_7 = 15\text{k}\Omega$, $R_{12} = 25\text{k}\Omega$

(c) $R_7 = 1\text{k}\Omega$, $R_{12} = 18\text{k}\Omega$

4 Simulations

The circuit shown in Fig. 3 is simulated. The waveforms of the chaotically modulated pulse and impulse with different values of R_7 and R_{12}, which are shown in Fig. 10, indicate that the modulation range of the pulse position becomes larger with a decrease in R_7 and R_{12}. Meanwhile, the waveforms of the chaotically modulated pulse with different values of R are given in Fig. 10, which shows that the probability distribution of the chaotic pulse position can be adjusted by R. Coincidental with Fig. 9, since $R = 1.988\,\mathrm{k}\Omega$, the values of the pulse position are concentrated in the area where they are relatively small. The values become uniform for $R = 1.988\,\mathrm{k}\Omega$, and are focused on the mean value for $R = 1.8\,\mathrm{k}\Omega$. Obviously, the pulse-position and impulse-position are modulated chaotically, and the simulations (Fig. 11) are basically consistent with the calculations.



(a) $R = 1.988\mathrm{k}\Omega$

(b) $R = 1.95\mathrm{k}\Omega$

(c) $R = 1.8\mathrm{k}\Omega$

Fig. 11 Waveforms of a chaotically modulated pulse

5 Conclusion

To promote the application of chaotic modulation, an analog CPPM circuit, which can be applied in a switching converter, DC motor control, and secure communication, is designed and simulated. The circuit, composed of standard components, is to generate a chaotic pulse position and chaotic impulse position signals. Meanwhile, the period of the pulse, the pulse-position modulation range, and the probability distribution of the pulse position all can be adjusted conveniently. In conclusion, a CPPM circuit with features of low cost, simplicity, and flexibility can be utilized widely in the electronic communications field.

References

1. Alexander, C., Alexander, C.K., Sadiku, M.: Fundamentals of Electric Circuits. Urban Media Comics (2006)
2. Azzaz, M.S., Tanougast, C., Sadoudi, S., Bouridane, A.: Synchronized hybrid chaotic generators: application to real-time wireless speech encryption. Commun. Nonlinear Sci. Numer. Simul. **18**(8), 2035–2047 (2013)
3. Chua, L.O., Kocarev, L., Eckert, K., Itoh, M.: Experimental chaos synchronization in Chua's circuit. Int. J. Bifurc. Chaos **2**(03), 705–708 (1992)
4. Dat, N.T., Hoang, T.M.: Implementation of chaotic pulse-position modulation on FPGA. In: Proceedings of the Joint INDS'11 & ISTET'11, pp. 1–4. IEEE (2011)
5. Fortuna, L., Frasca, M., Rizzo, A.: Chaotic pulse position modulation to improve the efficiency of sonar sensors. IEEE Trans. Instrum. Meas. **52**(6), 1809–1814 (2003)
6. Hussain, I., Shah, T., Gondal, M.A.: Application of S-box and chaotic map for image encryption. Math. Comput. Modell. **57**(9), 2576–2579 (2013)
7. Li, H., Li, Z., Zhang, B., Wang, F., Tan, N., Halang, W.: Design of analogue chaotic PWM for EMI suppression. IEEE Trans. Electromagn. Compat. **52**(4), 1001–1007 (2010)
8. Rulkov, N.F., Sushchik, M.M., Tsimring, L.S., Volkovskii, A.R.: Digital communication using chaotic-pulse-position modulation. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **48**(12), 1436–1444 (2001)
9. Tse, K.K., Ng, W.M., Chung, H.S.H., Hui, S.Y.R.: Evaluation of a chaotic switching scheme for power converters. In: Power Electronics Specialists Conference, 2000. PESC 00. 2000 IEEE 31st Annual, vol. 1, pp. 412–417. IEEE (2000)
10. Tse, K.K., Ng, R.W.M., Chung, H.S.H., Ron Hui, S.Y.: An evaluation of the spectral characteristics of switching converters with chaotic carrier-frequency modulation. IEEE Trans. Ind. Electron. **50**(1), 171–182 (2003)
11. Yang, H., Jiang, G.-P.: Delay-variable synchronized chaotic pulse position modulation for ultra-wide bandwidth communication. In: 2006 International Conference on Communications, Circuits and Systems (2006)
12. Zhang, Y.-Q., Wang, X.-Y.: Analysis and improvement of a chaos-based symmetric image encryption scheme using a bit-level permutation. Nonlinear Dyn. **77**(3), 687–698 (2014)

Chaos-Based Digital Communication Systems with Low Data-Rate Wireless Applications

Nguyen Xuan Quyen and Kyandoghere Kyamakya

Abstract This chapter presents a study on the modeling and performance evaluation of chaos-based coherent and incoherent systems, i.e., chaotic direct-sequence code-division multiple-access (CDS-CDMA) and differential chaos-shift keying (DCSK), for low-data-rate applications in wireless communications. This study is motivated by the design of a secure physical layer for wireless-based applications with low data rate and in small transmission areas. A wireless channel affected by noise, fading, multipath, and delay-spread for low-data-rate transmission of chaotically spreading signals is described and mathematically modeled. Discrete-time models for the transmitter and receiver of CDS-CDMA and DCSK systems under the impact of the wireless channel are developed. Bit error rate (BER) performance of the systems is estimated by means of both theoretical derivation and discrete integration. Simulated performances are shown and compared with the corresponding estimated ones, where the effects of the ratio E_b/N_0, spreading factor, number of users, sample rate, and the number of transmission paths on the BER are fully evaluated. The obtained results showed that the low-rate chaos-based systems can exploit the multipath nature of wireless channels in order to improve their BER performances. This feature indicates that chaos-based communication systems are a promising and robust solution for enhancing physical layer security in low-rate wireless personal area networks (LR-WPANs).

N. Xuan Quyen (✉)
School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam
e-mail: quyen.nguyenxuan@hust.edu.vn

K. Kyamakya
Institute for Smart System Technologies, Alpen-Adria University of Klagenfurt, Klagenfurt, Austria
e-mail: kyandoghere.kyamakya@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_11

1 Introduction

Chaos-based digital communication systems have received strong interest from researchers worldwide over the past decades [1, 2]. This is mainly due to the observation that chaotic signals not only can be simply generated [3] but also have some beneficial characteristics, i.e., aperiodic random behavior for increasing physical-layer security [4], good correlation properties for spread-spectrum and multiple-access performances [5, 6], immunity of the system to multipath degradation and self-interference [7, 8]. Among various digital communication systems using chaos, chaotic direct-sequence code-division multiple access (CDS-CDMA) [6, 8] and differential chaos shift keying (DCSK) [9, 10] have been the most widely studied. In the CDS-CDMA systems, the chaotic signal is used as a spreading sequence to spread the information-bearing signal, and the sequence synchronization is carried out on the receiving side for coherent demodulation. On the other hand, DCSK systems with incoherent receiver do not require sequence synchronization or channel estimation, but need only symbol or bit rate. Because of its simple structure, the DSCK system is one of the most promising chaos-based communication schemes for hardware implementation [11, 12].

Studies on CDS-CDMA systems can be divided into two main groups: (1) those that address the design [13], optimization [14], and synchronization of communication systems for chaotic spreading sequences [15–18], and (2) those that provide a theoretical and numerical analysis of the BER performance for multiple-access operation under different transmission channels [19–30]. Most of these previous works used correlator-type receivers to study the BER performance over noisy channels [19–25], because of their simple architecture. In contrast, the performance over multipath fading channels has been investigated only by means of more complicated rake receivers, which can combine multipath components to enhance the signal-to-noise ratio [26–29]. The error probability of CDS-CDMA systems is theoretically derived in the context of wide-band channels whose the fading coefficients and multipath delays vary according to random distributions [30]. The combination of DS-CDMA and chaotic pulse time modulation [31] was recently proposed in [32, 33], where the bit duration of input data is varied chaotically and then spread in the frequency domain by directly multiplying with the binary spreading sequences, i.e., pseudo-random noise (PN) or chaos-NRZ sequences. With respect to the DCSK systems, the performance of the conventional scheme over an additive white Gaussian noise (AWGN) channel and multipath fading channel is investigated in [34–36] and then extended to a multipath fading channel with delay spread in [37–39]. The work in [40] presents a study on ultra-wideband (UWB) direct chaotic communication technology using the DCSK scheme for LR-WPANs applications. However, the system performance is evaluated only by numerical simulations. Besides the conventional scheme, multiple extended DCSK schemes have been proposed, such as frequency-modulated DCSK [41], permutation-based DCSK [42], reference-modulated DCSK [43], high-data-rate DCSK [44], high-efficiency DCSK [45], multi-carrier DCSK [46], DCSK-ARQ/CARQ [47], improved DCSK [48], and so on, which aim at improving the system's properties, e.g., data rate, spectrum efficiency, BER performance, and physical layer security under different transmission environments. In particular, several works have recently been conducted focusing on improving the performance of DCSK communication schemes over wireless environments. The combination schemes of DCSK and diversity techniques, such as single-input multiple-output (SIMO) FM-DCSK [49], multiple-input multiple-output (MIMO) M-DCSK [50], MIMO-relay DCSK-CD [51], and cooperative communication DCSK [52], were proposed as robust solutions to achieve this improvement.

Wireless networks can be divided into two main categories, i.e., wireless local area networks (WLANs) and wireless personal area networks (WPANs) [53], where the former is designed for applications with high data rate and relatively long distance [54]. Most modern WLANs are based on IEEE 802.11 standards, marketed under the WiFi brand name [55]. In contrast, the latter focus mainly on low-data-rate and short-distance applications [56]. It is known that communication technologies today are robustly developing in the direction of increasingly enhancing data rate over a given frequency band. Nevertheless, this does not mean that the low-data-rate applications are less important. In fact, low-data-rate applications are more popular and closer to our daily lives, such as in industrial control and monitoring; environmental and health monitoring; home automation, entertainment, and toys; security, location, and asset tracking; emergency and disaster response; and control and communication in transportation [57]. The Bluetooth technique with the IEEE 802.15.1 standard was the first standard targeting at low-data-rate applications [58]. However, the complexity of this technique makes it unsuitable for simple applications requiring low cost and low power consumption. For this reason, IEEE released a new standard for low-rate WPANs (LR-WPANs), i.e., IEEE 802.15.4, which was intended to be simpler and less expensive than Bluetooth. A typical example of LR-WPANs with this standard is wireless sensor networks (WSNs) using Zigbee [59], which have been used widely in practical applications.

In this chapter, we study the modeling and performance evaluation of chaos-based communication systems for low-data-rate wireless applications [58], where two typical and popular systems, i.e., CDS-CDMA (coherent system) and DCSK (incoherent system), are selected for our investigation. The main motivation for this study is expressed by both application and academic aspects as follows: first, the design problem of low-complexity chaos-based communication systems that can exploit the flat fading characteristic of low-data-rate multipath channels in order to obtain a good performance for applications in LR-WPANs; second, most previous studies of the performance of CDS-CDMA and DCSK systems over multipath fading channels have been carried out under the context of high-data-rate transmission [21–30, 34–52], while this study presents and fully investigates these systems over low-data-rate multipath fading channels. The obtained models and analysis could motivate future studies on chaos-based low-rate wireless communications. Mathematical models in the discrete-time domain for wireless channels as well as the transmitter and receiver of CDS-CDMA and DCSK systems under the impact of noise, fading, multipath, and delay-spread for low-data-rate transmission of chaotic spreading signals are developed and analyzed. Theoretical BER expressions are theoretically determined by means of Gaussian approximation and then distribution histograms for the ratio of variable bit energy to noise power spectral density are numerically computed [24, 25]. BER performance is finally estimated by integrating the BER expressions over all possible values of the histograms. PC simulations are carried out and the simulated performances are shown in comparison with the corresponding estimated ones. The dependence of the BER on typical parameters, such as the ratio E_b/N_0, spreading factor, number of users, sample rate, and the number of paths, are evaluated in detail. It is found from the obtained results that chaos-based communication systems can perform well under the studied wireless channel, where the performances get better as the number of paths increases.

The rest of this chapter is organized as follows. In Sect. 2, the model of wireless channel for low data-rate transmission of chaotic spreading signals is described. Section 3 presents and analyses discrete-time models of CDS-CDMA and DCSK systems. In Sect. 4, theoretical performances are derived by means of Gaussian approximation. Obtained theoretical expressions are then used to compute BER performances through numerical integration in Sect. 5. The simulated results are shown in comparison with the estimated ones in Sect. 6. Our conclusion with remarks is given in Sect. 7.

2 Mathematical Modeling of Wireless Channel with Low-Rate Chaotic Spreading Signals

This section describes a simplified mathematical model for wireless channels in the context of low-data-rate transmission of chaotic spreading signals. Basically, the spread-spectrum process in chaos-based communication systems is carried out by multiplying the chaotic spreading sequence by the input data, where the chip duration T_c is much shorter than the bit duration T_b. The ratio $2\beta = T_b/T_c$ is called the spreading factor, i.e., the number of chips per bit. Suppose that the transmitted signal at the output of the transmitter is sent to the receiver over a closed space full of obstacles, which can be tables, chairs, trees, walls, and moving objects such as cars or people. The signals, after being reflected, refracted, or diffracted, can reach the receiver with various delays and attenuation and from various paths. The path having the shortest transmission period is considered the primary path, and all others are secondary paths with nonzero delays. A hypothetical scenario of the propagation in a multipath channel with a primary path and L secondary paths and its block diagram are presented in Fig. 1a, b, respectively. In our study, the transmitter and receiver are assumed to be stationary, thus phase variations of the received signals in primary and secondary channels can be ignored. The impulse response of the channel can be given by

$$h(n) = \sum_{j=0}^{L} \alpha_j \delta\left(n - \tau_j\right), \tag{1}$$

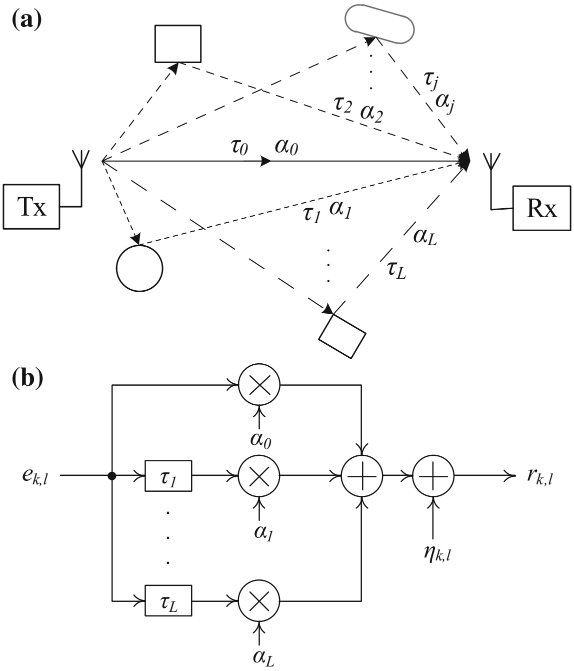


Fig. 1 Wireless channel with low-data-rate transmission of chaotic spreading signals: **a** hypothetical propagation scenario in a closed space, and **b** block diagram with a primary path and L secondary paths

where $\delta(t)$ is the Dirac impulse; τ_j and α_j are the time delay and fading coefficient of the jth path, respectively. The primary path ($j = 0$) has $\alpha_0 > 0$ and $\tau_0 = 0$. The secondary paths ($j = 1, 2, \ldots, L$) have $\alpha_j > 0$ and $\tau_j > 0$. Here, the fading coefficients α_j vary randomly according to the Rayleigh distribution [60] given by

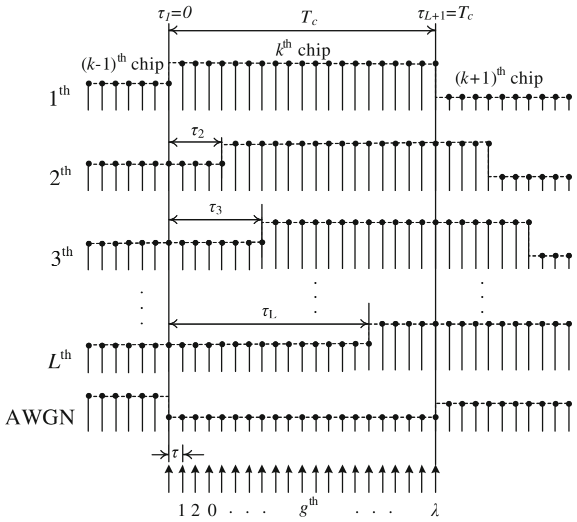
$$f(\alpha_j) = \frac{\alpha_j}{\sigma_j^2} e^{-\alpha_j^2/(2\sigma_j^2)}, \tag{2}$$

with σ_j the scale parameter of the distribution. The mean value of each fading coefficient is determined by $E[\alpha_j] = \sigma_j\sqrt{\pi/2}$. In the above parameters, the path delays τ_j, number of paths L, and spreading factor β are constants, while the fading coefficient α_j and Gaussian noise $\eta_{k,l}$ are considered independent random variables. In the scenario of low-rate transmission over a short distance, the delays τ_j are much shorter than the bit duration T_b and assumed to satisfy the condition $\tau_0 = 0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_L \le T_c$. Under this condition, the output signal of the channel can be expressed by the following sum:

$$r_{k,l} = \alpha_0 e_{k,l} + \alpha_1 e_{k-\tau_1,l} + \cdots + \alpha_L e_{k-\tau_L,l} + \eta_{k,l}, \tag{3}$$

where $e_{k,l}$ is the value of the kth chip in the lth bit in the transmitted signal, $e_{k-\tau_i,l}$ is the delayed version of $e_{k,l}$, $\alpha_j e_{k-\tau_j,l}$ is the signal on the jth secondary path, and $\eta_{k,l}$ is additive white Gaussian noise (AWGN). These signal components are illustrated in Fig. 2.

Fig. 2 Illustration of multipath components of the received signal and the sampling process in the receivers



3 Discrete-Time Modeling of Transmitter and Receiver

In this section, the discrete-time models of chaos-based communication systems over the aforementioned channel are presented and analyzed. Figure 3a, b show block diagrams of CDS-CDMA and DCSK systems, respectively.

3.1 CDS-CDMA System

In the transmitting side, the input bit streams of the K users are spread by K uncorrelated chaotic sequences produced by K chaotic generators, which use the same chaotic map with different initial conditions. Here, $b_l^{(i)} = \{\pm 1\}$ and $x_{k,l}^{(i)}$ respectively denote the lth bit of the ith user and the kth chip in the lth bit of the ith user. The output signal of the ith user in the kth chip duration is given by

$$c_{k,l}^{(i)} = b_l^{(i)} x_{k,l}^{(i)}. \tag{4}$$

The signal transmitted on the channel is the sum of all output signals from the K transmitters, which can be expressed as follows:

$$e_{k,l} = \sum_{i=1}^{K} c_{k,l}^{(i)} = \sum_{i=1}^{K} b_l^{(i)} x_{k,l}^{(i)}. \tag{5}$$

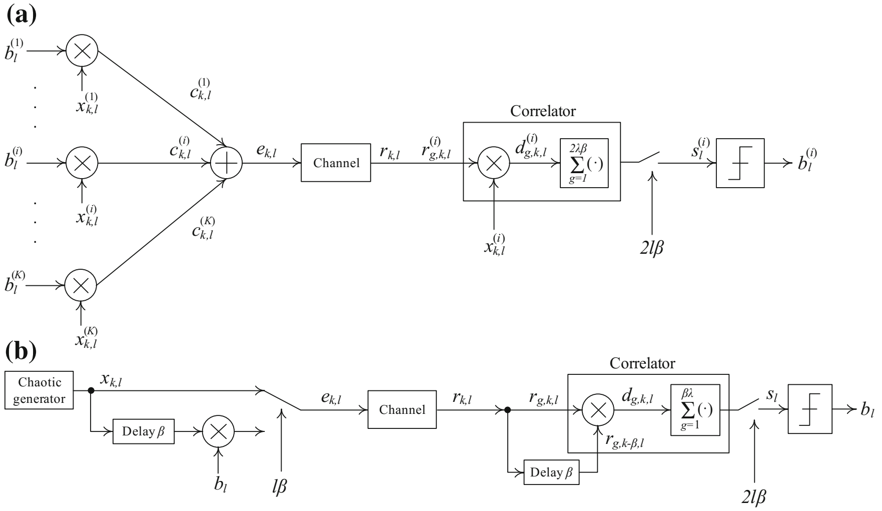


Fig. 3 Block diagrams of chaos-based communication systems under study: **a** CDS-CDMA system and **b** DCSK system

On the basid of Eq. (3), the output signal of the channel can be written as the following sum:

$$\begin{aligned} r_{k,l} =& \alpha_0 \sum_{i=1}^{K} b_l^{(i)} x_{k,l}^{(i)} + \alpha_1 \sum_{i=1}^{K} b_{l,\tau_1}^{(i)} x_{k,l,\tau_1}^{(i)} + \cdots + \alpha_L \sum_{i=1}^{K} b_{l,\tau_L}^{(i)} x_{k,l,\tau_L}^{(i)} \\ &+ \alpha_1 \sum_{i=1}^{K} b_{l,\tau_1}^{(i)} x_{k-1,l,\tau_1}^{(i)} + \cdots + \alpha_L \sum_{i=1}^{K} b_{l,\tau_L}^{(i)} x_{k-1,l,\tau_L}^{(i)} + \eta_{k,l}, \end{aligned} \tag{6}$$

where $b_{l,\tau_j}^{(i)}$ and $x_{k,l,\tau_j}^{(i)}$ are $b_l^{(i)}$ and $x_{k,l}^{(i)}$ after being delayed by a period τ_j, respectively.

In the ith receiver, the incoming signal $r_{k,l}$ is first sampled at a sampling cycle, i.e., $\tau = \frac{T_b}{2\beta\lambda} = \frac{T_c}{\lambda}$, with λ the number of samples in each chip duration T_c. With no loss of generality, the chip duration T_c and all channel delays τ_j are assumed to be equal to a multiple of the sampling cycle τ. For the sake of mathematical representation, we define $\lambda_j = \tau_j/\tau$ and $\tau_{L+1} = T_c$. This means that there are λ samples in each chip duration T_c and λ_j samples in the duration of τ_j. The oversampling process in the kth chip duration of the lth bit is also illustrated in Fig. 2. The despread spectrum process is then carried out by multiplying the sampling signal $r_{g,k,l}$ by the local chaotic sequence $x_{k,l}^{(i)}$, which is reproduced in the same way as in the transmitter and synchronized with the incoming signal by means of the synchronization methods proposed in [15–18]. Here, the integrate-and-dump block operates similarly as sum block. The consecutive samples at the input are added together in each bit duration to produce, at the output of the correlator, the signal $s_l^{(i)}$ as expressed by

$$s_l^{(i)} = \sum_{g=1}^{2\lambda\beta} d_{g,k,l}^{(i)} = \sum_{g=1}^{2\lambda\beta} r_{g,k,l}^{(i)} x_{k,l}^{(i)} = T + U + V, \tag{7}$$

where the three components T, U, V respectively are

$$T = A\left(\sum_{k=1}^{2\beta} b_l^{(i)} {x_{k,l}^{(i)}}^2 + \sum_{k=1}^{2\beta} \sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)} \right), \tag{8}$$

$$U = B\left(-\sum_{k=1}^{2\beta} b_l^{(i)} {x_{k,l}^{(i)}}^2 - \sum_{k=1}^{2\beta} \sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)} + \sum_{k=1}^{2\beta} \sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k-1,l}^{(m)} x_{k,l}^{(i)} \right), \tag{9}$$

and

$$V = \lambda \sum_{k=1}^{2\beta} \eta_{k,l} x_{k,l}^{(i)}, \tag{10}$$

with $A = \sum_{j=0}^{L} \lambda \alpha_j$ and $B = \sum_{j=1}^{L} \lambda_j \alpha_j$. (The detailed development of Eqs. (7)–(10) is presented in Appendix A.) In Eqs. (8)–(10), we can find that V is the noise component created by the AWGN. The sum of $(T + U)$ can be rewritten in another form as follows:

$$T + U = S + I, \tag{11}$$

where S is the component of the beneficial signal determined by

$$S = (A - B) \sum_{k=1}^{2\beta} b_l^{(i)} {x_{k,l}^{(i)}}^2, \tag{12}$$

and I is the interference component between different users determined by

$$I = (A - B) \sum_{k=1}^{2\beta} \sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)} + B \sum_{k=1}^{2\beta} \sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k-1,l}^{(m)} x_{k,l}^{(i)}. \tag{13}$$

Based on Eqs. (10), (12), (13), it is easy to find that the values of the components S, I, and V increase as the number of delay paths L or the number of samples per chip λ increases. However, owing to the low cross-correlation between the different chaotic sequences and between the chaotic sequences and AWGN, the increasing number of the components I and V is much less than that of the component S. Therefore, the SNR of the output signal $s_l^{(i)}$ increases. The increment of the SNR leads to improvement in the system's performance. The binary value of the lth bit is finally recovered by the decision circuit using a sign function as follows:

$$b_l^{(i)} = \begin{cases} 1 & s_l^{(i)} \geq 0, \\ -1 & s_l^{(i)} < 0. \end{cases} \tag{14}$$

3.2 DCSK System

In the DCSK transmitter, each transmitted bit duration T_b is divided into two equal time intervals. The first interval is used to transmit the chaotic reference sequence. The second one is to send the data-bearing sequence. During the second interval, if a bit $+1$ is transmitted, the chaotic reference sequence is repeated, while an inverted copy of the reference sequence is sent if a bit -1 is transmitted. The output signal $e_{k,l}$ of the transmitter in the lth bit duration is given by

$$e_{k,l} = \begin{cases} x_{k,l}, & k = 1, \ldots, \beta, \\ b_l x_{k-\beta,l}, & k = \beta+1, \ldots, 2\beta, \end{cases} \tag{15}$$

where b_l is the binary value of the lth bit, $x_{k,l}$ is the value of the kth chip in the lth bit in the reference sequence, and $x_{k-\beta,l}$ is the delayed version of $x_{k,l}$.

In the receiver, the signal components and the sampling process of the incoming signal $r_{k,l}$ are the same as in Fig. 2. Within the duration of the kth chip in the second interval of lth bit, we find that the received signal consists of three components, i.e., the kth chip and its delayed parts with duration of $(T_c - \tau_j)$, the delayed parts with duration τ_j of the $(k-1)$th chip, and AWGN. It also can be found that all samples in the duration of $(t_j, t_{j+1}]$ have the same value. The value of samples falling into the duration of $(t_j, t_{j+1}]$ is determined by

$$\begin{aligned} r_{g,k,l} &= \alpha_0 e_{k,l} + \cdots + \alpha_j e_{k,l} + \alpha_{j+1} e_{k-1,l} + \cdots + \alpha_L e_{k-1,l} + \eta_{k,l} \\ &= e_{k,l} \sum_{i=0}^{j} \alpha_i + e_{k-1,l} \sum_{i=j+1}^{L} \alpha_i + \eta_{k,l} \\ &= b_l x_{k-\beta,l} \sum_{i=0}^{j} \alpha_i + b_l x_{(k-1)-\beta,l} \sum_{i=j+1}^{L} \alpha_i + \eta_{k,l}, \end{aligned} \tag{16}$$

where $r_{g,k,l}$ is the value of the gth sample in the kth chip duration, satisfying $\lambda_j < g \leq \lambda_{j+1}$. Similarly, we can express the value of samples in $(t_j, t_{j+1}]$ of the signal at the output of the delay β block as follows:

$$r_{g,k-\beta,l} = x_{k-\beta,l} \sum_{i=0}^{j} \alpha_i + x_{(k-1)-\beta,l} \sum_{i=j+1}^{L} \alpha_i + \eta_{k-\beta,l}. \tag{17}$$

Based on the results in Eqs. (16) and (17), the value of the samples in $(t_j, t_{j+1}]$ at the output of the multiplier is determined by

$$\begin{aligned} d_{g,k,l} &= r_{g,k,l} r_{g,k-\beta,l} \\ &= \left(\sum_{i=0}^{j} \alpha_i\right)^2 b_l x_{k-\beta,l}^2 + \left(\sum_{i=j+1}^{L} \alpha_i\right)^2 b_l x_{(k-1)-\beta,l}^2 + 2\sum_{i=0}^{j} \alpha_i \sum_{i=j+1}^{L} \alpha_i \, b_l x_{k-\beta,l} x_{(k-1)-\beta,l} \\ &+ \left(\sum_{i=0}^{j} \alpha_i \, x_{k-\beta,l} + \sum_{i=j+1}^{L} \alpha_i \, x_{(k-1)-\beta,l}\right) \times \left(\eta_{k,l} + b_l \eta_{k-\beta,l}\right) + \eta_{k,l}\eta_{k-\beta,l}. \end{aligned} \tag{18}$$

In the sum block, the consecutive samples at the input are added together in the second half of each bit duration to produce a decision variable s_l as follows:

$$\begin{aligned} s_l &= \sum_{g=1}^{\beta\lambda} d_{g,k,l} = \sum_{k=1}^{\beta}\sum_{g=1}^{\lambda} d_{g,k,l} = \sum_{k=1}^{\beta}\sum_{j=0}^{L} (\lambda_{j+1} - \lambda_j) d_{g,k,l} \\ &= A b_l \sum_{k=1}^{\beta} x_{k-\beta,l}^2 + B b_l \sum_{k=1}^{\beta} x_{(k-1)-\beta,l}^2 + C b_l \sum_{k=1}^{\beta} x_{k-\beta,l} x_{(k-1)-\beta,l} \\ &+ D \sum_{k=1}^{\beta} x_{k-\beta,l} \left(\eta_{k,l} + b_l \eta_{k-\beta,l}\right) + E \sum_{k=1}^{\beta} x_{(k-1)-\beta,l} \left(\eta_{k,l} + b_l \eta_{k-\beta,l}\right) + \lambda \sum_{k=1}^{\beta} \eta_{k,l}\eta_{k-\beta,l}, \end{aligned} \tag{19}$$

where

$$C = \sum_{j=0}^{L} (\lambda_{j+1} - \lambda_j) \left(\sum_{i=0}^{j} \alpha_i\right)^2, \tag{20}$$

$$D = \sum_{j=0}^{L} (\lambda_{j+1} - \lambda_j) \left(\sum_{i=j+1}^{L} \alpha_i\right)^2, \tag{21}$$

$$E = 2\sum_{j=0}^{L} (\lambda_{j+1} - \lambda_j) \sum_{i=0}^{j} \alpha_i \sum_{i=j+1}^{L} \alpha_i, \tag{22}$$

$$F = \sum_{j=0}^{L} (\lambda_{j+1} - \lambda_j) \sum_{i=0}^{j} \alpha_i, \tag{23}$$

and

$$G = \sum_{j=0}^{L} (\lambda_{j+1} - \lambda_j) \sum_{i=j+1}^{L} \alpha_i. \tag{24}$$

We find that the signal s_l contains the following components: the component of the beneficial signal, denoted by

$$W = Cb_l \sum_{k=1}^{\beta} x_{k-\beta,l}^2 + Db_l \sum_{k=1}^{\beta} x_{(k-1)-\beta,l}^2; \tag{25}$$

the interference component between the current chip with its delayed version, denoted by

$$X = Eb_l \sum_{k=1}^{\beta} x_{k-\beta,l} x_{(k-1)-\beta,l}; \tag{26}$$

the noise component created by the impact of the AWGN, denoted by

$$\begin{aligned} Y = F \sum_{k=1}^{\beta} x_{k-\beta,l} \left(\eta_{k,l} + b_l \eta_{k-\beta,l}\right) \\ + G \sum_{k=1}^{\beta} x_{(k-1)-\beta,l} \left(\eta_{k,l} + b_l \eta_{k-\beta,l}\right); \end{aligned} \tag{27}$$

and the noise component created by the AWGN only, denoted by

$$Z = \lambda \sum_{k=1}^{\beta} \eta_{k,l} \eta_{k-\beta,l}. \tag{28}$$

It can be seen from Eqs. (20)–(28) that the values of the components W, X, Y, and Z increase with the number of delay paths L, and the increase in the number of components X, Y, and Z is much less than that of the component W. This makes the SNR of the output signal s_l increase. Finally, the binary value of the lth bit is recovered according to the following rule:

$$b_l = \begin{cases} 1 & s_l \geq 0, \\ -1 & s_l < 0. \end{cases} \tag{29}$$

4 Theoretical Derivation of BER Performance

This section presents the theoretical derivation of BER performance for the CDS-CDMA and DCSK systems. First, the static BER expression is derived based on Gaussian approximation. The dynamic expression is then produced by reflecting the chaotic variation in the static one.

4.1 Static BER Expression

In the above equations, the components, i.e., T, U, V of the variable $s_l^{(i)}$, and W, X, Y, Z of the variable s_l, are random functions. Owing to the independence between fading coefficients and delays in the primary and secondary paths and their independence to AWGN, these components are also independent.

Firstly, the statistics of the components, i.e., T, U, V and W, X, Y, Z, in the case of a $+1$ bit transmitted are determined, under the assumption that the value of the spreading factor is high enough that the correlation values of the independent variables in these components are approximately equal to zero. The mean and mean squared values of these components are determined as follows (see Appendix B and Appendix C):

$$E[T|b_l^{(i)} = +1] = 2A\beta E_c, \tag{30}$$

$$E[U|b_l^{(i)} = +1] = -2B\beta E_c, \tag{31}$$

$$E[V|b_l^{(i)} = +1] = \lambda E\left[\sum_{k=1}^{2\beta} \eta_{k,l} x_{k,l}^{(i)}\right] = 0, \tag{32}$$

$$E[T^2|b_l^{(i)} = +1] = A^2\left(2\beta E_{c4} + 2\beta(2\beta + K - 2)E_c^2\right), \tag{33}$$

$$E[U^2|b_l^{(i)} = +1] = B^2\left(2\beta E_{c4} + 2\beta(2\beta + 2K - 2)E_c^2\right), \tag{34}$$

$$E[V^2|b_l^{(i)} = +1] = \lambda^2 E_b \frac{N_0}{2}, \tag{35}$$

and

$$E[W|b_l = +1] = (C + D)\beta E_c, \tag{36}$$

$$E[X|b_l = +1] = E[Y|b_l = +1] = E[Z|b_l = +1] = 0, \tag{37}$$

$$E[W^2|b_l = +1] = (C^2 + D^2)\beta(E_{c4} + (\beta - 1)E_c^2) + 2CD\beta^2 E_c^2, \tag{38}$$

$$E[X^2|b_l = +1] = E^2\beta E_c^2, \tag{39}$$

$$E[Y^2|b_l = +1] = (F^2 + G^2)\beta N_0 E_c, \tag{40}$$

$$E[Z^2|b_l = +1] = \lambda^2\beta\frac{N_0^2}{4}, \tag{41}$$

where $E_c = E\left[x_{k,l}^{(i)2}\right] = E[x_{k,l}^2]$, $E_{c4} = E\left[x_{k,l}^{(i)4}\right] = E[x_{k,l}^4]$, $E_b = 2\beta E_c$, and $N_0 = 2E\left[\eta_{k,l}^2\right]$.

Based on the resulting values above, the variances of the components are calculated by

$$\begin{aligned} Var[T|b_l^{(i)}=+1] &= E[T^2|b_l^{(i)}=+1]-E^2[T|b_l^{(i)}=+1] \\ &= A^2\left(2\beta E_{c4}+2\beta(K-2)E_c^2\right), \end{aligned} \tag{42}$$

$$\begin{aligned} Var[U|b_l^{(i)}=+1] &= E[U^2|b_l^{(i)}=+1]-E^2[U|b_l^{(i)}=+1] \\ &= B^2\left(2\beta E_{c4}+2\beta(2K-2)E_c^2\right), \end{aligned} \tag{43}$$

$$Var[V|b_l^{(i)}=+1]=E[V^2|b_l^{(i)}=+1]-E^2[V|b_l^{(i)}=+1]=\lambda^2 E_b\frac{N_0}{2}, \tag{44}$$

and

$$Var[W|b_l=+1]=E[W^2|b_l=+1]-E^2[W|b_l=+1]=(C^2+D^2)\beta(E_{c4}-E_c^2), \tag{45}$$

$$Var[X|b_l=+1]=E[X^2|b_l=+1]-E^2[X|b_l=+1]=E^2\beta E_c^2, \tag{46}$$

$$Var[Y|b_l=+1]=E[Y^2|b_l=+1]-E^2[Y|b_l=+1]=(F^2+G^2)\beta N_0 E_c, \tag{47}$$

$$Var[Z|b_l=+1]=E[Z^2|b_l=+1]-E^2[Z|b_l=+1]=\lambda^2\beta\frac{N_0^2}{4}. \tag{48}$$

Due to the statistical independence between the components, i.e., T, U, V and W, X, Y, Z, the mean value and variance of the decision variables, $s_l^{(i)}$ and s_l, are respectively determined as

$$E[s_l^{(i)}|b_l^{(i)}=+1]=E[T|b_l^{(i)}=+1]+E[U|b_l^{(i)}=+1]+0, \tag{49}$$

$$Var[s_l^{(i)}|b_l^{(i)}=+1]=Var[T|b_l^{(i)}=+1]+Var[U|b_l^{(i)}=+1]+Var[V|b_l^{(i)}=+1], \tag{50}$$

and

$$\begin{aligned} E[s_l|b_l=+1] &= E[W|b_l=+1]+E[X|b_l=+1]+E[Y|b_l=+1]+E[Z|b_l=+1] \\ &= E[W|b_l=+1], \end{aligned} \tag{51}$$

$$\begin{aligned} Var[s_l|b_l=+1] = {} & Var[W|b_l=+1]+Var[X|b_l=+1]+Var[Y|b_l=+1], \\ & + Var[Z|b_l=+1]. \end{aligned} \tag{52}$$

Secondly, the case of a -1 bit transmitted is considered. Analogously, we have

$$E[s_l^{(i)}|b_l^{(i)}=-1]=-E[s_l^{(i)}|b_l^{(i)}=+1], \tag{53}$$

$$Var[s_l^{(i)}|b_l^{(i)}=-1]=Var[s_l^{(i)}|b_l^{(i)}=+1], \tag{54}$$

and

$$E[s_l|b_l=-1]=-E[s_l|b_l=+1], \tag{55}$$

$$Var[s_l|b_l=-1]=Var[s_l|b_l=+1], \tag{56}$$

assuming that each bit, either $+1$ or -1, appears at the output of the data source with a probability of $1/2$. Based on the obtained results above and according to the central limit theorem (CLT) [61], the statical BER expressions for CDS-CDMA and DCSK systems can be respectively derived by Gaussian approximation as follows:

$$\begin{aligned} BER_{CDS-CDMA} &= \frac{1}{2}\Pr(s_l^{(i)}\leq 0|b_l^{(i)}=+1)+\frac{1}{2}\Pr(s_l^{(i)}>0|b_l^{(i)}=-1)\\ &= \Pr(s_l^{(i)}\leq 0|b_l^{(i)}=+1)=Q\left(\frac{Var[s_l^{(i)}|b_l^{(i)}=+1]}{E^2[s_l^{(i)}|b_l^{(i)}=+1]}\right)^{-\frac{1}{2}}\\ &= Q\left(\frac{\frac{E_{c4}}{E_c^2}+K-2}{2\beta\left(1-\frac{B}{A}\right)^2}+\frac{\frac{E_{c4}}{E_c^2}+2K-2}{2\beta\left(\frac{A}{B}-1\right)^2}+\frac{1}{2\frac{E_b}{N_0}\left(\frac{A-B}{\lambda}\right)^2}\right)^{-\frac{1}{2}}, \end{aligned} \tag{57}$$

and

$$\begin{aligned} BER_{DCSK} &= \frac{1}{2}\Pr(s_l\leq 0|b_l=+1)+\frac{1}{2}\Pr(s_l>0|b_l=-1)\\ &= \Pr(s_l\leq 0|b_l=+1)=Q\left(\frac{Var[s_l|b_l=+1]}{E^2[s_l|b_l=+1]}\right)^{-\frac{1}{2}}\\ &= Q\left(\frac{(C^2+D^2)(\frac{E_{c4}}{E_c^2}-1)+E^2}{\beta(C+D)^2}+\frac{\frac{2(F^2+G^2)}{(C+D)\lambda}}{\frac{(C+D)}{\lambda}\frac{E_b}{N_0}}+\frac{\beta}{\left(\frac{(C+D)}{\lambda}\frac{E_b}{N_0}\right)^2}\right)^{-\frac{1}{2}}, \end{aligned} \tag{58}$$

where the function $Q(\cdot)$ is defined by $Q(\epsilon)=\frac{1}{\sqrt{2\pi}}\int\limits_{\epsilon}^{\infty}\exp(y^2/2)dy$, and $\frac{E_b}{N_0}$ is the ratio of average bit energy to noise power spectral density.

4.2 Dynamical BER Expression

In the studied channel, since the fading coefficients vary randomly according to the Rayleigh distribution, the elements in parentheses in Eqs. (57) and (58), respectively denoted by $\phi_{CDS-CDMA} = \frac{\frac{E_{c4}}{E_c^2}+K-2}{2\beta\left(1-\frac{B}{A}\right)^2}$, $\psi_{CDS_CDMA}=\frac{\frac{E_{c4}}{E_c^2}+2K-2}{2\beta\left(\frac{A}{B}-1\right)^2}$, $\gamma_{CDS-CDMA} = \frac{E_b}{N_0}\left(\frac{A-B}{\lambda}\right)^2$, and $\phi_{DCSK} = \frac{(C^2+D^2)(\frac{E_{c4}}{E_c^2}-1)+4E^2}{\beta(C+D)^2}$, $\psi_{DCSK} = \frac{2(F^2+G^2)}{(C+D)\lambda}$, $\gamma_{DCSK} = \frac{C+D}{\lambda}\frac{E_b}{N_0}$, also randomly vary in the communication process. Here, the elements, i.e., $\gamma_{CDS-CDMA}$ and γ_{DCSK}, are considered the main elements because they fully depend on the bit energy E_b and all parameters of the channel, i.e., L, α_j, λ_j, N_0. To simplify our analysis, the elements ϕ_{CDS_CDMA}, ψ_{CDS_CDMA}, and ϕ_{DCSK}, ψ_{DCSK} are approximated to constants, which are respectively equal to $\overline{\phi}_{CDS-CDMA} = \frac{\frac{E_{c4}}{E_c^2}+K-2}{2\beta\left(1-\frac{\overline{B}}{\overline{A}}\right)^2}$, $\overline{\psi}_{CDS-CDMA} = \frac{\frac{E_{c4}}{E_c^2}+2K-2}{2\beta\left(\frac{\overline{A}}{\overline{B}}-1\right)^2}$, and $\overline{\phi}_{DCSK} = \frac{(\overline{C}^2+\overline{D}^2)(\frac{E_{c4}}{E_c^2}-1)+4E^2}{\beta(\overline{C}+\overline{D})^2}$, $\overline{\psi}_{DCSK} = \frac{2(\overline{F}^2+\overline{G}^2)}{(\overline{C}+\overline{D})\lambda}$. The constants, i.e., $\overline{A}$, $\overline{B}$, $\overline{C}$, $\overline{D}$, $\overline{E}$, $\overline{F}$, and $\overline{G}$, are respectively obtained by replacing the variable coefficients α_j in A, B, C, D, E, F, and G by their mean values, i.e., $\overline{\alpha}_j = \sigma_j\sqrt{\pi/2}$.

The static BER expressions in Eq. (57) and (58) are approximated by

$$BER_{CDS-CDMA} \approx Q\left(\overline{\phi}_{CDS-CDMA} + \overline{\psi}_{CDS-CDMA} + \frac{1}{2\gamma_{CDS-CDMA}}\right)^{-\frac{1}{2}} \tag{59}$$

and

$$BER_{DCSK} \approx Q\left(\overline{\phi}_{DCSK} + \frac{\overline{\psi}_{DCSK}}{\gamma_{DCSK}} + \frac{\beta}{\gamma_{DCSK}^2}\right)^{-\frac{1}{2}}. \tag{60}$$

On the basis of Eqs. (59) and (60), the dynamic BER expressions reflecting the variations of chaotic sequence and fading coefficients corresponding to the two systems are obtained as follows:

$$\begin{aligned} BER_{CDS-CDMA}(\gamma_{CDS-CDMA}) \approx \int_0^\infty Q\left(\overline{\phi}_{CDS-CDMA} + \overline{\psi}_{CDS-CDMA} + \frac{1}{2\gamma_{CDS-CDMA}}\right)^{-\frac{1}{2}} \\ \times f(\gamma_{CDS-CDMA})d\gamma_{CDS-CDMA} \end{aligned} \tag{61}$$

and

$$BER_{DCSK}(\gamma_{DCSK}) \approx \int_0^\infty Q\left(\overline{\phi}_{DCSK} + \frac{\overline{\psi}_{DCSK}}{\gamma_{DCSK}} + \frac{\beta}{\gamma_{DCSK}^2}\right)^{-\frac{1}{2}} \times f(\gamma_{DCSK})d\gamma_{DCSK}, \tag{62}$$

with $f(\gamma_{CDS-CDMA})$ and $f(\gamma_{DCSK})$ the probability density functions (PDF) of the elements $\gamma_{CDS-CDMA}$ and γ_{DCSK}, respectively.

5 Performance Estimation Using Numerical Integration

First, we consider the performances of CDS-CDMA and DCSK systems in the special case of a multipath fading channel, i.e., one-path Rayleigh fading channel, which is equivalent to the studied channel in the simple case of $L = 0$ and $\alpha_0 > 0$. Under this condition, the parameters of two systems have the following values: $\lambda_0 = 0$, $\lambda_1 = \lambda$, $A = \lambda\alpha_0$, $B = 0$, $C = \lambda\alpha_0^2$, and $F = \lambda\alpha_0$, $D = E = G = 0$. The elements in the dynamic BER expressions become as follows: $\overline{\phi}_{CDS-CDMA} = \left(\frac{E_{c4}}{E_c^2} + K - 2\right)/(2\beta)$, $\overline{\psi}_{CDS-CDMA} = 0$, $\overline{\phi}_{DCSK} = \left(\frac{E_{c4}}{E_c^2} - 1\right)/\beta$, $\overline{\psi}_{DCSK} = 2$, and $\gamma_{CDS-CDMA} = \gamma_{DCSK} = \alpha_0^2 \frac{E_b}{N_0}$. According to [37, 39], the PDFs of $\gamma_{CDS-CDMA}$ and γ_{DCSK} are given by

$$f(\gamma_{CDS-CDMA}) = f(\gamma_{DCSK}) = \frac{1}{2\sigma_0^2 \frac{E_b}{N_0}}\, e^{-\frac{\alpha_0^2}{2\sigma_0^2}}. \tag{63}$$

The dynamic BER expressions in this special case are determined by

$$BER_{CDS-CDMA} \approx \int_0^\infty Q\left(\frac{\frac{E_{c4}}{E_c^2} + K - 2}{2\beta} + \frac{1}{2\alpha_0^2 \frac{E_b}{N_0}}\right)^{-\frac{1}{2}} \frac{1}{2\sigma_0^2 \frac{E_b}{N_0}}\, e^{-\frac{\alpha_0^2}{2\sigma_0^2}}\, d\left(\alpha_0^2 \frac{E_b}{N_0}\right) \tag{64}$$

and

$$BER_{DCSK} \approx \int_0^\infty Q\left(\frac{\frac{E_{c4}}{E_c^2} - 1}{\beta} + \frac{2}{\alpha_0^2 \frac{E_b}{N_0}} + \frac{\beta}{\left(\alpha_0^2 \frac{E_b}{N_0}\right)^2}\right)^{-\frac{1}{2}} \frac{1}{2\sigma_0^2 \frac{E_b}{N_0}}\, e^{-\frac{\alpha_0^2}{2\sigma_0^2}}\, d\left(\alpha_0^2 \frac{E_b}{N_0}\right). \tag{65}$$

It can be seen from the above analysis that since we know the PDFs exactly, i.e., $f(\gamma_{CDS-CDMA})$ and $f(\gamma_{DCSK})$, in the simple case of $L = 0$, the BER of the systems can be totally determined by means of the theoretical expressions. However, in our study, the channel under investigation is generalized with the number of paths $L \geq 1$. Therefore, the greater the number of secondary paths L, the more complicated, if not impossible, the theoretical determination of PDFs becomes. For this reason, a simpler approach to estimate the PDFs and performances using numerical integration is presented. In this approach, the numerical computation method proposed in [24, 25] is used to determine histograms of the value distributions of $\gamma_{CDS-CDMA}$ and γ_{DCSK} instead of theoretically determining their PDFs. The sample values of $\gamma_{CDS-CDMA}$ and γ_{DCSK} are first computed and then used to build the histogram of the value distribution. With the assumption that the values of $\gamma_{CDS-CDMA}$ and γ_{DCSK} are the outputs of stationary random processes, the obtained histograms can be considered a good estimate of the PDFs [62]. Based on the BER expressions theoretically obtained in

Eqs. (61) and (62) as well as the distribution histograms obtained by the computation, the BER performances of the two proposed receivers can be respectively calculated by the following numerical integrations:

$$\begin{aligned} &BER_{CDS-CDMA}(\gamma_{CDS-CDMA}) \\ &\approx \sum_{m=1}^{N} Q\left(\overline{\phi}_{CDS-CDMA} + \overline{\psi}_{CDS-CDMA} + \frac{1}{2\gamma_{CDS-CDMA,m}}\right)^{-\frac{1}{2}} P(\gamma_{CDS-CDMA,m}) \end{aligned} \tag{66}$$

and

$$BER_{DCSK}(\gamma_{DCSK}) \approx \sum_{m=1}^{N} Q\left(\overline{\phi}_{DCSK} + \frac{\overline{\psi}_{DCSK}}{\gamma_{DCSK,m}} + \frac{\beta}{\gamma^2_{DCSK,m}}\right)^{-\frac{1}{2}} P(\gamma_{DCSK,m}), \tag{67}$$

where N is the number of classes of the histogram, and $P(\gamma_{CDS-CDMA,m})$ and $P(\gamma_{DCSK,m})$ are the probabilities of having the energy in intervals centered on $\gamma_{CDS-CDMA,m}$ and $\gamma_{DCSK,m}$, respectively.

6 Simulation Results

In this section, the performances of two systems obtained by the analysis according to Eqs. (66), (67) and the corresponding numerical simulations are displayed in the same graphs for comparison. The figures display the BER performance according to the typical parameters, i.e., the ratio E_b/N_0, spreading factor 2β, number of users K, and number of samples per chip λ, as the number of paths L gradually increases from 1 to 4. The chaotic map used for generating the chaotic sequence is the Chebyshev polynomial function of order 2 [63] given by

$$x_k = f(x_{k-1}) = 2x^2_{k-1} - 1, \tag{68}$$

with the invariant PDF of x, denoted by $\rho(x)$, being

$$\rho(x) = \begin{cases} \frac{1}{\pi\sqrt{1-x^2}} & |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{69}$$

The values of E_c and E_{c4} are determined by

$$E_c = E\left[x^{(i)\,2}_{k,l}\right] = \int_{-\infty}^{\infty} x^2 \rho(x) dx = \int_{-1}^{1} x^2 \frac{1}{\pi\sqrt{1-x^2}} dx = \frac{1}{2}, \tag{70}$$

$$E_{c4} = E\left[x_{k,l}^{(i)4}\right] = \int_{-\infty}^{\infty} x^4 \rho(x)dx = \int_{-1}^{1} x^4 \frac{1}{\pi\sqrt{1-x^2}}dx = \frac{3}{8}. \tag{71}$$

The parameters of the channel are set as follows: $\sigma_0 = 0.7$ for the primary path; $\sigma_1 = 0.6, \tau_1 = 5\tau, \sigma_2 = 0.5, \tau_2 = 10\tau, \sigma_3 = 0.4, \tau_3 = 15\tau$, and $\sigma_4 = 0.3, \tau_4 = 20\tau$, for the secondary paths.

Figure 4a, b show respectively the histograms of the value distributions of $\gamma_{CDS-CDMA}$ and γ_{DCSK} obtained by numerical computation for the case of $E_b/N_0 = 10dB$ and $\lambda = 30$. Each histogram was plotted with $N = 1000$ classes, which are calculated statistically from 100000 samples. It clearly appears that incrementing L leads to changes in value distribution, specifically, the average values and variation ranges of $\gamma_{CDS-CDMA,m}$ and $\gamma_{DCSK,m}$ increase. Due to the property of the Q function, we can find from Eqs. (66) and (67) that these changes will make the BERs decrease. In addition, with the same value of L, the average value and variation range of $\gamma_{DCSK,m}$ are always higher than those of $\gamma_{CDS-CDMA,m}$.

In Fig. 5a, b, we study the effect of the ratio E_b/N_0 and number of paths L on the performance of the two systems in the case of $\lambda = 30$ and $2\beta = 64$. The BER performances obtained in Eqs. (64) and (65) for the special case of the channel with $L = 0$ are also plotted. It can be seen that there is a good match between the analysis and simulation performances. The performance of two systems is significantly improved when the number of paths increases, where the performance of CDS-CDMA gets better than that of DSCK with a higher value of L. For example, at the same $E_b/N_0 = 10dB$, the BER values of the CDS-CDMA and DCSK systems corresponding to $L = 0, 1, 2, 3, 4$ are $1.4 \cdot 10^{-2}$, $1.1 \cdot 10^{-3}$, $6.2 \cdot 10^{-4}$, $1.2 \cdot 10^{-5}$, $4.2 \cdot 10^{-6}$; and $1.9 \cdot 10^{-1}$, $4.2 \cdot 10^{-2}$, $7.9 \cdot 10^{-3}$, $1.7 \cdot 10^{-3}$, $5.8 \cdot 10^{-4}$, respectively.

The dependence of the CDS-CDMA performance on the number of users K with different values of E_b/N_0 is shown in Fig. 6. We can observe that the BER performance becomes significantly worse with incrementing K. Specifically, at the same value of $E_b/N_0 = 10dB$, the BERs obtained from the simulation increase from $9.1 \cdot 10^{-4}$ to $1.1 \cdot 10^{-2}$, corresponding to K changing from 1 to 10. In general, the simulation results agree with those obtained by our estimation. However, with $K = 5$, slight differences between them begin to be visible. If we continue to increase $K \geq 5$, these differences become more and more clear. The reason is that the variables A, B, C, D, E, F, G in Eqs. (57), (58) are respectively approximated by the fixed values $\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}, \overline{F}, \overline{G}$ in Eqs. (61), (62) in order to make the BER expressions simpler.

The effect of the value of the spreading factor 2β on the BER performances of CDS-CDMA and DCSK systems in the case of $E_b/N_0 = 10dB$ and $\lambda = 30$ is shown in Fig. 7a, b, respectively. For the CDS-CDMA system, the BER performance becomes slightly better with the increment of 2β. For example, at the same value of $E_b/N_0 = 10dB$, the obtained BERs decrease from $1.1 \cdot 10^{-4}$ to $8.1 \cdot 10^{-5}$, corresponding to 2β changing from 16 to 128. With respect to the DCSK system, we can see that in the value range from 24 to 64, the value increments of 2β and corresponding BERs are directly proportional to each other. For example, in the case

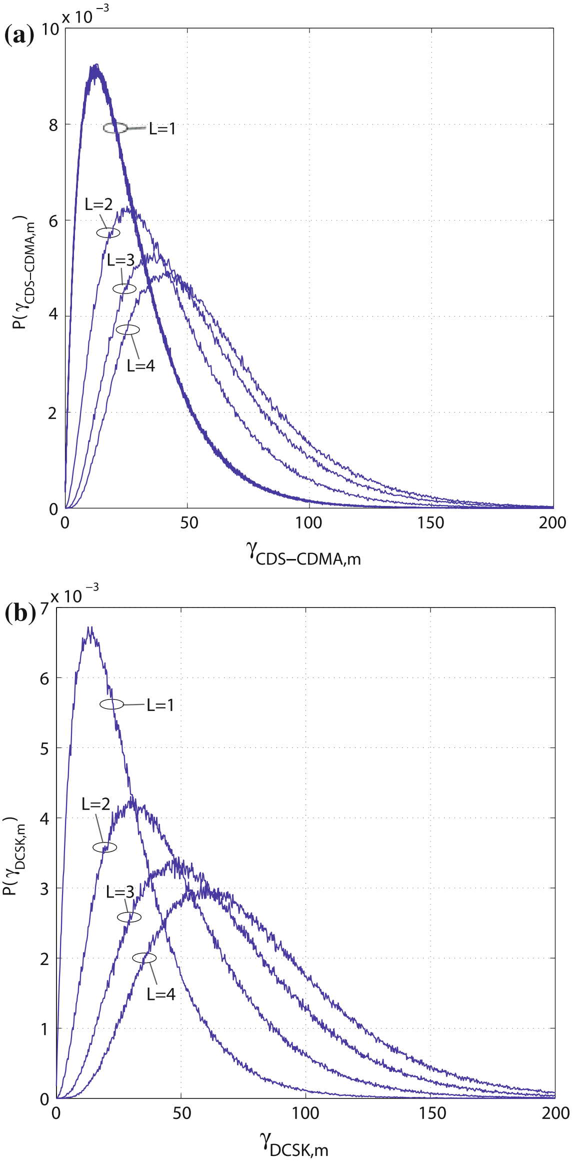


Fig. 4 Histograms of the value distribution of the elements **a** $\gamma_{CDS-CDMA}$ and **b** γ_{DCSK}, in the case of $E_b/N_0 = 10dB$ and $\lambda = 30$

of $L = 3$, the BER values increase respectively from $1.5 \cdot 10^{-3}$ to $3.4 \cdot 10^{-3}$, corresponding to β changing from 24 to 64. In contrast, in the value range from 4 to 24, an increment of 2β makes the BERs decrease. In particular, the minimum values of BER are obtained with the value of 2β within 8 and 16, which means that good performance is obtained at low values of the spreading factor, or in other words, the DCSK system can perform well even with a moderate bandwidth. Generally, the

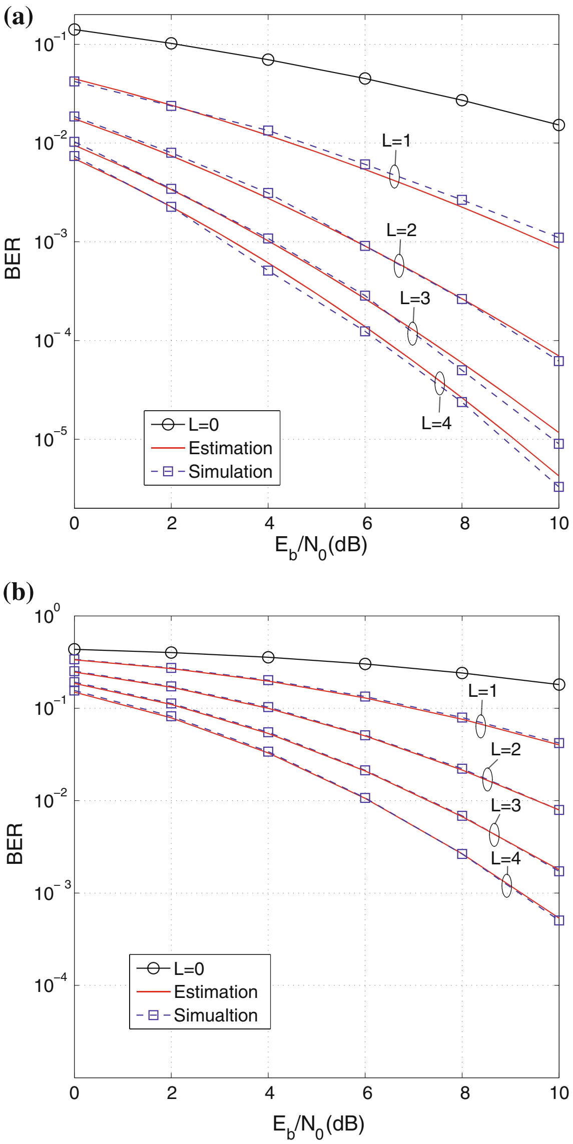
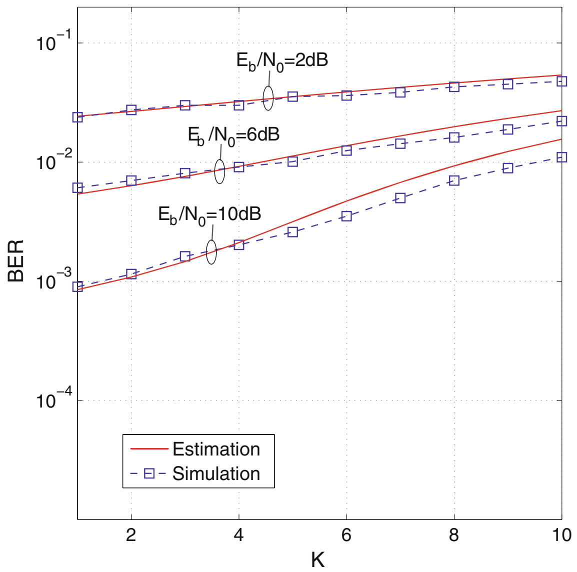


Fig. 5 BER values against the ratio E_b/N_0 with an increment in the number of secondary paths L in the case of $2\beta = 64$ and $\lambda = 30$: **a** CDS-CDMA system and **b** DCSK system

simulation results agree exactly with the analysis except for the ranges of $2\beta \leq 32$ and $2\beta \leq 24$ in the cases of the CDS-CDMA and DCSK systems, respectively. The differences become more pronounced at higher values of L. There are two causes for these differences. First, our assumption in Sect. 4, i.e., that the correlation values of the independent variables are approximately equal to zero, is no longer satisfied with low spreading factor. Second, the elements, ϕ_{SDS} and ϕ_{DSS}, ψ_{DSS}, in Eqs. (57)

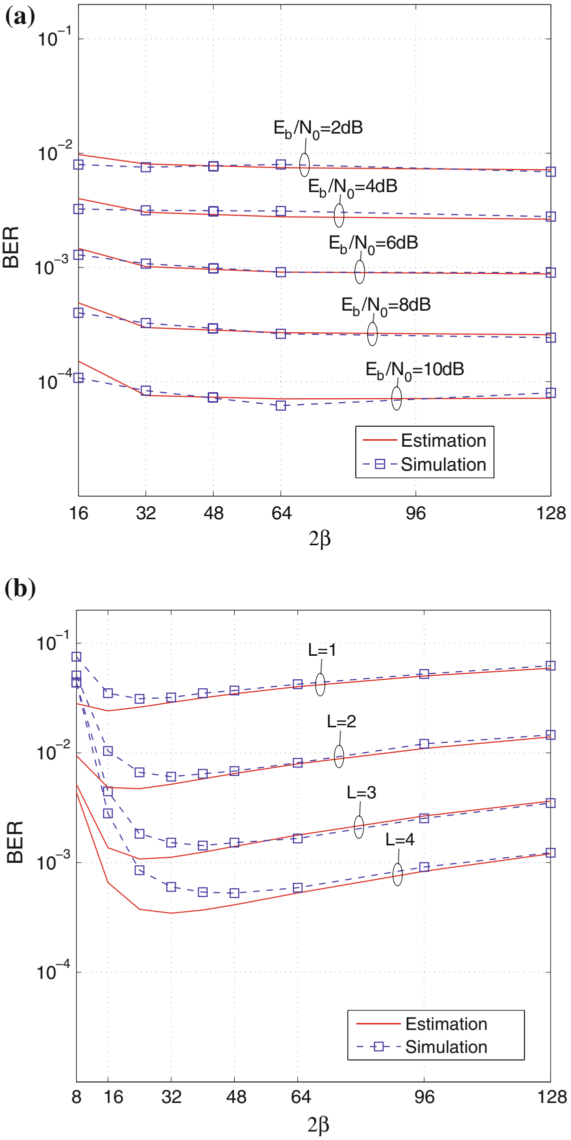
Fig. 6 BER values against the number of users K in the CDS-CDMA system with an increment of E_b/N_0 in the case $L = 1, 2\beta = 64, \lambda = 30$



and (58) are approximated by the constants $\overline{\phi}_{SDS}$ and $\overline{\phi}_{DSS}$, $\overline{\psi}_{DSS}$, in Eqs. (61) and (62), respectively. Note that these assumptions and approximations aim to simplify our BER analysis, but simultaneously, they also create differences.

In the CDS-CDMA and DCSK systems, with fixed channel delays τ_j, the number of samples per chip λ can be changed in the following two cases: (1) keeping the chip duration fixed and changing the sampling clock in the receiver and (2) keeping the sampling clock constant and changing the chip duration. In the first case, in spite of the sampling clock changing, the ratio $\lambda_j/\lambda = \tau_j/T_c$ is still unvaried. Notice in Eqs. (57) and (58) that the BERs depend only on the ratio λ_j/λ, but not on the specific value of λ_j or λ. Therefore, this case does not result in the BER being changed. In the second case, the simultaneous change of chip duration T_c in both transmitter and receiver leads to a change in the ratio λ_j/λ, and thus the BER of the system also varies. Figure 8a, b respectively show the dependence of the BERs on λ of the two systems for the second case. The simulation results point out that the system outperforms when λ is increased. For example, in the case $L = 4$, the BER values of the CDS-CDMA and DCSK systems respectively decrease from $1.2 \cdot 10^{-3}$ and $6.2 \cdot 10^{-3}$ to $2.9 \cdot 10^{-5}$ and $6.3 \cdot 10^{-4}$, corresponding to λ increasing from 10 to 50. It can be seen that the estimated results agree completely with the simulated result except for the cases $L = 3, 4$ with $\lambda = 10$. The main reason of this mismatch is that the simulation parameters chosen for these cases, i.e., $\tau_3 = 15\tau$, $\tau_4 = 20\tau > T_c = \lambda\tau = 10\tau$, are incorrect with the channel condition in theory.

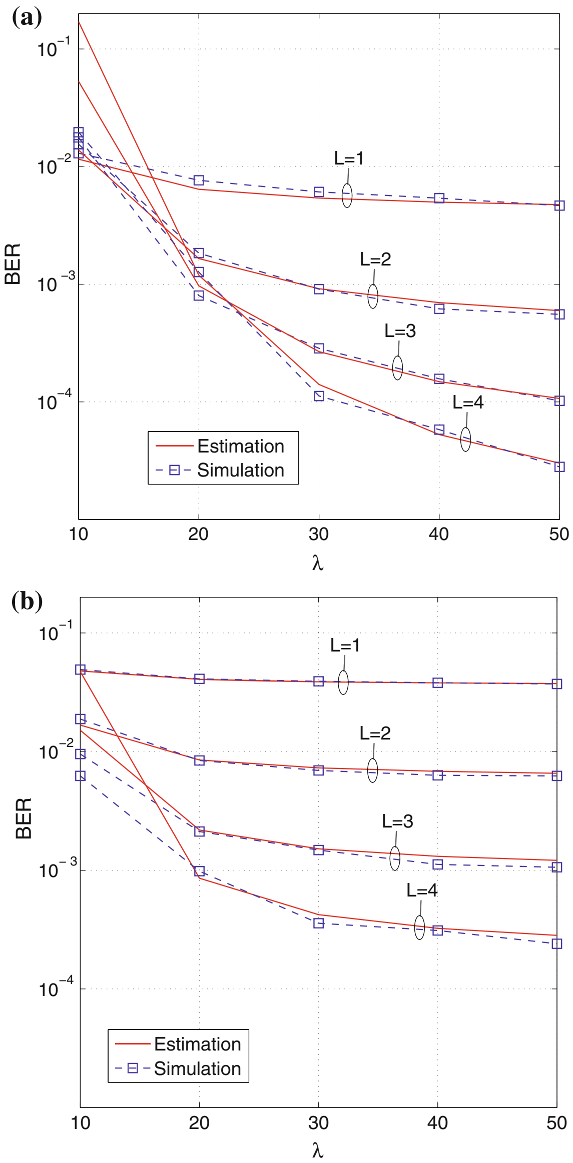
Fig. 7 BER values against the spreading factor 2β: **a** CDS-CDMA system with an increment of E_b/N_0 in the case of $K = 1, L = 2$, $\lambda = 30$, and **b** DCSK system with an increment of the number of secondary paths L in the case of $E_b/N_0 = 10$ and $\lambda = 30$



7 Conclusion

The performance of chaos-based communication systems over wireless channels in the context of low-rate transmission was investigated in this chapter. The mathematical models in the discrete time domain for the wireless channel, transmitter, and receiver of the two most typical systems, i.e., CDS-CDMA and DCSK, were

Fig. 8 BER values against the number of samples per chip λ with an increment of the number of secondary paths L: **a** CDS-CDMA system in the case of $K = 1$, $2\beta = 64$, $E_b/N_0 = 6dB$ **b** DCSK system in the case of $2\beta = 64$ and $E_b/N_0 = 10$



described. Based on the theoretical BER expressions and distribution histograms obtained, the BER performances of these systems were estimated by means of numerical integration. The simulations agree with the estimates. It can be seen from the obtained results that first, the chaos-based communication systems can perform well in the studied wireless channel; especially, the BER performances are significantly enhanced when the number of secondary paths increases. This means that the systems can exploit the low-rate transmission characteristic of chaotically spreading signals to improve their performance. Second, it is observed that owing to the requirement of chaotic sequence synchronization, the structure of the CDS-CDMA receiver is more complicated than that of the DCSK receiver. In return, our comparison showed that under the same channel condition and system parameters, the CDS-CDMA system outperforms the DCSK, especially at a higher number of secondary paths. Last but not least, since the transmitter and receiver of the two systems can perform the spreading and despreading processes based on discrete-sample processing in the time domain, it is suitable to be implemented on high-speed programmable ICs such as FPGA and DSP. All blocks in the proposed schemes can be implemented by available or created modules on the same chip. For example, the sampler at the input can be done by an ADC module, in which the received signal is sampled and converted into binary values stored in registers. The delay block can be realized by a counter or a timer. The multiplication, summation, and threshold comparison of discrete samples are then done by corresponding calculations with respect to the binary values in the registers. Besides the simple structure of the systems, the low-rate transmission also does not require too high a hardware processing speed. Therefore, the low-rate CDS-CDMA and DCSK communication systems can achieve low power consumption, low cost, and small size. The aforementioned features make chaos-based communication systems promising and robust in the design of a secure physical layer for wireless applications with low rate, low complexity, and small size for communications between nodes in LR-WPANs.

Annex: List of Symbols

Symbol	Definition
$\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}, \overline{F}, \overline{G}$	Values of A, B, C, D, E, F, G when α_j is replaced by $\overline{\alpha_j}$
A, B	Variables in the expression of $s_l^{(i)}$
$BER_{CDS-CDMA}$	Static bit-error rate of CDS-CDMA system
$BER_{CDS-CDMA}(\gamma_{CDS-CDMA})$	Dynamic bit-error rate of CDS-CDMA system
BER_{DCSK}	Static bit-error rate of DCSK system
$BER_{DCSK}(\gamma_{DCSK})$	Dynamic bit-error-rate of DCSK system
C, D, E, F, G	Variables in the expression of s_l
E_b	Energy per bit
E_b/N_0	Energy per bit to noise power spectral density ratio
E_c	Energy per chip
E_{4c}	Constant defined by $E\left[{x_{k,l}^{(i)}}^4\right] = E[x_{k,l}^4]$
$E[\cdot]$	Mean value function
$E[(\cdot)^2]$	Squared mean value function
I	Interference component between different CDS-CDMA users
K	Number of users in CDS-CDMA system
L	Number of secondary paths of wireless channel
N	Number classes of distribution histogram

Symbol	Definition
N_0	Noise power spectral density
$P(\gamma_{CDS-CDMA})$	Probability of having the energy in interval centered on $\gamma_{CDS-CDMA,m}$
$P(\gamma_{DCSK})$	Probability of having the energy in interval centered on $\gamma_{DCSK,m}$
$Q(\cdot)$	Q function
S	Component of the beneficial signal in CDS-CDMA receiver
T, U	Components of the signal $s_l^{(i)}$
T_b	Bit duration
T_c	Chip duration
V	Noise component created by the AWGN in CDS-CDMA receiver
W	Component of the beneficial signal in DCSK receiver
X	Interference component between chips in DCSK receiver
Y	Noise component created by the impact of AWGN in DCSK receiver
Z	Noise component created by AWGN in DCSK receiver
$b_l, b_l^{(i)}$	Binary value of lth bit
$c_{k,l}^{(i)}$	Output signal of ith user in kth chip duration in CDS-CDMA transmitter
$d_{g,k,l}, d_{g,k,l}^{(i)}$	Value of samples at output of the multiplier in the receivers
$e_{k,l}$	Value of kth chip in lth bit in the transmitted signal
$e_{k-\tau_j,l}$	Delayed version of $e_{k,l}$
$h(n)$	Impulse response of the channel
$r_{k,l}$	Input signal of the receiver
$r_{g,k,l}, r_{g,k,l}^{(i)}$	Input signal of the correlator in the receivers
$s_l, s_l^{(i)}$	Decision variable in the receivers
$x_{k,l}$	Value of kth chip in lth bit in the reference sequence
$x_{k-\beta,l}$	Delayed version of $x_{k,l}$
α_i	Fading coefficients of Rayleigh distribution
2β	Spreading factor of the systems
$\gamma_{CDS-CDMA}, \phi_{CDS-CDMA}, \psi_{CDS-CDMA}$	Elements of the static BER expression of CDS-CDMA system
$\overline{\gamma}_{CDS-CDMA}, \overline{\phi}_{CDS-CDMA}, \overline{\psi}_{CDS-CDMA}$	Values of $\gamma_{CDS-CDMA}, \phi_{CDS-CDMA}, \psi_{CDS-CDMA}$ on replacing α_j by $\overline{\alpha_j}$
$\gamma_{DCSK}, \phi_{DCSK}, \psi_{DCSK}$	Elements of the static BER expression of DCSK system
$\overline{\gamma}_{DCSK}, \overline{\phi}_{DCSK}, \overline{\psi}_{DCSK}$	Values of $\gamma_{DCSK}, \phi_{DCSK}, \psi_{DCSK}$ on replacing α_j by $\overline{\alpha_j}$
$\delta(t)$	Dirac impulse
$\eta_{k,l}$	Additive white Gaussian noise (AWGN)
λ	Number of samples per chip
λ_j	Number of samples in duration τ_j
τ	Sampling cycle of the receivers
τ_j	Delay of the ith secondary path

Appendix

A. Expression for the output signal $s_l^{(i)}$ of the CDS-CDMA correlator:

$$
\begin{aligned}
s_l^{(i)} &= \sum_{g=1}^{2\lambda\beta} d_{g,k,l}^{(i)} = \sum_{k=1}^{2\beta}\sum_{g=1}^{\lambda} d_{g,k,l}^{(i)} = \sum_{k=1}^{2\beta}\sum_{g=1}^{\lambda} r_{g,k,l} x_{k,l}^{(i)} \\
&= \sum_{k=1}^{2\beta}\Bigg(\lambda\alpha_0 x_{k,l}^{(i)} \sum_{i=1}^{K} b_l^{(i)} x_{k,l}^{(i)} + (\lambda-\tau_1)\alpha_1 x_{k,l}^{(i)} \sum_{i=1}^{K} b_l^{(i)} x_{k,l}^{(i)} + \cdots + (\lambda-\tau_L)\alpha_L x_{k,l}^{(i)} \sum_{i=1}^{K} b_l^{(i)} x_{k,l}^{(i)} \\
&\quad + \tau_1\alpha_1 x_{k,l}^{(i)} \sum_{i=1}^{K} b_l^{(i)} x_{k-1,l}^{(i)} + \cdots + \tau_L\alpha_L x_{k,l}^{(i)} \sum_{i=1}^{K} b_l^{(i)} x_{k-1,l}^{(i)} + \lambda\eta_{k,l} x_{k,l}^{(i)}\Bigg) \\
&= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)\Bigg(\sum_{k=1}^{2\beta} b_l^{(i)} x_{k,l}^{(i)\,2} + \sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)}\Bigg) + \Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg)\Bigg(-\sum_{k=1}^{2\beta} b_l^{(i)} x_{k,l}^{(i)\,2} \\
&\quad - \sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)} + \sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k-1,l}^{(m)} x_{k,l}^{(i)}\Bigg) + \lambda\sum_{k=1}^{2\beta}\eta_{k,l} x_{k,l}^{(i)}
\end{aligned}
\tag{72}
$$

B. Statistical calculation for the mean and mean squared values of T, U, V:

$$
\begin{aligned}
E[T|b_l^{(i)} = +1] &= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)\Bigg(E\Bigg[\sum_{k=1}^{2\beta} x_{k,l}^{(i)\,2}\Bigg] + E\Bigg[\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)}\Bigg]\Bigg) \\
&= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)\sum_{k=1}^{2\beta} E\left[x_{k,l}^{(i)\,2}\right] = \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg) 2\beta E_c,
\end{aligned}
\tag{73}
$$

$$
\begin{aligned}
E[U|b_l^{(i)} = +1] &= \Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg)\Bigg(-E\Bigg[\sum_{k=1}^{2\beta} x_{k,l}^{(i)\,2}\Bigg] - E\Bigg[\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k,l}^{(m)} x_{k,l}^{(i)}\Bigg] \\
&\quad + E\Bigg[\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)} x_{k-1,l}^{(m)} x_{k,l}^{(i)}\Bigg]\Bigg) \\
&= \Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg)\Bigg(-\sum_{k=1}^{2\beta} E\left[x_{k,l}^{(i)\,2}\right]\Bigg) = -\Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg) 2\beta E_c,
\end{aligned}
\tag{74}
$$

and

$$
\begin{aligned}
E[T^2|b_l^{(i)}=+1] &= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)^2 \Bigg(E\Bigg[\Bigg(\sum_{k=1}^{2\beta}{x_{k,l}^{(i)}}^2\Bigg)^2\Bigg] + 2E\Bigg[\sum_{k=1}^{2\beta}{x_{k,l}^{(i)}}^2\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)}x_{k,l}^{(m)}x_{k,l}^{(i)}\Bigg]\\
&\quad +E\Bigg[\Bigg(\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)}x_{k,l}^{(m)}x_{k,l}^{(i)}\Bigg)^2\Bigg]\Bigg)\\
&= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)^2 \Bigg(\sum_{k=1}^{2\beta}E\Big[{x_{k,l}^{(i)}}^4\Big] + \sum_{k=1}^{2\beta-1}E\Big[{x_{k,l}^{(i)}}^2\Big]\sum_{p=1}^{2\beta}E\Big[{x_{p,l}^{(i)}}^2\Big]\\
&\quad +\sum_{k=1}^{2\beta}E\Big[{x_{k,l}^{(i)}}^2\Big]\sum_{m=1,m\neq i}^{K}E\Big[{x_{k,l}^{(m)}}^2\Big]\Bigg)\\
&= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)^2\Big(2\beta E_{c4} + 2\beta(2\beta-1)E_c^2 + 2\beta(K-1)E_c^2\Big)\\
&= \Bigg(\sum_{j=0}^{L}\lambda\alpha_j\Bigg)^2\Big(2\beta E_{c4} + 2\beta(2\beta+K-2)E_c^2\Big).
\end{aligned}
\tag{75}
$$

$$
\begin{aligned}
E[U^2|b_l^{(i)}=+1] &= \Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg)^2 \Bigg(E\Bigg[\Bigg(\sum_{k=1}^{2\beta}x_{k,l}^{(i)}\Bigg)^2\Bigg] + E\Bigg[\Bigg(\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)}x_{k,l}^{(m)}x_{k,l}^{(i)}\Bigg)^2\Bigg]\\
&\quad +E\Bigg[\Bigg(\sum_{k=1}^{2\beta}\sum_{m=1,m\neq i}^{K} b_l^{(m)}x_{k-1,l}^{(m)}x_{k,l}^{(i)}\Bigg)^2\Bigg]\Bigg)\\
&= \Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg)^2\Big(2\beta E_{c4} + 2\beta(2\beta-1)E_c^2 + 2\beta(K-1)E_c^2 + 2\beta K E_c^2\Big)\\
&= \Bigg(\sum_{j=1}^{L}\tau_j\alpha_j\Bigg)^2\Big(2\beta E_{c4} + 2\beta(2\beta+2K-2)E_c^2\Big).
\end{aligned}
\tag{76}
$$

$$
E[V^2|b_l^{(i)}=+1] = \lambda^2 E\Bigg[\Bigg(\sum_{k=1}^{2\beta}\eta_{k,l}x_{k,l}^{(i)}\Bigg)^2\Bigg] = \lambda^2\sum_{k=1}^{2\beta}E\big[\eta_{k,l}^2\big]E\Big[{x_{k,l}^{(i)}}^2\Big] = \lambda^2 E_b\frac{N_0}{2}. \tag{77}
$$

C. Statistical mean and mean squared values of the components W, X, Y, Z:

$$
E[W|b_l=+1] = A\sum_{k=1}^{\beta}E[x_{k-\beta,l}^2] + B\sum_{k=1}^{\beta}E[x_{(k-1)-\beta),l}^2] = (A+B)\beta E_c, \tag{78}
$$

$$
E[X|b_l=+1] = C\sum_{k=1}^{\beta}E[x_{k-\beta,l}x_{(k-1)-\beta),l}] = 0, \tag{79}
$$

$$\begin{aligned}E[Y|b_l = +1] &= D\sum_{k=1}^{\beta} E[x_{k-\beta,l}\eta_{k,l}] + D\sum_{k=1}^{\beta} E[x_{k-\beta,l}\eta_{k-\beta,l}]\\ &+ E\sum_{k=1}^{\beta} E[x_{(k-1)-\beta,l}\eta_{k,l}] + E\sum_{k=1}^{\beta} E[x_{(k-1)-\beta,l}\eta_{k-\beta,l}] = 0,\end{aligned} \tag{80}$$

$$E[Z|b_l = +1] = \lambda\sum_{k=1}^{\beta} E[\eta_{k,l}\eta_{k-\beta,l}] = 0, \tag{81}$$

and

$$\begin{aligned}E[W^2|b_l^{(i)} = +1] &= A^2\sum_{k=1}^{\beta} E[x_{k-\beta,l}^4] + A^2\sum_{k=1}^{\beta} E[x_{k-\beta,l}^2]\sum_{m=1}^{\beta-1} E[x_{m-\beta,l}^2]\\ &+ B^2\sum_{k=1}^{\beta} E[x_{(k-1)-\beta),l}^2] + B^2\sum_{k=1}^{\beta} E[x_{(k-1)-\beta,l}^2]\sum_{m=1}^{\beta-1} E[x_{(m-1)-\beta,l}^2]\\ &+ 2AB\sum_{k=1}^{\beta} E[x_{k-\beta,l}^2]\sum_{k=1}^{\beta} E[x_{(k-1)-\beta,l}^2]\\ &= (A^2+B^2)\beta E_{c4} + (A^2+B^2)\beta(\beta-1)E_c^2 + 2AB\beta^2E_c^2\\ &= (A^2+B^2)\beta(E_{c4} + (\beta-1)E_c^2) + 2AB\beta^2E_c^2,\end{aligned} \tag{82}$$

$$\begin{aligned}E[X^2|b_l = +1] &= C^2\sum_{k=1}^{\beta} E[x_{k-\beta,l}^2]E[x_{(k-1)-\beta,l}^2] + C^2\sum_{k=1}^{\beta} E[x_{k-\beta,l}x_{(k-1)-\beta,l}]\\ &\times\sum_{m=1}^{\beta-1} E[x_{m-\beta,l}x_{(m-1)-\beta,l}] = C^2\beta E_c^2,\end{aligned} \tag{83}$$

$$\begin{aligned}E[Y^2|b_l = +1] &= D^2\sum_{k=1}^{\beta} E[x_{k-\beta,l}^2]E[\eta_{k,l}^2] + D^2\sum_{k=1}^{\beta} E[x_{k-\beta,l}^2]E[\eta_{k-\beta,l}^2]\\ &+ E^2\sum_{k=1}^{\beta} E[x_{(k-1)-\beta,l}^2]E[\eta_{k,l}^2] + E^2\sum_{k=1}^{\beta} E[x_{(k-1)-\beta,l}^2]E[\eta_{k-\beta,l}^2]\\ &= (D^2+E^2)\beta N_0E_c,\end{aligned} \tag{84}$$

$$E[Z^2|b_l = +1] = \lambda^2\sum_{k=1}^{\beta} E[\eta_{k,l}^2]E[\eta_{k-\beta,l}^2] = \lambda^2\beta\frac{N_0^2}{4}. \tag{85}$$

References

1. Lau, F.C.M., Tse, C.K.: Chaos-based Digital Communication Systems: Operating Principles, Analysis Methods, and Performance Evaluation. Springer, Berlin (2003)
2. Stavroulakis, P.: Chaos Applications in Telecommunications. CRC Press, Boca Raton (2005)
3. Quyen, N.X., Quyet, B.T., Yem, V.V., Dzung, N.T., Hoang, T.M.: Simulation and implementation of improved chaotic Colpitts circuit for UWB communications. In: Proceedings of the International Conference on Communications and Electronics (ICCE'10), Nha Trang-Vietnam, pp. 307–312 (2010)
4. Yu, J., Yao, Y.-D.: Detection performance of chaotic spreading LPI waveforms. IEEE Trans. Wirel. Commun. **4**(2), 390–396 (2005)
5. Heidari-Bateni, G., McGillem, C.D.: A chaotic direct-sequence spread-spectrum communication system. IEEE Trans. Commun. **42**(234), 1524–1527 (1994)
6. Cong, L., Shaoquian, L.: Chaotic spreading sequences with multiple access performance better than random sequence. IEEE Trans. Circuits. Syst. I **47**(3), 394–397 (2000)
7. Kolumban, G., Kis, G., Jako, Z., Kennedy, M.P.: FM-DCSK: a robust modulation scheme for chaotic communications. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E81**(A(9)), 1798–1802 (1998)
8. Mazzini, G., Setti, G., Rovatti, R.: Chaotic complex spreading sequences for asynchronous DS-CDMA, I, system modeling and results. IEEE Trans. Circ. Syst. I **44**(10), 937–947 (1997)
9. Kolumban, G., Vizvari, B., Schwarz, W., Abel, A.: Differential chaos shift keying: a robust code for chaos communication. In: Proceedings of the 4th International Workshop on Nonlinear Dynamics Electronic System, pp. 87–92 (1996)
10. Lau, F.C.M., Yip, M.M., Tse, C.K., Hau, S.F.: A multiple-access technique for differential chaos shift keying. IEEE Trans. Circuits Syst. **49**(1), 96–104 (2002)
11. Mandal, S., Banerjee, S.: Analysis and CMOS implementation of a chaos-based communication system. IEEE Trans. Circuits Syst. I **51**(9), 1708–1722 (2004)
12. Delgado-Restituto, M., Acosta, A.J., Rodriguez-Vazquez, A.: A mixed-signal integrated circuit for FM-DCSK modulation. IEEE J. Solid-State Circuits **40**(7), 1460–1471 (2005)
13. Chen, C., Yao, K., Umeno, K., Biglieri, E.: Design of spread-spectrum sequences using chaotic dynamical systems and ergodic theory. IEEE Trans. Circuits Syst. **48**, 1110–1114 (2001)
14. Rovatti, R., Setti, G., Mazzini, G.: Toward sequence optimization for chaos-based asynchronous DS-CDMA systems. In: Proceedings of the IEEE GLOBECOM Sydney, pp. 2174–2179 (1998)
15. Setti, G., Rovatti, R., Mazzini, G.: Synchronization mechanism and optimization of spreading sequences in Chaos-based DS-CDMA systems. IEICE Trans. Fundam. Elec. Commun. Comput. Sci. **E82–A**, 1737–1746 (1999)
16. Jovic, B., Unsworth, C., Sandhu, G., Berber, S.: A robust sequence synchronization unit for multi-user DS-CDMA chaos-based communication systems. Sig. Process. **87**, 1692–1708 (2007)
17. Kaddoum, G., Roviras, D., Charge, P., Fournier-Prunaret, D.: Robust synchronization for asynchronous multi-user chaos-based DS-CDMA. Sig. Process. **89**, 807–818 (2009)
18. Vali, R., Berber, S., Nguang, S.: Effect of Rayleigh fading on noncoherent sequence synchronization for multi-user chaos based DS-CDMA. Sig. Process. **90**, 1924–1939 (2010)
19. Quyen, N.X., Cong, L.V., Long, N.H., Yem, V.V.: An OFDM-based chaotic DSSS communication system with MPSK modulation. In: Proceedings of the International Conference on Communications and Electronics (ICCE'14), Danang-Vietnam, pp. 106–111 (2014)
20. Quyen, N.X., Duong, T.Q., Vo, N.S., Xie, Q., Shu, L.: Chaotic direct-sequence spread-spectrum with variable symbol period: a technique for enhancing physical layer security. Comput. Netw. (2016). doi:10.1016/j.comnet.2016.06.022
21. Vitali, S., Rovatti, R., Setti, G.: On the performance of chaos-based multicode DS-CDMA systems. Circuits, Syst. Signal Process. **24**, 475–495 (2005)
22. Quyen. N.X., Nguyen, C.T., Pere, B-R., Reiner, D.: A novel approach to security enhancement of chaotic DSSS systems. In: Proceedings of the International Conference on Communications and Electronics (ICCE'16), Halong-Vietnam, pp. 471–476 (2016)

23. Tam, W., Lau, F., Tse, C., Lawrance, A.: Exact analytical bit error rates for multiple access chaos-based communication systems. IEEE Trans. Circuits Syst. **II**(51), 473–481 (2004)
24. Kaddoum, G., Charge, P., Roviras, D., Fournier-Prunaret, D.: A methodology for bit error rate prediction in chaos-based communication systems. Circuits Syst. Signal Process. **28**, 925–944 (2009)
25. Kaddoum, G., Charge, P., Roviras, D.: A generalized methodology for bit-error-rate prediction in correlation-based communication schemes using chaos. IEEE Comm. Lett. **13**, 567–569 (2009)
26. Rovatti, R., Mazzini, G., Setti, G.: Enhanced rake receivers for chaos-based DS-CDMA. IEEE Trans. Circuits Syst. **I**(48), 818–829 (2001)
27. Mazzini, G., Rovatti, R., Setti, G.: Chaos-based asynchronous DS-CDMA systems and enhanced rake receivers: measuring the improvements. IEEE Trans. Circuits Syst. **I**(48), 1445–1453 (2001)
28. Kaddoum, G., Roviras, D., Charge, P., Fournier-Prunaret, D.: Accurate bit error rate calculation for asynchronous chaos-based DS-CDMA over multipath channel. EURASIP J. Adv. Signal Process. **2009**(571307), 12 (2009)
29. Kaddoum, G., Coulon, M., Roviras, D., Charge, P.: Theoretical performance for asynchronous multi-user chaos-based communication systems on fading channels. Elsevier Signal Process. **90**, 2923–2933 (2010)
30. Berber, S.M.: Probability of error derivatives for binary and chaos-based CDMA systems in wide-band channels. IEEE Trans. Wireless Comm. **13**, 5596–5606 (2014)
31. Quyen, N.X., Yem, V.V., Hoang, T.M., Kyamakya, K.: MxN-ary chaotic pulse width position modulation: an effective combination method for improving bit rate. Int. J. Comput. Math. Electr. Electron. Eng. (COMPEL) **32**(3), 776–793 (2013)
32. Quyen, N.X., Yem, V.V., Hoang, T.M.: A chaos-based direct-sequence/spread-spectrum communication scheme. In: Proceedings of the International Symposium on Theoretical Electrical Engineering (ISTET'13), III-1/III-2 (2013)
33. Quyen, N.X., Yem, V.V., Duong, T.Q.: Design and analysis of a spread-spectrum communication system with chaos-based variation of both phase-coded carrier and spreading factor. IET Commun. **9**(12), 1466–1473 (2015)
34. Sushchik, M., Tsimring, L.S., Volkovskii, A.R.: Performance analysis of correlation-based communication schemes utilizing chaos. IEEE Trans. Circuits Syst. **47**, 1684–1691 (2000)
35. Kaddoum, G., Charge, P., Roviras, D., Fournier-Prunaret, D.: Performance analysis of differential Chaos shift keying over an AWGN channel. ACTEA **2009**, 255–358 (2009)
36. Quyen, N.X.: An oversampling-based correlator-type receiver for DCSK communication systems over generalized flat rayleigh fading channels. REV J. Electron. Commun. **6**(1–2), 1–12 (2016)
37. Xia, Y., Tse, C.K., Lau, F.C.M.: Performance of differential chaos-shift-keying digital communication systems over a multipath fading channel with delay spread. IEEE Trans. Circuits Syst. II, Express Briefs **51**, 680–684 (2004)
38. Quyen, N.X., Duong, T.Q., Nallanathan, A.: Modeling, analysis and performance comparison of two direct sampling DCSK receivers under frequency nonselective fading channels. IET Commun. **10**(11), 1466–1473 (2016)
39. Zhou, Z., Zhou, T., Wang, J.: Performance of multiple access DCSK communication over multipath fading channel with delay spread. Circuits, Syst. Signal Process. **27**, 507–518 (2008)
40. Chong, C.C., Yong, S.K.: UWB direct chaotic communication technology for low-rate WPAN applications. IEEE Trans. Veh. Technol. **57**(3), 1527–1536 (2008)
41. Kolumban, G., Kis, G., Jako, Z., Kennedy, M.P.: FM-DCSK: A robust modulation scheme for chaotic communications. IEICE Trans. Fund. Electron. Comm. Comput. Sci. **E81–A**(9), 1798–1802 (1998)
42. Lau, F.C.M., Cheong, K.Y., Tse, C.K.: Permutation-based DCSK and multiple access DCSK systems. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **50**(6), 733–742 (2003)
43. Yang, H., Jiang, G.P.: Reference-modulated DCSK: a novel chaotic communication scheme. IEEE Trans. Circuits Syst. II **60**(4), 232–236 (2013)

44. Kaddoum, G., Gagnon, F.: Design of a high-data-rate differential chaos-shift keying system. IEEE Trans. Circuits Syst. II Express Briefs **57**(9), 448–452 (2012)
45. Yang, H., Jiang, G.P.: High-efficiency differential-chaos-shift keying scheme for chaos-based noncoherent communication. IEEE Trans. Circuits Syst. II: Express Briefs **59**(5), 312–316 (2012)
46. Kaddoum, G., Richardson, F.D., Gagnon, F.: Design and analysis of a multi-carrier differential chaos shift keying communication system. EEE Trans. Commun. **61**(8), 3281–3291 (2013)
47. Fang, Y., Wang, L., Chen, P., Xu, J., Chen, G., Xu, W.: Design and analysis of a DCSK-ARQ/CARQ system over multipath fading channels. IEEE Trans. Circuits Syst. I: Reg. Papers **62**(6), 1637–1647 (2015)
48. Kaddoum, G., Soujeri, E., Arcila, C., Eshteiwi, K.: I-DCSK: an improved non-coherent communication system architecture. IEEE Trans. Circuits Syst. II: Express Briefs **62**(9), 901–905 (2015)
49. Wang, L., Min, X., Chen, G.: Performance of SIMO FM-DCSK UWB system based on chaotic pulse cluster signals. IEEE Trans. Circuits Syst. I **58**(9), 2259–2268 (2011)
50. Wang, S., Wang, X.: M-DCSK-based chaotic communications in MIMO multipath channels with no channel state information. IEEE Trans. Circuits Syst. II Express Briefs **57**(12), 1001–1005 (2010)
51. Fang, Y., Xu, J., Wang, L., Chen, G.: Performance of MIMO relay DCSK-CD systems over Nakagami fading channels. IEEE Trans. Circuits Syst. I **60**(3), 757–767 (2013)
52. Xu, W., Wang, L., Chen, G.: Performance of DCSK cooperative communication systems over multipath fading channels. IEEE Trans. Circuits Syst. I **58**(1), 196–204 (2011)
53. Prasad, R., Munoz, L.: WLANs and WPANs towards 4G Wireless. Artech House, Boston (2003)
54. Zheng, J., Lee, M.J., Anshel, M.: Toward secure low rate wireless personal area networks. IEEE Trans. Mobile Comput. **5**(10), 1361–1373 (2006)
55. IEEE 802.11: Wireless LANs. http://standards.ieee.org (2013)
56. Zheng, J., M.J. Lee, M.J.: Will IEEE 802.15.4 make ubiquitous networking a reality?-A discussion on a potential low power, low bit rate standard. IEEE Comm. Mag. **42**(6), 140–146 (2004)
57. Bisdikian, C.: An overview of the Bluetooth wireless technology. IEEE Commun. Mag. **39**(12), 86–94 (2001)
58. IEEE Standard for Local and Metropolitan Area Networks-Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs). IEEE Standard 802.15.4-2011(Revision of IEEE Std 802.15.4-2006. IEEE: New York, NY, USA, 2011, pp. 1-294
59. Wheeler, A.: Commercial applications of wireless sensor networks using ZigBee. IEEE Commun. Mag. **45**(4), 70–77 (2007)
60. Siddiqui, M.M.: Statistical inference for Rayleigh distributions. J. Res. Natl. Bur. Stand. **68D**, 1007 (1964)
61. Grinstead, C.M., Snell, J.L.: Central Limit Theorem, Introduction to Probability, 2nd edn. American Mathematical Society, Providence (1997)
62. Isabelle, S.H., Wornell, G.W.: Statistical analysis and spectral estimation techniques for one-dimensional chaotic signals. IEEE Trans. Signal Process. **45**, 1495–1497 (1997)
63. Geisel, T., Fairen, V.: Statistical properties of chaos in Chebyshev maps. Phys. Lett. **105A**(6), 263–266 (1984)

Nonlinear Programming Approach for Design of High Performance Sigma–Delta Modulators

Valeri Mladenov and Georgi Tsenov

Abstract In this chapter we present a nonlinear programming approach to the design of third-order sigma–delta modulators with respect to maximization of the signal-to-noise ratio, taking into account the modulator's stability. The proposed approach uses an analytic formula for calculation of the signal-to-noise ratio and an analytic formula for stability of the modulator. Thus the goal function becomes maximization of the signal-to-noise ratio and constraints come from stability issues and bounds of the modulator noise transfer function coefficients. The results are compared with the optimal third-order modulator design provided by DStoolbox. The proposed procedure has low computation requirements. It is described for third-order modulators with one real pole of the loop filter transfer function and can be extended easily and generalized to higher-order modulators.

Keywords Sigma-delta modulators · Digital signal processing · Stability · Analog-to-digital conversion · Signal-to-noise ratio

1 Introduction

The basic SDM structure consists of a filter with transfer function followed usually for implementation simplicity by a one-bit quantizer in a feedback loop. For sigma–delta modulators (SDM) with single-bit quantization levels, it is usually difficult to design high-order sigma–delta modulator loop filter transfer functions that both are stable and provide high signal-to-noise ratio levels [1–3]. This is caused by the nonlinearity of the modulator, i.e., SDMs are nonlinear systems, due to the quantizer utilized. Nowadays, sigma–delta modulators have become the standard for analog-

V. Mladenov (✉) · G. Tsenov
Department of Theoretical Electrical Engineering, Technical University of Sofia,
8 Kliment Ohridski St., 1000 Sofia, Bulgaria
e-mail: valerim@tu-sofia.bg

G. Tsenov
e-mail: gogotzenov@tu-sofia.bg

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_12

to-digital conversion. SDMs can achieve a very high signal-to-noise ratio (SNR) even with a small number of quantization levels, and this is achieved using very high oversampling ratios. With the feedback loop and the transfer function, SDMs shape the noise and push it to frequencies higher than the operational band of interest. Thanks to its simplicity, single-bit code-shaping SDMs are of great interest, because their performance is influenced mostly by the loop filter transfer function and the modulator's oversampling ratio (OSR), and the modulator output is encoded into a bitstream. Until recent years, the modulator's maximal stable DC input signal range and its SNR were determined mostly with the use of simulations, which left a zone of uncertainty. Many engineers experimented with the loop filter transfer function coefficients in order to achieve higher SNR while keeping the modulator stable [3–9]. The realistic high-performance loop filter transfer functions have the poles grouped into complex conjugate pairs and one real pole when the modulator order is odd, and complex conjugate pairs for even loop filter orders. In order to increase modulator performance, some authors move one of the complex conjugate pairs of poles or the real pole a little bit outside the unit circle while keeping the other poles inside, resulting in increased SNR and reduced stability limit for maximal DC input signal amplitude beyond which the modulator becomes unstable. Moving a single pole or complex-conjugate pair a little bit outside the unit circle does not necessarily makes the SDM unstable, since it is a nonlinear system, which makes the SDM behavior analysis harder for those cases. In recent years, approximation formulas for determining both the SNR and the maximum stable DC input levels based on the loop filter transfer function coefficients and the input variables without the need for simulation models have been published. Using a parallel decomposition form of the loop filter allows approximation of the maximal stable DC input signal value without the need for simulations for single-bit quantizer modulators, while SNR calculation from a derivation of the loop filter noise transfer function also provides no need of the modulator's output bitstream to obtain it, resulting in no need for SDM simulations to determine both stability and performance. This makes possible the design of optimal loop filter transfer functions with respect to maximizing the SNR while keeping the modulator stable with nonlinear optimization procedure as the modulator loop filter transfer function coefficients and their respective ranges set as constraints. We are presenting such a design approach for a higher odd-order SDM, taking into account SDM stability and SNR performance. The procedure is backed up with examples and results made for third-order loop filter sigma–delta modulators, and it gives the performance impact when the poles of the noise transfer function are varied when optimized zeros are used. This loop filter function design is computed with fast theoretical calculation of the signal-to-noise ratio with a mathematical formula, instead of an approximation based on simulations and combined with theory that presents an approximate value for the modulator's maximal stable DC input signal, resulting in a design without the need for simulations of the modulator's output bitstream.

The chapter is organized as follows. We provide the theoretical background necessary for understanding the approach in the following section. The formulation of the nonlinear programming procedure is given in the third section. Results and comparisons are presented in the fourth section, followed by a conclusion and remarks given in the last section.

2 Theoretical Background

In order to better understand the design approach, the results of [10] will be briefly recalled as they are used. The basic structure of an SDM, shown in Fig. 1, comprises a filter with a transfer function $G(z)$. A one-bit quantizer in a feedback loop follows it in the SDM structure.

The system works in discrete time, and a discrete-time sequence $u(n) \in [-1, 1]$ is the input to the loop. It appears in quantized form at the output. The output of the filter is the discrete-time sequence $x(n)$. The same sequence is applied to the quantizer input. When the input is a positive quantity, the quantizer produces an output of $+1$, and when its input is negative, the quantizer output is˘1 (single-bit). In this case, the quantizer will not provide a good approximation to its input signal: that is why a feedback loop has been used. In this way, the quantization noise is shifted away from a specified frequency band. If a given input signal that lies in this frequency band is applied to the loop, then a great deal of the noise due to the quantization process will appear outside the frequency band of interest. Thus a good approximation to the input signal will be received. This process is called noise shaping.

The stability of the modulator could be acquired without simulation. The authors of [10] examine an Nth-order modulator of the form

$$G(z) = \frac{a_1 z^{-1} + \cdots + a_N z^{-N}}{1 + d_1 z^{-1} + d_2 z^{-2} + \cdots + d_N z^{-N}}. \tag{1}$$

In the most common case, the loop filter transfer function has complex-conjugate roots. Without any loss of generality, only one pair of complex-conjugate roots will be considered. Then Eq. (1) becomes

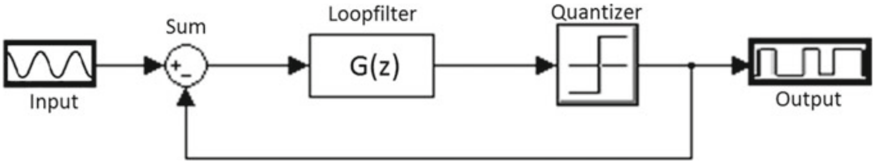


Fig. 1 Basic sigma–delta modulator structure

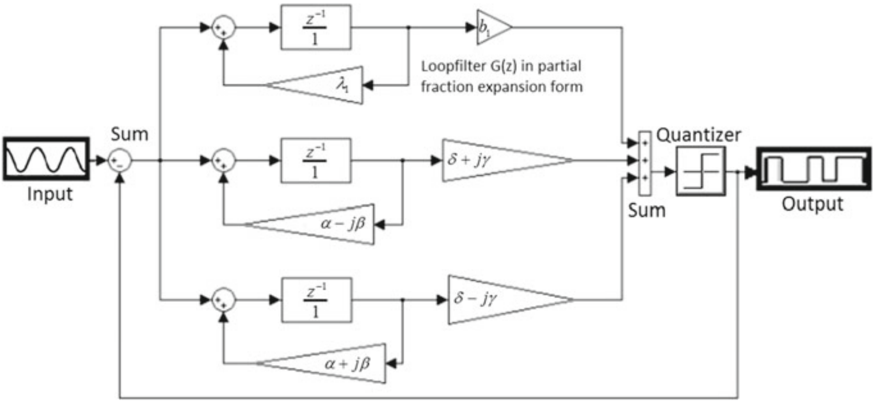


Fig. 2 Block diagram of higher-order SDM with parallel loop filter form

$$G(z) = \frac{b_1 z^{-1}}{1 - \lambda_1 z^{-1}} + \cdots + G_2(z) = \frac{b_1 z^{-1}}{1 - \lambda_1 z^{-1}} + \cdots + \frac{B_{N-1} z^{-1} + B_N z^{-2}}{1 - d_1 z^{-1} - d_2 z^{-2}}. \tag{2}$$

In Eq. (2), the coefficients b_i, $i = 1, 2, \ldots, N$, of the fractional components can be easily found by the well-known expression $b_i = \frac{(1-\lambda_i z^{-1})}{z^{-1}} G(z)\Big|_{z=\lambda_i}$. The denominator of the last term of Eq. (2) has a complex-conjugate pair of roots, and Eq. (2) becomes

$$G(z) = \frac{b_1 z^{-1}}{1 - \lambda_1 z^{-1}} + \cdots + \frac{b_{N-1} z^{-1}}{1 - \lambda_{N-1} z^{-1}} + \frac{b_N z^{-1}}{1 - \lambda_N z^{-1}}, \tag{3}$$

where

$$\begin{aligned} \lambda_{N-1} &= \alpha + j\beta, \;\; \lambda_N = \alpha - j\beta, \\ b_{N-1} &= \delta - j\gamma, \;\; b_N = \delta + j\gamma, \end{aligned} \tag{4}$$

i.e., λ_{N-1}, λ_N, and b_{N-1}, b_N are complex-conjugates.

Due to this, in Eq. (2) a parallel presentation of the third-order modulator is used (see Fig. 2). The values of the last two blocks are complex numbers, but their output signals are real numbers. They correspond to a second-order SDM with complex-conjugate poles of the loop filter transfer function $G(z)$. The signals x_2 and x_3 are complex-conjugates:

$$\begin{aligned} x_2(k+1) &= m(k+1) + jn(k+1), \\ x_3(k+1) &= m(k+1) - jn(k+1). \end{aligned} \tag{5}$$

Due to this, the input of the quantizer is real:

$$(\delta - j\gamma)x_2(k) + (\delta + j\gamma)x_3(k) = 2\delta m(k) + 2\gamma n(k). \tag{6}$$

We can consider the modulator to be three first-order modulators through the quantizer function only. The connected signals with two modulators are complex quantities, while at the same time, the input and output signals (u and y) are the "true" signals of the modulator. As emphasized in [10], both modulators work cooperatively, since their signals are conjugated. The real SDM does not have these modulators: they are introduced to facilitate the behavior of the whole system.

The advantage of this modulator description is the possibility to determine whether the modulator is stable by this criterion [10]:

$$\frac{(2-\lambda_1)}{\lambda_1}\frac{b_1}{(\lambda_1-1)} > -\sum_{i=2}^{N-2}\frac{|b_i|}{\lambda_i-1} + \frac{2|\delta(1-\alpha)+\gamma\beta|}{(1-\alpha)^2+\beta^2}. \tag{7}$$

The maximal range of input signal that ensures the stability expressed by Δu can be determined as well:

$$\Delta u < \frac{\sum_{i=2}^{N-2}\frac{|b_i|}{\lambda_i-1} - \frac{2|\delta(1-\alpha)+\gamma\beta|}{(1-\alpha)^2+\beta^2} + \frac{b_1(2-\lambda_1)}{\lambda_1(\lambda_1-1)}}{\frac{b_1}{\lambda_1-1} - \sum_{i=2}^{N-2}\frac{|b_i|}{\lambda_i-1} + \frac{2|\delta(1-\alpha)+\gamma\beta|}{(1-\alpha)^2+\beta^2}}. \tag{8}$$

The last inequality that will be used for fast signal-to-noise ratio (SNR) calculations is received following the procedure [11]: The quantization theory and the corresponding noise are well established. The distance between two successive quantization levels is called the quantization step size Q. A quantizer with a given number of bits covering the range from $+1$ to -1 needs 2^{bits} quantization levels, and the width of every quantization step is

$$Q = \frac{2}{(2^{bits}-1)}. \tag{9}$$

The quantizer assigns each input sample $u(n)$ to the nearest quantization level. The quantization error is the difference between the input and output to the quantizer, $e_q = y(u) - u$, and is bounded by

$$-\frac{Q}{2} \le e_q(n) \le \frac{Q}{2} \quad . \tag{10}$$

The quantization noise power is given by

$$\sigma_e^2 = \frac{1}{Q}\int_{-\frac{Q}{2}}^{\frac{Q}{2}} e_q^2 de_q = \frac{Q^2}{12} = \frac{1}{3(2^{bits}-1)^2}. \tag{11}$$

Many authors propose that σ_e^2 be approximated with

$$\sigma_e^2 \approx \frac{1}{3\cdot 2^{2bits}}. \tag{12}$$

This quantization error is on the order of one least-significant bit in amplitude, and it is quite small compared to full-amplitude signals.

The average power of a sinusoidal signal of amplitude A, $x(t) = A\cos(2\pi t/T)$, is

$$\sigma_x^2 = \frac{1}{T}\int_0^T (A\cos(2\pi t/T))^2 dt = \frac{A^2}{2}. \tag{13}$$

If the signal is assumed to be oversampled, then instead of acquiring the signal at the Nyquist rate, $2f_B$, the actual sampling rate is $f_s = 2^{r+1} f_B$, and the oversampling ratio is $OSR = 2^r = f_s/2f_B$. In this case, the quantizing noise is spread over a larger frequency range; yet we are still primarily interested in the noise below the Nyquist frequency.

Now, most of the noise power is located outside the signal band. The quantization noise power in the band of interest is decreased by a factor OSR. The signal power occurs over the signal band only. It remains unchanged and is given by Eq. (13).

The linear SDM model for analysis is used by many authors. This model has two inputs: the input signal $X(z)$ and the quantization error $E(z)$. In the basic model shown in Fig. 1, a filter is placed in front of the quantizer, known as the "loop filter," and the output of quantization is fed back and subtracted from the input signal, as shown in Fig. 3.

This may be represented by transfer functions applied to both the input signal and the quantization noise. The Z-domain output may be represented as

$$Y(z) = STF(z)X(z) + NTF(z)E(z), \tag{14}$$

where STF is the *signal transfer function*, and NTF is the *noise transfer function*. The input to the loop filter is $X(z) - E(z)$, so that $Y(z) = G(z)[X(z) - Y(z)] + E(z)$. Rearranging terms, we have:

$$STF(z) = \frac{G(z)}{1 + G(z)}, \quad NTF(z) = \frac{1}{1 + G(z)}. \tag{15}$$

Utilizing the linear model of SDM, the noise shaping in SDM implies a variable noise power in the baseband:

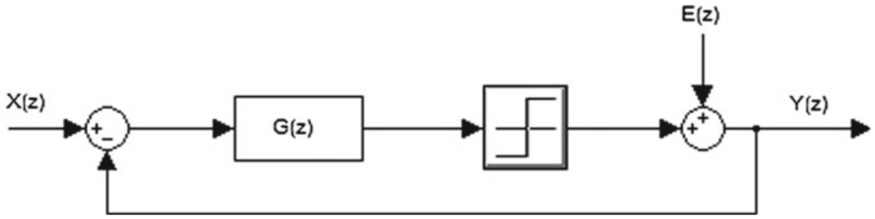


Fig. 3 Representation of a sigma–delta modulator using the linear model

$$\sigma_n^2 = \int_{-f_B}^{f_B} S_e^2(f)|NTF(f)|^2 df, \tag{16}$$

where $S_e^2(f) = \sigma e^2/f_s$ is the power spectral density of the unshaped quantization noise. The total noise power, σe^2, remains unchanged, but an appropriate choice of $NTF(z)$ pushes the noise up to the high frequencies.

By definition, SNR is calculated on the basis of

$$SNR(\text{dB}) = 10\log_{10}\frac{\sigma_x^2}{\sigma_n^2}. \tag{17}$$

We can take the σ_e^2, σ_x^2, and σ_n^2 terms and substitute them to get the general formula for the signal-to-noise ratio of any sigma–delta modulator:

$$\begin{aligned} SNR(\text{dB}) &= 10\log_{10}\tfrac{\sigma_x^2}{\sigma_n^2} = 10\log_{10}\frac{\frac{A^2}{2}}{\int_{-f_B}^{f_B} S_e^2(f)|NTF(f)|^2 df} \\ &= 10\log_{10}\frac{A^2 \cdot f_s}{2\sigma_e^2 \int_{-f_B}^{f_B} |NTF(f)|^2 df}. \end{aligned} \tag{18}$$

Applying the approximation of σ_e^2, we get the formula

$$\begin{aligned} SNR(\text{dB}) &\approx 10\log_{10}\frac{3\cdot 2^{2bits}\cdot A^2 \cdot f_s}{2\int_{-f_B}^{f_B} |NTF(f)|^2 df} \\ &\approx 10\log_{10} 3 \cdot 2^{2bits} A^2 f_s - 10\log_{10} 2\int_{-f_B}^{f_B} |NTF(f)|^2 df. \end{aligned} \tag{19}$$

Making use of numerical integration, this equation can be solved for an SDM with any noise transfer function, oversampling rate, and bit length. These calculations for the SNR approximation by computer are very fast and precise, and they are realized much faster than the SNR approximate estimation when the modulator output bitstream from simulations is used. If a loop filter transfer function is of odd order, using the function coefficients in Eq. (8), then the stability can be calculated without simulations. This provides the tools for SDM analysis without the need for SDM model simulations.

The practical relationship between Δu and modulator stability that depends on its signal value for one case is given in Fig. 4. Here we have a loop filter transfer function that produces losses of stability for input signals with amplitudes less than unity if unity is the scaled maximal input signal.

When a higher input signal is applied, the SNR usually rises. With raising the test sine wave amplitude it can be observed that at a given point, there is an SNR decrease and perhaps a loss of stability. This means that the modulator can be stable for input signals greater than Δu. Unfortunately, we cannot guarantee its stability ($\Delta u = 0.68$ in this example).

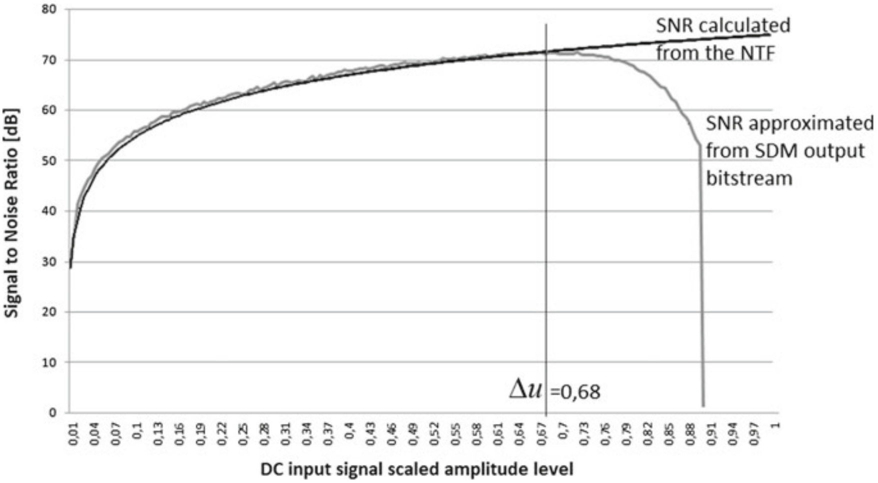


Fig. 4 Relationship between Δu and DC input signal value in terms of stability and SNR

3 Design Approach

Some authors [12, 13] claim that if the NTF zeros (loop filter poles) are not placed on the DC, but spread in the baseband, then the SNR is higher, and the quantization noise is spread evenly in the baseband, as shown in Fig. 5.

The Delta Sigma Toolbox [14] for MATLAB, created by this author, can generate such NTFs of arbitrary order. This toolbox was used to generate an exemplary noise transfer function with optimized zeros for a third-order SDM filter order that has the following form:

$$NTF(z) = \frac{z^3 - 2.999z^2 + 2.999z - 1}{z^3 - 2.1992z^2 + 1.6876z - 0.4441}. \tag{20}$$

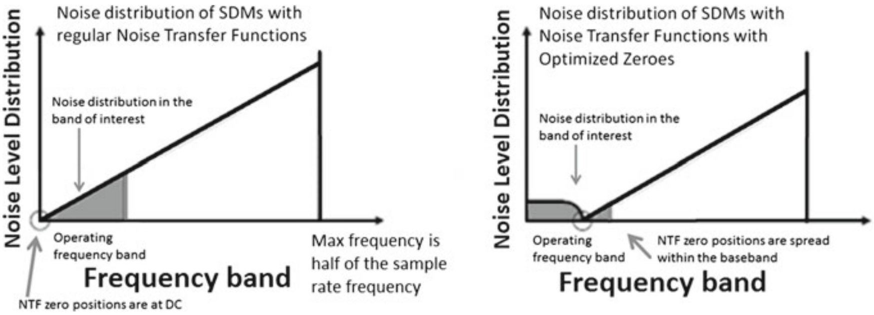

Noise Level Distribution
Noise distribution of SDMs with regular Noise Transfer Functions
Noise distribution in the band of interest
Operating frequency band
Frequency band
Max frequency is half of the sample rate frequency
NTF zero positions are at DC
Noise Level Distribution
Noise distribution of SDMs with Noise Transfer Functions with Optimized Zeroes
Noise distribution in the band of interest
Operating frequency band
NTF zero positions are spread within the baseband
Frequency band

Fig. 5 SDM noise distribution comparison with and without optimized NTF zeros

Our goal was to initiate a nonlinear constrained optimization [15] programming approach [14, 16] with maximization with respect to the SNR while maintaining a maximal reasonable limit for the maximal minimum level for the DC input level Δu that ensures stable modulator behavior. This is an approach similar to finding stable SDM NTFs of third order in terms of finding them with high SNR like the one in [17], where the poles were varied inside the unit circle and zeros were optimized from DSToolbox and kept constant. For this type of noise transfer function with Eq. (8) we can determine the SDM stability ranges even while moving the real pole outside the unit circle, and we can find the SNR very rapidly with the following constrained optimization procedure:

$$\begin{aligned}
&\max\,(SNR)\\
&SNR = 10\log_{10} 3\cdot 2^{2bits} A^2 f_s - 10\log_{10} 2\int_{-f_B}^{f_B} |NTF(f)|^2 df \quad [\mathrm{dB}],\\
&NTF = \frac{(1-zero\_ntf_1 * z^{-1})*(1-zero\_ntf_2 * z^{-1})*(1-zero\_ntf_3 * z^{-1})}{(1-pole_1 * z^{-1})*(1-pole_2 * z^{-1})*(1-pole_3 * z^{-1})},\\
&\text{Subject to}:\\
&\Delta u < \frac{\frac{b_1(2-zero\_ntf_1)}{zero\_ntf_1(zero\_ntf_1-1)} - \frac{2|\delta(1-\alpha)+\gamma\beta|}{(1-\alpha)^2+\beta^2}}{\frac{b_1}{zero\_ntf_1-1} + \frac{2|\delta(1-\alpha)+\gamma\beta|}{(1-\alpha)^2+\beta^2}}\\
&zero\_ntf_1 \geq 1,\ \ zero\_ntf_1 \leq 1,5\\
&pole_1 \geq 0,\ \ pole_1 \leq 1\\
&a \geq 0,\ \ a \leq 1,\ \ c \geq 0,\ \ c \leq 1\\
&b \geq 0^0,\ \ b \leq 90^0,\ \ d \geq 0^0,\ \ d \leq 90^0,
\end{aligned} \tag{21}$$

where

$$\begin{aligned}
&NTF(z) = \frac{1}{1+G(z)};\quad G(z) = \frac{b_1 z^{-1}}{1-zero\_ntf_1 z^{-1}} + \frac{b_2 z^{-1}}{1-zero\_ntf_2 z^{-1}} + \frac{b_3 z^{-1}}{1-zero\_ntf_3 z^{-1}}\\
&zero\_ntf_2 = \alpha + j\beta,\ \ zero\_ntf_3 = \alpha - j\beta,\ \ b_2 = \delta - j\gamma,\ \ b_3 = \delta + j\gamma,\\
&zero\_ntf_2 = a * \cos(b) + j * c * \sin(d),\ \ zero\_ntf_3 = a * \cos(b) - j * c * \sin(d)\\
&pole_2 = conj(pole_3),\ \ pole_3 = conj(pole_2).
\end{aligned}$$

This procedure deals with the transformation of a polynomial to the rooted form of the NTF and with specifying boundaries for the poles and zeros and one additional constraint for Δu. This constrained optimization problem was programmed with the `fmincon` function in MATLAB. Since that function searches for a global minimum, in the programing approach the goal function is the SNR, but the SNR was specified to be presented with negative values. With this form of procedure, we are to find the NTF with maximal SNR for third-order SDMs that is also stable. The theoretical peak SNR for third-order SDMs for 64 times oversampling ratio (OSR), scaled input amplitude of level 0.5, scaled sampling frequency Fs equal to 1, and a 1-bit quantizer is 106.8 dB. This NTF, however, provides unstable modulator behavior. The NTF function with optimized zeros generated from the DSToolbox provides a little over 85 dB SNR for the same parameters and is stable, and the candidate NTF in [17] provides around 90 dB for the same OSR, amplitude level, and quantizer bits, and 95 dB peak SNR for the case in which the input signal has maximal amplitude level.

Our goal is to find an NTF with better SNR performance in comparison to the existing known good NTFs for SDMs with high SNR. For this reason, the NTF coefficients from Eq. (20) were used as an initial conditions vector starting point.

4 Simulation Results

We have implemented the sigma–delta modulator design procedure mentioned in the previous section in MATLAB with the `fmincon` optimization function. The end result was an NTF function with stability and result suggesting that there should be a 95 dB SNR for third-order SDM with 64 times OSR for an amplitude of 0.5 for a 1-bit quantizer that is also stable.

The candidate noise transfer function that provides a maximal SNR from Eq. (21) is the following, when transformed from rooted to polynomial form (with numbers rounded up to the fourth digit):

$$NTF(z) = \frac{z^3 - 2.9983z^2 + 2.998z - 0.9997}{z^3 - 1.8933z^2 + 1.3352z - 0.3341} \tag{22}$$

In order to verify the correctness of the results from the implementation of the design approach procedure with respect to maximization of the SNR, we have tested the resulting candidate NTF obtained from the constrained nonlinear optimization procedure with simulations on a sigma–delta modulator model implemented also in MATLAB. We have observed stable SDM behavior without loss of stability, as can be seen in Fig. 6, where are plotted a sine wave input signal with amplitude 0.5 and the resulting bitstream on the output.

In order to better compare the NTFs from Eqs. (20) and (22) of Fig. 7, we have plotted the resulting SNR for input signal amplitude variation for SDM with 64 times OSR and a 1-bit quantizer. This difference will apply when changing the OSR or increasing the output quantization levels.

One example of a modulator power spectrum shape obtained after simulations, scaled with respect to sampling frequency ($Fs = 1$), when using 64 times oversampling ratio and test sine wave with scaled amplitude of 0.5 for the sigma–delta modulator with noise transfer function from Eq. (22), is shown in Fig. 8. In this case, the effective SNR obtained from simulations is also around 95 dB, proving the correctness of the solution. The resulting data suggests (Table 1) that the NTF from Eq. (22) provides around 10 dB better SNR than the DSToolbox generated NTF and 5 dB better SNR from the resulting NTF from [17].

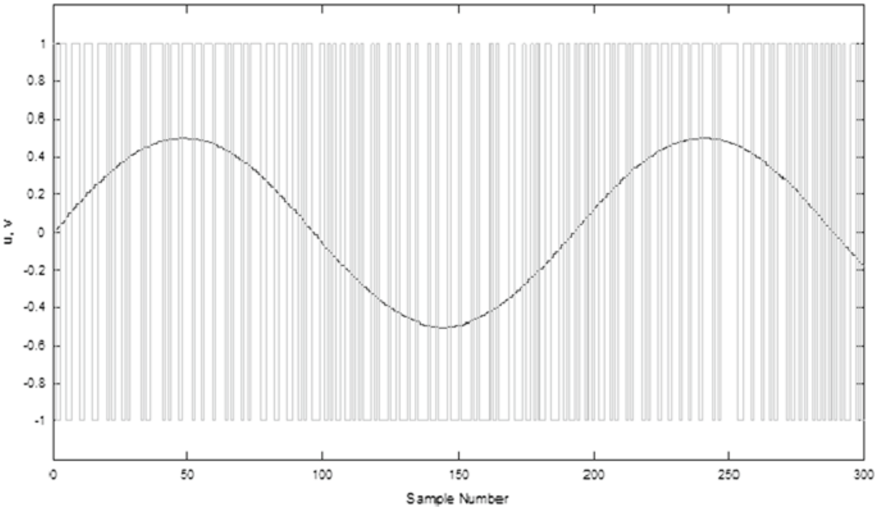
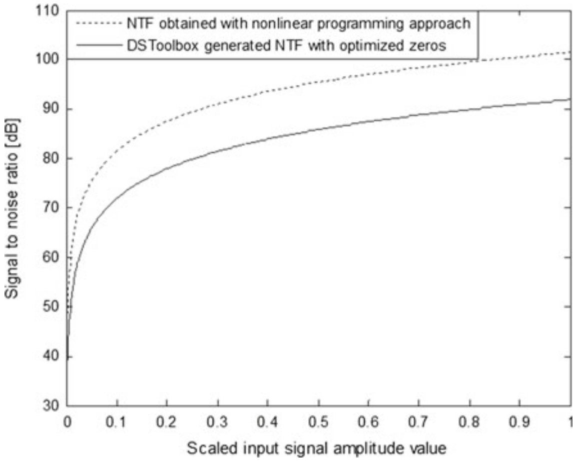


Fig. 6 SDM input and output of simulated SDM with third-order NTF from Eq. (22)

Fig. 7 Input signal and resulting SNR for SDMs when using $NTF(z)$ from Eqs. (20) and (22) for 64 OSR and quantization level of 1 bit



5 Conclusion

In the paper, a design approach for s stable high-performance third-order sigma–delta modulator was presented. The approach is based on formulating and solving an appropriate constrained optimization problem. The goal function is maximization of the signal-to-noise ratio, and the constraints come from stability issues and the ranges of the noise transfer function coefficients. This is possible due to the use of an analytic formula for calculation of the signal-to-noise ratio and results for stability and maximum stable DC input signal. Based on the presented approach, a third-order

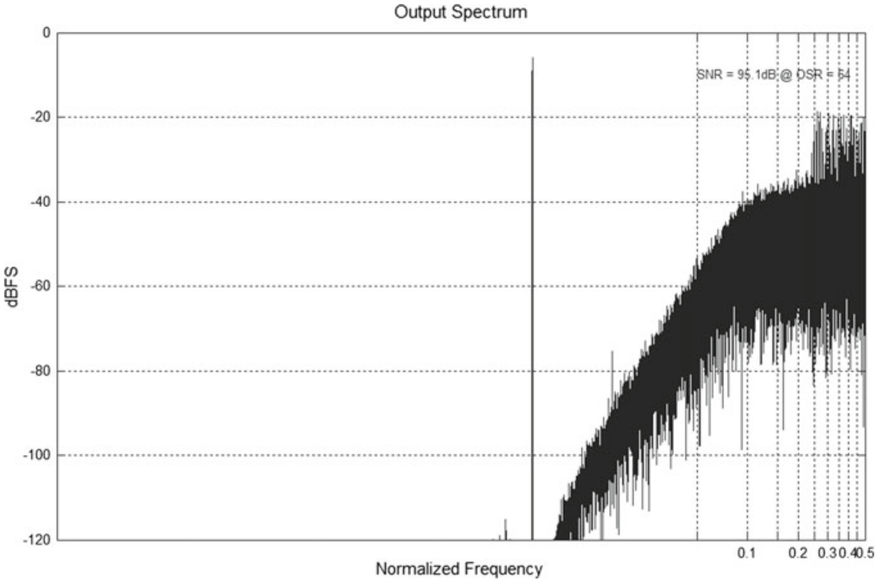


Fig. 8 SDM spectrum of simulated SDM output with third-order NTF from Eq. (22)

Table 1 SNR performance comparison between three stable NTFs

NTF	Resulting rounded SNR for input with amplitude level of 0.5 and 1-bit quantizer (dB)
DSToolbox generated NTF $NTF(z) = \frac{z^3-2.999z^2+2.999z-1}{z^3-2.1992z^2+1.6876z-0.4441}$	85
The resulting NTF from [17] $NTF(z) = \frac{z^3-2.997z^2+2.996z-0.999}{z^3-1.997z^2+1.535z-0.467}$	90
The resulting NTF from Eq. (22) $NTF(z) = \frac{z^3-2.9983z^2+2.998z-0.9997}{z^3-1.8933z^2+1.3352z-0.3341}$	95

stable SDM with reasonable performance in the sense of SNR and stable DC input signal range was obtained. The approach can be easily generalized and extended to higher-order modulators.

Acknowledgements This work is supported by N.W.O. visitor travel grant Nr. 040.11.493 for 2015.

References

1. Schreier, R., Temes, G.C.: Understanding Delta-Sigma Data Converters. Wiley, New Jersey (2005)
2. Reefman, D., Janssen, E.: Signal processing for direct stream digital: a tutorial for digital sigma delta modulation and 1-bit digital audio processing. Philips Research, Eindhoven, White Paper (2002). Accessed 18 Dec 2002
3. Jantzi, S., Schreier, R., Snelgrove, M.: Bandpass sigma-delta analog-to-digital conversion. IEEE Trans. Circuits Syst. **38**(11), 1406–1409 (1991)
4. Gore, A., Chakrabartty, S.: A min-max optimization framework for designing $\Sigma\Delta$ learners: theory and hardware. IEEE Trans. Circuits Syst. I **57**(3), 604–617 (2010)
5. Ho, C.-F., Ling, B., Reiss, J., Liu, Y.-Q., Teo, K.-L.: Design of interpolative sigma delta modulators via semi-infinite programming. IEEE Trans. Signal Process. **54**(10), 4047–4051 (2006)
6. Nagahara, M., Yamamoto, Y.: Optimal noise shaping in delta-sigma modulators via generalized KYP lemma. In: Proceedings of IEEE ICASSP, pp. 3381–3384 (2009)
7. Yu, S.-H.: Analysis and design of single-bit sigma-delta modulators using the theory of sliding modes. IEEE Trans. Control Syst. Technol. **14**(2), 336–345 (2006)
8. McKernan, J., Gani, M., Yang, F., Henrion, D.: Optimal low-frequency filter design for uncertain 2–1 sigma-delta modulators. IEEE Signal Process. Lett. **16**(5), 362–365 (2009)
9. Nagahara, M., Yamamoto, Y.: Frequency domain min-max optimization of noise-shaping delta-sigma modulators. IEEE Trans. Signal Process. **60**(6), 2828–2839 (2012)
10. Mladenov, V., Hegt, H. , Roermund, A.v.: On the stability analysis of sigma-delta modulators. In: 16th European Conference on Circuit Theory and Design ECCTD 2003, Cracow, Poland, pp. I-97–I-100 (2003)
11. Mladenov, V., Karampelas, P., Tsenov, G., Vita, V.: Approximation formula for easy calculation of signal-to-noise ratio of sigma-delta modulators. ISRN Signal Process. Article ID 731989 (2011). doi:10.5402/2011/731989
12. Schreier, R.: An empirical study of high-order single-bit delta-sigma modulators. IEEE Trans. Circuits Syst.-11: Analog Digit. Signal Process. **40**(8), 461–466 (1993)
13. Tsenov, G., Mladenov, V., Reiss, J.D.: A comparison of theoretical, simulated, and experimental results concerning the stability of sigma delta modulators. In: 124th AES Convention (2008)
14. Schreier, R.: Delta sigma toolbox. http://www.mathworks.com/matlabcentral/fileexchange/19
15. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution: A Practical Approach to Global Optimization, 1st edn. Springer, Berlin (2005)
16. Sarker, Ruhul: Masoud Mohammadian and Xin Yao, Evolutionary Optimization, 1st edn. Springer, Berlin (2002)
17. Tsenov, G., Mladenov, V.: A design procedure for stable high order. High performance sigma-delta modulator loopfilters. Springer Communications in Computer and Information Science, vol. 438, pp. 114–124 (2014)

Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review

Mouhannad Ali, Ahmad Haj Mosa, Fadi Al Machot and Kyandoghere Kyamakya

Abstract Emotions play an extremely important role in how we make decisions, in planning, in reasoning, and in other human mental states. The recognition of a driver's emotions is becoming a vital task for advanced driver assistance systems (ADAS). Monitoring drivers' emotions while driving offers drivers important feedback that can be useful in preventing accidents. The importance comes from the fact that driving in aggressive moods on road leads to traffic accidents. Emotion recognition can be achieved by analyzing facial expression, speech, and various other biosignals such as electroencephalograph (EEG), blood volume pulse (BVP), electrodermal skin resistance (EDA), electrocardiogram (ECG), etc. In this chapter, a comprehensive review of the state of-the-art methodologies for emotion recognition based on physiological changes and speech is presented. In particular, we investigate the potential of physiological signals and driver's speech for emotion recognition and their requirements for ADAS. All steps of an automatic recognition system are explained: emotion elicitation, data preprocessing such as noise and artifacts removal, features extraction and selection, and finally classification.

1 Introduction

Human emotion is a state involving various physical structures; it is either gross or fine-grained behavior, and it occurs in particular situations [21]. The ability to understand and discern a driver's emotions while driving and to perform the

M. Ali (✉) · A.H. Mosa · F.A. Machot · K. Kyamakya
Institute of Smart System Technologies, Alpen-Adria University, Klagenfurt, Austria
e-mail: Mouhannad.Ali@aau.at

A.H. Mosa
e-mail: AhmadHaj.Mosa@aau.at

F.A. Machot
e-mail: FadiAl.Machot@aau.at

K. Kyamakya
e-mail: Kyandoghere.Kyamakya@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_13

appropriate actions has been identified as one of the key focus areas listed by international research groups for improving intelligent transportation systems [4]. However, recognition of the emotional state and response during driving is an extremely difficult task and is still a scientific challenge. One of the main difficulties is that the emotion-relevant signal patterns may differ widely from person to person or from one specific situation to another. Moreover, it is hard to find an exact correlation between classes (patterns) due to the problem of precise definition of emotions and their meanings [22].

Nevertheless, the emotions and reaction(s) of a driver can be captured and measured using appropriate biosensors. Most researchers in the field of emotion recognition have focused on the analysis of data originating from a single sensor, such as audio (speech) or video (facial expression) data [9]. Lately, many studies in the emotion recognition field have begun to combine multiple-sensor data in order to build a robust emotion recognition system. The main target of using a combination of multiple sensors is that we as humans use a combination of different modalities in our body to express emotional states during human interaction. The human modalities are divided into audiovisual (facial expression, voice, gesture, posture, etc.) and physiological (respiration, skin temperature, etc.) [21].

Computers can be made to understand human emotions by capturing these modalities, extracting a set of useful features from them, and fusing those features in order to infer an accurate emotional state. There is a growing number of sensors that can capture various physical manifestations of emotion: video recordings of facial expressions [13], vocal inflection changes [2], EEG, skin-surface sensing of muscle tension, electrocardiogram (ECG), electrodermal activity (EDA), body temperature, etc.

In this chapter, an overview of the recent state of emotion recognition approaches involving speech signals and different physiological signals is presented. In particular, it is focused on emotion elicitation scenarios, features extraction, and selection and classification methodologies. The main goal of this review is to get an idea about the current situation (state of the art) of emotion recognition approaches and the current advancement in this filed.

This chapter is organized as follows: Sect. 2 describes the theories of emotion and the different categories of emotion types. Section 3 presents the physiological measures of human emotion recognition. Section 4 discusses the implementation steps of an emotion recognition system using physiological and speech signals. The overview of previous research work on emotion recognition using speech and physiological signals is presented in Sect. 5. Finally, a set of concluding remarks is given in Sect. 6.

2 Theories of Emotion

What is an emotion? "Everyone knows what an emotion is, until asked to give a definition" [14]. Emotion by definition is awareness of situations as relevant, urgent, and meaningful with respect to ways of dealing with it. According to cognitive theory, people's experience of emotion depends on the way they appraise or evaluate the events around them [28]. For example, a person sees a snake. His brain starts to process the situation as a dangerous one, then his heart rate increases, and he then feels afraid. In general, emotion is a complex concept involving three components [40]:

- Subjective experience: There is a number of basic universal emotions [12] experienced by all humans regardless of culture and race. However, the way of experiencing these emotions is highly subjective [22].
- Emotion expressions: These are observable and nonverbal behaviors that illustrate an affective or internal emotion state. For example, a smile indicating happiness or pleasure and a frown indicating sadness or displeasure. In general, expressions include audiovisual such as face, gesture, posture, voice intonation, breathing noise.
- Physiological response: This is a biological arousal or a physical reaction the body experience during an emotion. For example, when we are frightened, our heart races, our breathing becomes rapid, our mouth becomes dry, our muscles tense, our palms become sweaty, and we may want to run [28].

Emotions can be categorized into various types. The two most frequently applied models for emotion classification are the "discrete emotion model" proposed by Ekman [12] and the "two-dimensional valance arousal model" proposed by Lang [27]. The discrete emotional model categorizes emotions into six basic emotions—happiness, sadness, surprise, anger, disgust, and fear [12]. These emotions are biologically fixed and universal to all humans. They are widely accepted. The dimensional model assumes that emotions are a combination of several psychological dimensions. The best-known dimensional model is the "valance arousal dimensional model." Valance represents the pleasure level and ranges from negative to positive. Arousal indicates the physiological and psychological level of being awake and ranges from low to high [23].

3 Physiological and Speech Signals

The general way to recognize the emotional state of a subject is through his speech, facial expression, or gesture. The speech signal can carry the emotional state of the speaker [31]. Williams and Stevens [43] found that when the sympathetic nervous system is aroused with the emotions of anger, fear, or joy, speech becomes loud, fast, and enunciated with strong high-frequency energy. Moreover, when a subject is sad, his parasympathetic nervous system is aroused and his speech becomes slow with high-frequency energy. According to the cited authors, emotion affects overall energy, energy distribution across the frequency spectrum, and the frequency and duration of pauses of speech signals.

However, detecting the physiological patterns of a subject can also give information about his emotional state, because when a subject is positively or negatively excited, the sympathetic nerves of the autonomic nervous system are activated [43].

This sympathetic activation increases respiration rate, raises heart rate, decreases heart rate variability, and raises blood pressure [41]. The most common physiological signals used for emotion recognition include:

- **Electromyography (EMG)**: This refers to the muscle activity or frequency of muscle tension of a certain muscle. EMG detects the electrical potential generated by muscle cells when these cells are electrically or neurologically activated [36]. High muscle tension often occurs under stress. It can also be measured on the face to distinguish between negative and positive emotions. Figure 1 shows an example of an EMG signal recorded at the smiling muscle (zygomaticus major muscle) during three consecutive muscle contractions.

- **Electrodermal activity (EDA)**: This refers to skin conductivity (SC); it basically measures the conductivity of the skin, which increases if the skin is sweaty. This signal was found to be a good and sensitive indicator of stress as well as other stimuli and also helps to differentiate between conflict and no-conflict situations or between anger and fear. The problem with this signal, however, is that it is also influenced by external factors such as outside temperature. It therefore needs reference measurements and calibration [16]. Figure 2 illustrates the skin conductance response (SCR) in an EDA signal that is occurring in reaction to a stimulus [20].

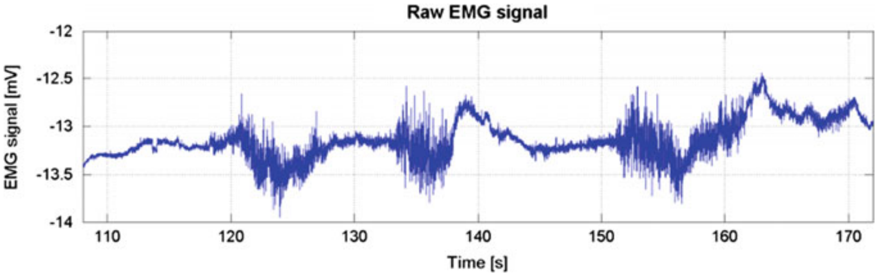
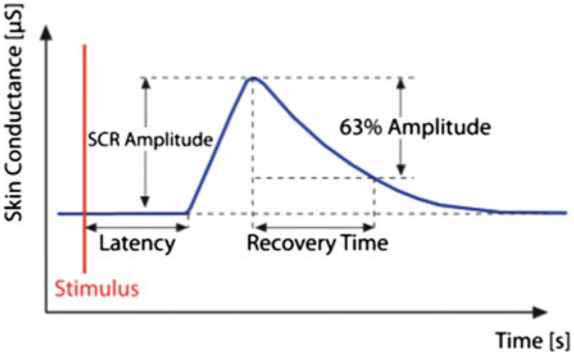


Fig. 1 Example of an EMG signal recorded at the zygomaticus major (smiling muscle), showing three muscle contractions [19]

Fig. 2 Ideal skin conductance response (SCR) in the EDA signal [20]



- **Skin temperature**: This is a measure of the peripheral skin temperature. The skin temperature depends on the blood flow in the underlying blood vessels. Since muscles are tense under strain, the blood vessels will be contracted, and therefore the temperature will decrease. In general, it is a relatively slow indicator of changes in emotional state [16]. Nevertheless, during happiness or anger, the temperature increases, and it decreases during sadness or fear.

- **Blood volume pulse (BVP)**: This is a measure to determine the amount of blood currently running though the vessels using a photoplethysmogram (PPG). A PPG consists of a light source and a photo sensor, which are attached to the skin. The source bounces infrared light against a skin surface and measures the amount of reflected light. BVP is used for emotion recognition, whereby the BV increases during anger or stress and decreases during sadness and relaxation. Moreover, BVP can be used to measure vasoconstriction and heart rate [16].

- **Electrocardiogram (ECG)**: Each healthy heartbeat has an orderly progression of depolarization that begins in the sinoatrial node, which generates an electrical impulse. This impulse spreads through the heart muscle and causes the contraction of the heart. The accumulation of action potentials traveling along the heart muscle generates electrical potential fluctuations. The electrical impulses generated by the heart can be measured on the surface of the skin over a period of time with electrodes [15]. This process of recording is called ECG. The ECG signal is a recurring pattern, as schematically depicted in Fig. 3. The ECG signal consists

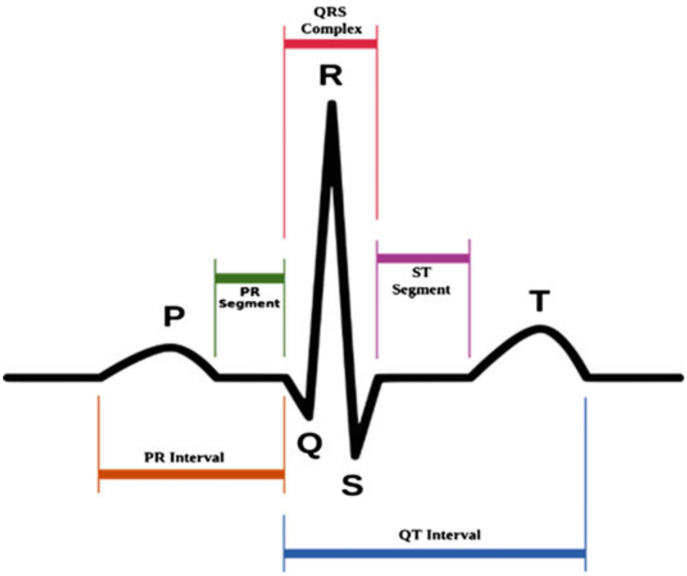


Fig. 3 Typical ECG signal with P, QRS, and T waves. [3]

of three main waves. The first is known as the P wave, which corresponds to the depolarization of the atrium. The is the QRS wave, which indicates the start of ventricular contraction. Finally, after the ventricles have contracted for a few milliseconds, the third wave, known as the T wave, occurs when the ventricular muscle repolarizes [15].

The R-peak represents the most prominent attribute of the ECG, and its time stamp can be precisely determined. It can be used to measure heart rate (HR) and interbeat intervals (IBI) to determine heart rate variability (HRV). A low HRV can indicate a state of relaxation, whereas an increased HRV can indicate a potential state of mental stress or frustration [19].

- **Electroencephalogram (EEG)**: An electroencephalography signal is the measurement of brain waves. It can be used to evaluate brain disorders. The brain waves are generated by current flow during synaptic excitation of the dendrites of many pyramidal neurons in the cerebral cortex [38]. The EEG signals are measured using small, flat metal disks (electrodes) attached to the scalp. There are five major brain waves distinguished by their different frequency ranges. These frequency bands from low to high frequencies are called respectively [38]:

1. Delta (δ) waves, which lie within the range of 0.5–4 Hz. They are usually present during deep sleep and may appear in the waking state.
2. Theta (θ) waves, which lie within the range of 4–7.5 Hz. They are usually present during drowsiness and are associated with increased learning, creativity, and deep

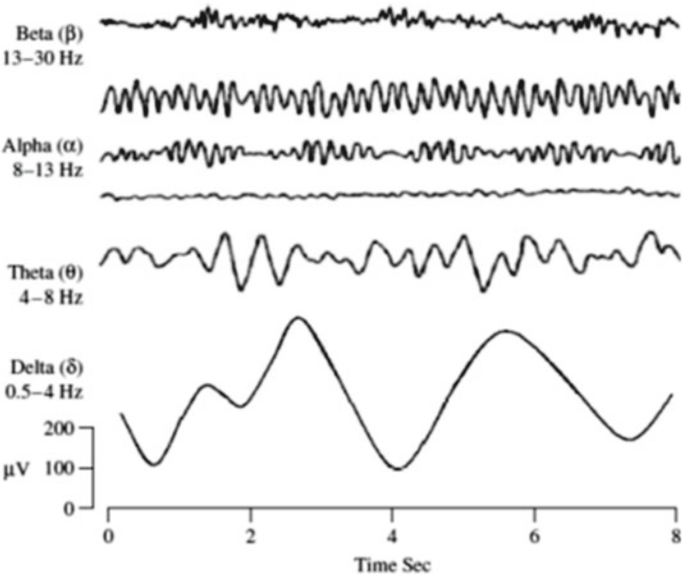


Fig. 4 Four typical brain waves, from high to low frequencies. [38]

meditation. They allow access to the unconscious. Moreover, a theta wave seems to be related to the level of arousal.

3. Alpha (α) waves lie within the range of 8–13 Hz. In general, an alpha wave appears as a round or sinusoidal signal; they are usually associated with relaxation and super-learning.
4. Beta (β) waves lie within the range of 14–26 Hz. They are usually associated with active thinking, active attention, or solving concrete problems.
5. Gamma (γ) waves correspond to the frequencies above 30 Hz. The detection of these waves can be used for confirmation of certain brain diseases.

Figure 4 illustrates the first four brain waveforms with their usual amplitude levels.

- **Respiration**: this measurement indicates the breathing rhythm of a person. It is captured be applying a rubber band around the chest. Usually, fast and deep breathing can point to anger or fear but sometimes also joy. Rapid shallow breathing can indicate fear, panic, or concentration. Moreover, slow and deep breathing refers to relaxation, and slow and shallow breathing can indicate depression or calm happiness [16].

The combination of these signals can be used to derive a set of features that can be used to build a robust classifier. This is then used to automatically detect the emotional state of different subjects.

4 The Implementation Steps of an Emotion Recognition System Based on Physiological and Speech Signals

4.1 Emotion Elicitation

The creation of a high-quality dataset (i.e., a reference database) of speech and biological signals still represents a necessary task for researchers in the field of emotion recognition. Different scenarios have been used to elicit emotion. Current rsearch highlights seven criteria that help with the selection and use of these emotion elicitation scenarios [7]:

1. Intensity: Do the elicitation scenarios lead to intense negative and positive emotion?
2. Complexity: Is the scenario a simple one such as showing a silent image such as a fixation cross, or is it more complex, such as a dynamic visual and auditory sequence?
3. Attention capture: Do the scenarios require much attention?
4. Demand characteristics: Does the scenario include a specific instruction such as "please watch this video carefully." In general, it is content-dependent?

5. Standardization: Is the scenario going to affect all participants in the same way? For example, is it possible to guarantee that showing a specific video will be equally effective for all participants?
6. Temporal consideration: Emotions can be considered relatively rapid phenomena with onsets and offsets over seconds. Generally, film clips are much lower in temporal resolution and range from about 1 to 10 min.
7. Ecological validity: To what extent will emotion elicitation procedures elicit emotions in the way that many stimuli encountered in real life do?

Moreover, the most-used scenarios for emotion elicitation are the following [7]:

Film clips: This scenario works by showing many short film clips to participants. The advantage of this method is that it is a rich source of discrete emotions (love, anger, fear, joy, etc.), which can be self-reported by participants. On the other hand, the disadvantages of this method are that it is necessary to extract particular periods of interest from the film. Furthermore, due the fact that emotions are considered evanescent phenomena, any delay between the activation of emotion and the assessment of it by an experimenter can introduce an error in the measurement.

Pictures: This scenario works by showing a set of pictures to participants. The advantages of this method are that it is easy and fast to apply and can also be self-reported by participants. However, the disadvantages of the method are that it is not a rich source of discrete emotions compared to film clips, and the time for stimulating emotion is too short.

Music: This scenario applies by playing music to participants. The advantages are that it is simple and highly standardized, and emotions develop over time (15–20 min). The drawbacks are that the music tastes of the participant might influence experienced emotions. In addition, this scenario gives only the moods (positive or negative), not the discrete emotions.

Emotional behaviors as emotional stimuli: This scenario involves the manipulation of the target person's behavior or the person's understanding of that behavior in order to change his/her feelings. The advantage of this method consists in the large sources to produce emotions (posture, eye gaze, tone of voice, breathing, and emotional actions). On the other hand, the disadvantage is that in some cases, it would be easy to manipulate the subject, but others might be more difficult (for example, making the participant angry).

Dyadic interaction tasks: Emotion is here elicited through interaction with different types of dyads (friends, romantic partner, family member, etc.). The advantages of this method are that it elicits a range of emotional responses and it studies emotion in social contexts. The disadvantages of this method are that (1) it requires significant resources, for example, dyadic interaction procedures can take 2–4 h; (2) some procedures my not be completed (the participant changes the topic to avoid a high level of emotional intensity; finally, (3) it provides just a snapshot sampling of emotions.

In general, choosing the appropriate elicitation scenario or stimuli depends on the target emotions and available sensors. For example, if we need to extract a speech signal from a subject, then music and picture scenarios will not be useful. The useful scenarios in this case are "emotional behaviors as emotional stimuli" and "dyadic interaction" tasks. Moreover, if we want to extract discrete emotions from a subject, we cannot choose the music scenario, because it just induces moods (positive or negative).

4.2 Preprocessing of Involved Signals

Both speech and physiological signals always contain unwanted modifications during capture, processing, or transmission. These unwanted modifications are noises and other external interferences such as artifacts that appear because of the electrostatic devices and muscular movements [24]. These noises and artifacts should be removed from the signal by the use of different types of appropriate filtering techniques. The appropriate filter is defined according to the signal's type and noise patterns. Low-pass filters such as adaptive filters, elliptic filters, and Butterworth filters generally are used to preprocess the raw ECG and facial EMG signals. And smoothing filters are used to preprocess the raw GSR signals [5, 21]. Also, low- and high-pass filters are used to preprocess EEG signals. Moreover, different methods have been used to remove artifacts from physiological signals. Principal component analysis (PCA) [37] and independent component analysis (ICA) [11] are the best-known methods for removing artifacts from EEG, ECG, and EMG.

Moving to speech signals, preprocessing including preemphasis, framing, and windowing processes have to be integrated [25].

4.3 Feature Extraction

Once the signals have been preprocessed, it is necessary to extract useful information or features from these signals in order to use them in pattern classification to detect the emotional state. The features to be extracted are chosen according to the signal type. Some features (such as mean, standard deviation, minimum, maximum, and range) can be extracted from most of the recorded sensor signals, whereas some special features are extracted only from a specific signal type.

From the speech signal: the best-known speech features usually extracted for emotion recognition include prosodic and spectral features. Prosodic features include pitch, pitch histogram, intensity, formant frequency, and voice quality [25]. Spectral features include Mel-frequency cepstral coefficients (MFCC), Daubechies wavelets, coefficient histogram [44], linear prediction cepstral coefficients (LPC), log fre-

Table 1 ECG features

Feature	Domain
Standard deviation of all RR interval	Time
Root mean Square of differences between RR intervals	Time
Power in low frequency	Frequency
Power in high frequency	Frequency

quency power coefficients (LFPC), and perceptual linear prediction (PLP) coefficients [44].

From the electrocardiogram signal: The most-used ECG features are heart rate (HR) and heart rate variability (HRV). Moreover, [1] used the Hilbert instantaneous frequency and a measure of local oscillation as feature extraction. Table 1 lists different ECG features in the frequency and time domains.

From the electrodermal activity signal: The extracted features from EDA are the average, skin resistance, zero crossing rate of skin conductance, average of absolute derivative, skin conductance response (SCR), and the nonspecific skin conductance response [21].

From the respiration signal: Respiration rate, average, and breathing rhythm are the best-known respiration features. Moreover, the average breath depth and spectral power are also commonly used [19, 21].

From the electromyogram signal: The most frequently extracted features from the EMG are mean value, root mean square, and the power [19, 21].

From the electroencephalogram EEG signal: EEG power spectra at distinct frequency bands, such as delta, theta, alpha, beta, and gamma, are commonly used as indices for assessing the correlates of specific ongoing cognitive processes in EEG research [25, 38], fast Fourier transform analysis, wavelet analysis, and high-order crossing [33].

4.4 Feature Reduction

After the features are extracted from the signals, it is useful to determine which features are most relevant to differentiate well between emotional states. Reducing the dimension of the feature space has two advantages:

- The computational costs are lowered, which leads also to shorter training times [18].
- Overfitting is reduced and prediction performance is improved by excluding irrelevant or noisy features in the learning process [46].

The most used methods for selection of features are principal component analysis (PCA) [37], independent component analysis (ICA) [11], random subset feature selection (RSFS) [35], and sequential floating forward selection (SFFS) [34].

4.5 Classification

After extracting and selecting the features that are appropriate for a best possible differentiation of emotional states, the next step is to use these features to train a classifier and test whether it can classify different emotional states. This section describes the most popular classification algorithms from the literature. We have selected the following classifiers:

4.5.1 Quadratic Discriminant Analysis (QDA)

Quadratic discriminant analysis (QDA) is the most commonly used method. It assumes that the likelihood of each class is normally distributed and uses the posterior distributions to estimate the class for a given test point [17]. The normal (Gaussian) parameters of each class are usually estimated from training points with maximum likelihood (ML) estimation [39].

4.5.2 k-Nearest Neighbor (KNN)

KNN classifies unlabeled samples (testing data) by their similarity with the training data. In general, given an unlabeled sample X, the KNN classifier finds the K closest neighborhood samples in the training data, and it labels the sample X with the class label that appears most frequently in the K closet neighborhood samples in the training data [32].

4.5.3 Support Vector Machines (SVMs)

The SVM is a classifier that separates a set of objects into classes so that the distance between the class borders is as large as possible. The idea of SVM is to separate two classes with a hyperplane so that the minimal distance between elements of both classes and the hyperplane is maximal [8].

4.5.4 Artificial Neural Network

Artificial neural networks are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown [45]. Artificial neural networks are a computational model that uses the idea of natural neurons that receive signals and exchange messages between each other. Each connection between neurons (nodes) has a numeric weight that can be tuned based on experience, thus making the network capable of learning. In general, ANNs start out with randomized weights for all their neurons. For that, they must be trained to solve a particular problem. The training phase can be applied using two methods, depending on the problem the ANNs must solve. The first method is self-organizing ANN. It uses a large amount of data and tries to discover patterns and relationships in those data. The second method is back-propagation ANN, which is trained by humans to perform specific tasks [6, 47].

5 Previous Works

Emotion recognition has become an important research topic mainly in the field of human–machine interaction. Several studies have been done on emotion recognition using speech and physiological signals. Initially, researchers tried with subject-dependent approaches, where the emotion recognition system is performed only on one user, and it needs to be retrained or recalibrated in order to perform well on another user/subject. Nowadays, the focus has shifted in the direction of subject-independent approaches, where the emotion recognition system is tested on fully different subjects and not those on which it had been trained; i.e., it is tested with unknown speech and physiological signals. Table 2 illustrates a short review of previous works in emotion recognition using speech and physiological signals. The table shows which signals were analyzed, which stimuli were used for emotion elicitation, which emotion were recognized, the number of subjects involved in the experiments, and which features and classification method were applied. Also, the accuracy of the recognition approaches is included in the table.

We can observe that regarding the subject-dependent approaches, the maximum accuracy values reached are 96.58% for recognizing three arousal levels (high, medium, and low), 95% for four emotions (joy, anger, sorrow, and pleasure) and 91.7% for six emotions (amusement, frustration, anger, fear, sadness, and surprise). On the other hand, for subject-independent approaches, the maximum accuracies were 99.5% for recognizing one emotion (stress), 86% for two emotions (joy and sadness), and 70% for detecting four emotional states (joy, anger, sadness, and pleasure). We can also notice, for physiological signals, that besides the feature extraction and classification approaches, the emotion stimulus type involved also has an effect on the classification accuracy. In general, the sensors used, number of subjects, emotional states, stimuli used, feature extraction and classification methods are the required building blocks and parameters to build a robust and reliable emotion recognition system.

Table 2 Literature review on emotion recognition using physiological and speech signals

Ref. No	Signals	Features	Classifiers	Emotions	Stimuli	No of subjects	Accuracy in %
[23]	EMG ECG EDA RSP	Statistical, energy, subband spectrum, entropy	Linear discriminant analysis	Joy, anger, sadness, pleasure	Music	3, MIT-database	95 (sub dependent) 70 (sub independent)
[29]	GSR HR ST	No specific features stated	KNN, discriminant function analysis, Marquardt backpropagation	Sadness, anger, fear, surprise, frustration, amusement	Movies	14	91.7 (sub dependent)
[16]	EMG EDA BVP ECG RSP	Running mean running standard deviation slope	NN	Arousal, valance	IAPS (visual affective picture system)	1	96.58 Arousal 89.93 valence (sub dependent)
[42]	ECG	Fast fourier	Tabu search	Joy, sadness	Movies	154	86 (sub independent)
[10]	EDA HR	No specific features stated	fuzzy logic	Stress	Hyperventilation Talk preparation	80	99.5 (sub independent)
[30]	BVP EMG ST EDA RSP	Statistical features	SVM, Fisher LDA	Amusement, contentment, disgust, fear, sad, neutral	IAPS	10	90 (sub dependent) 92 (sub dependent)
[22]	EMG EDA ECG BVP ST RSP SPEECH	Statistical features, BRV, zero-crossing, MFCCs	KNN	Arousal, valance	Quiz dataset	3	92 (sub dependent) 55 (sub independent)
[26]	EDA HR EMG	No specific features stated	HMM	Arousal, valance	Robot actions	36	81 (sub dependent) 66 (sub independent

6 Conclusion

This chapter has reviewed and presented the main steps toward human emotion recognition systems using physiological and speech signals. It should be emphasized that building a generalized system (i.e., subject-independent) for classifying different emotional states is still a big challenge, particularly because emotions are highly subjective. Most of the state-of-the art methodologies are based on the subject-dependent approach. Upgrading to a subject-independent approach needs more sophisticated features, more robust classifiers, and eventually more sensory data for training and testing. For good referencing, the collection of sensor data is done during a specific emotion elicitation scenario. Hence, the emotion elicitation scenario plays an important role in defining the target emotional states and how strongly those emotions should be elicited.

Moreover, the respective techniques for feature extraction, feature selection and classification are also very important steps and building blocks towards a robust and reliable subject-independent emotion recognition system.

Some future research avenues that are worth mentioning are related to:

- Emotion state forecasting for short-, middle-, and long-term horizons. Here the human emotional system is considered a dynamical system that is externally excited by emotion elicitation-related elements of the contextual environment. The short-term time horizon covers some seconds to several minutes. The middle-term horizon should cover hours. And the long-term horizon should cover several days. It is evident that a reliable forecasting of the emotional states is a core enabling unit for some form of emotion-related early warning system.
- To improve robustness, reliability, and accuracy, our hypothesis is that neurocomputing-based classifier concepts offer the greatest potential for best-possible performance. A series of our won ongoing works is developing, optimizing, and benchmarking cellular neural network-based neurocomputing classifier concepts, which integrate related recent paradigms such as deep learning, reservoir computing, and echo state.

References

1. Agrafioti, F., Hatzinakos, D., Anderson, A.K.: ECG pattern analysis for emotion detection. IEEE Trans. Affect. Comput. **3**(1), 102–115 (2012)
2. Al Machot, F., Mosa, A.H., Fasih, A., Schwarzlmüller, C., Ali, M., Kyamakya K.: A novel real-time emotion detection system for advanced driver assistance systems. In: Autonomous Systems: Developments and Trends, pp. 267–276. Springer, Berlin (2012)
3. Atkielski, A.: Schematic diagram of normal sinus rhythm for a human heart as seen on ECG. http://commons.wikimedia.org/wiki/File:SinusRhythmLabels.png/ (2006). Accessed 17-March 2016
4. Burns, P.C., Lansdown, T.C.: E-distraction: the challenges for safe and usable internet services in vehicles. In: Internet Forum on the Safety Impact of Driver Distraction When Using In-Vehicle Technologies (2000)

5. Chang, C.-Y., Zheng, J.-Y., Wang, C.-J.: Based on support vector regression for emotion recognition using physiological signals. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2010)
6. Cigizoglu, H.K., Kişi, Ö.: Flow prediction by three back propagation techniques using k-fold partitioning of neural network training data. Hydro. Res. **36**(1), 49–64 (2005)
7. Coan, J.A., Allen, J.J.B.: Handbook of Emotion Elicitation and Assessment. Oxford University Press, Oxford (2007)
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
9. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
10. de Santos Sierra, A., Sánchez Ávila, C., Casanova, J.G., Del Pozo, G.B.: A stress-detection system based on physiological signals and fuzzy logic. IEEE Trans. Ind. Electron. **58**(10), 4857–4865 (2011)
11. Ekenel, H.K., Sankur, B.: Pattern Recognit. Lett. Feature selection in the independent component subspace for face recognition. **25**(12), 1377–1388 (2004)
12. Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. J. Personal. Soc. Psychol. **53**(4), 712 (1987)
13. Essa, I.A., Pentland, A.: A vision system for observing and extracting facial action parameters. In: Proceedings CVPR'94. 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994, pp. 76–83. IEEE (1994)
14. Fehr, B., Russell, J.A.: Concept of emotion viewed from a prototype perspective. J. Exp. Psychol.: General **113**(3), 464 (1984)
15. Fox, S.I.: Human Physiology 9th edn. McGraw Hill, New York (1996)
16. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: first steps towards an automatic system. In: ADS, pp. 36–48. Springer, Berlin (2004)
17. Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R.: The Elements of Statistical Learning, vol. 2. Springer, Berlin (2009)
18. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 112. Springer, Berlin (2013)
19. Kappeler-Setz, C.: Multimodal emotion and stress recognition. Dissertation, Eidgenössische Technische Hochschule ETH Zürich, Nr. 20086 (2012)
20. Kappeler-Setz, C., Gravenhorst, F., Schumm, J., Arnrich, B., Tröster, G.: Towards long term monitoring of electrodermal activity in daily life. Pers. Ubiquitous Comput. **17**(2), 261–271 (2013)
21. Katsis, C.D., Katertsidis, N., Ganiatsas, G., Fotiadis, D.I.: Toward emotion recognition in car-racing drivers: a biosignal processing approach. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **38**(3), 502–512 (2008)
22. Kim, J.: Bimodal emotion recognition using speech and physiological changes. Citeseer (2007)
23. Kim, J., André, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2067–2083 (2008)
24. Kim, K.H., Bang, S.W., Kim, S.R.: Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**(3), 419–427 (2004)
25. Konar, A., Chakraborty, A.: Emotion Recognition: A Pattern Analysis Approach. Wiley, New York (2014)
26. Kulic, D., Croft, E.A.: Affective state estimation for human–robot interaction. IEEE Trans. Robot. **23**(5), 991–1000 (2007)
27. Lang, P.J.: The emotion probe: studies of motivation and attention. Am. Psychol. **50**(5), 372 (1995)
28. Lazarus, R.S.: Progress on a cognitive-motivational-relational theory of emotion. Am. Psychol. **46**(8), 819 (1991)

29. Lisetti, C.L., Nasoz, F.: EURASIP J. Adv. Signal Process. Using noninvasive wearable computers to recognize human emotions from physiological signals. **2004**(11), 1–16 (2004)
30. Maaoui, C., Pruski, A.: Emotion recognition through physiological signals for human-machine communication. INTECH Open Access Publisher (2010)
31. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. Speech Commun. **41**(4), 603–623 (2003)
32. Peterson, L.E: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
33. Petrantonakis, P.C., Hadjileontiadis, L.J.: Emotion recognition from EEG using higher order crossings. IEEE Trans. Inf. Technol. Biomed. **14**(2), 186–197 (2010)
34. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognit. Lett. **15**(11), 1119–1125 (1994)
35. Räsänen, O., Pohjalainen, J.: Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In: INTERSPEECH, pp. 210–214 (2013)
36. Robertson, G., Caldwell, G., Hamill, J., Kamen, G., Whittlesey, S.: Research Methods in Biomechanics, vol. 2E. Human Kinetics, Champaign (2013)
37. Rocchi, L., Chiari, L., Cappello, A.: Feature selection of stabilometric parameters based on principal component analysis. Med. Biol. Eng. Comput. **42**(1), 71–79 (2004)
38. Sanei, S., Chambers, J.A.: EEG Signal Processing. Wiley, New York (2013)
39. Scholz, F.W.: Maximum likelihood estimation. Encyclopedia of Statistical Sciences (1985)
40. Thoits, P.A.: The sociology of emotions. Annu. Rev. Soc. **15**, 317–342 (1989)
41. Van Der Vloed, G., Berentsen, J.: Measuring emotional wellbeing with a non-intrusive bed sensor. In: Human-Computer Interaction–INTERACT 2009, pp. 908–911. Springer, Berlin (2009)
42. Wan-Hui, W., Yu-Hui, Q., Guang-Yuan, L.: Electrocardiography recording, feature extraction and classification for emotion recognition. In: 2009 WRI World Congress on Computer Science and Information Engineering, vol. 4, pp. 168–172. IEEE (2009)
43. Williams, C.E., Stevens, K.N.: Vocal correlates of emotional states. Speech Evaluation in Psychiatry, pp. 221–240 (1981)
44. Yang, Y.-H., Chen, H.H.: Music Emotion Recognition. CRC Press, Boca Raton (2011)
45. Yegnanarayana, B.: Artificial Neural Networks. PHI Learning Pvt. Ltd., New Delhi (2009)
46. Yoon, H.J., Chung, S.Y.: EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm. Comput. Biol. Med. **43**(12), 2230–2237 (2013)
47. Zhang, J.-R., Zhang, J., Lok, T.-M., Lyu, M.R.: A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. Appl. Math. Comput. **185**(2), 1026–1037 (2007)

A Hybrid Reasoning Approach for Activity Recognition Based on Answer Set Programming and Dempster–Shafer Theory

Fadi Al Machot, Heinrich C. Mayr and Suneth Ranasinghe

Abstract This chapter discusses a promising approach for multisensor-based activity recognition in smart homes. The research originated in the domain of active and assisted living, particularly in the field of supporting people in mastering their daily life activities. The chapter proposes (a) a reasoning method based on answer set programming that uses different types of features for selecting the optimal sensor set, and (b) a fusion approach to combine the beliefs of the selected sensors using an advanced evidence combination rule of Dempster–Shafer theory. In order to check the overall performance, this approach was tested with the HBMS dataset on an embedded platform. The results demonstrated a highly promising accuracy compared to other approaches.

1 Introduction

Active and assisted living (AAL) [31] aims at helping persons in mastering their daily life activities [18] by employing intelligent technical means to compensate for disabilities. One of the major issues of AAL systems is to recognize the behavior of a person (i.e., what the person is currently doing) robustly in order to be able to provide optimal support. Activity theory conceptualizes a person's behavior as activities that consist of series of simple events such as walking, running, pushing a button, and grabbing something.

Consequently, activity recognition systems use different types of sensors that extract low-level features from the environment. For a structured view on that envi-

F. Al Machot (✉) · H.C. Mayr · S. Ranasinghe
Institute for Applied Informatics, Application Engineering, Alpen-Adria-Universität,
Universitätsstrasse 65-67, 9020 Klagenfurt, Austria
e-mail: Fadi.AlMachot@aau.at

H.C. Mayr
e-mail: Heinrich.Mayr@aau.at

S. Ranasinghe
e-mail: Suneth.Ranasinghe@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_14

ronment, researchers define an aggregation of contexts according to the task context [22], the personal context, the environmental context, the social context, and the spatiotemporal context. These contexts have to be analyzed and interpreted in order to identify the current activity and subsequently the whole activity. For example, in smart home environments, it is common that different activities may share many similar sensors, e.g., preparing a meal and preparing a drink activities can share the same simple events such as entering the kitchen, opening the cupboard, and opening the fridge. Thus, such situations form a kind of uncertainty that can cause bad decisions.

In this chapter, an uncertainty handling approach that allows better decisions in such situations and that can be implemented in embedded platforms is presented. It has been performed within the realm of the Human Behavior Monitoring and Support[1] (HBMS) project [32], that aims at deriving support services from integrated models of abilities and episodic knowledge that an individual has had or has temporarily forgotten.

The chapter is organized as follows: Sect. 2 gives an overview of the state-of-the-art approaches and their limitations. Section 3 covers a wide range of uncertainty handling approaches. Section 4 explains the answer set programming paradigm. Section 5 discusses the overall architecture of our activity recognition system. Section 6 presents the obtained results and the overall performance evaluation. The chapter ends in Sect. 7 with a discussion about uncertainty handling with respect to the proposed approach. Finally, a conclusion is provided in Sect. 8.

2 Related Work

During the last decade, different approaches to human activity recognition under uncertainty have been reported. They can be classified into three major categories along with their underlying model types: knowledge-based context models, graphical models, and syntactic models. Figure 1 provides an overview of activity recognition approaches under uncertainty.

Knowledge-based context models use expressions and rules to describe context properties such as entities, their properties, and the relationship between them. To recognize complex human activities, for example, the Ontology Web Language (OWL) [1] and answer set programming (ASP) [2, 3] are used for ontology representation and knowledge base (KB) creation, respectively.

Graphical models are used to describe complex activities in a higher-level representation, e.g., Bayesian dynamic networks [46], hidden Markov models [49], Dempster–Shafer [29], conditional random fields (CRFs) [44], and Gaussian mixture models (GMM) [36].

Syntactic models describe real-world events by structuring them with the use of a set of production rules, e.g., rough set theory [45] and fuzzy logic [11]. The Ontology

[1] This work was funded by the Klaus Tschira Stiftung GmbH, Heidelberg.

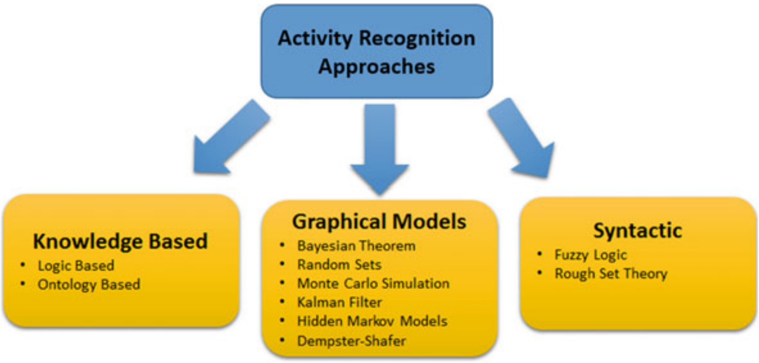


Fig. 1 An overview of complex event detection approaches under uncertainty

Web Language (OWL 2) is still a major research area [30, 35], and fortunately, OWL 2 ontologies are supported by a fuzzy logic-based reasoner to handle uncertainty.

Bayesian inference suffers from difficulty in defining a priori probabilities and the inability to consider general uncertainty [21]. Hidden Markov models (HMM) showed promising results in the field of activity recognition, but they do not perform perfectly, since human behavior is not Markovian [37]. The fuzzy logic sensor fusion method provides an effective means to handle requirements of human daily life [17]. However, fuzzy logic sensor fusion defines membership functions and production rules that are extremely domain- and problem-specific.

To overcome the limitations of the Bayesian inference method, the Dempster–Shafer method generalizes Bayesian theory to allow for distributing support not only to a single hypothesis but also to a union of hypotheses [23]. The Dempster–Shafer and Bayesian methods produce identical results when all the hypotheses are singletons (not nested) and mutually exclusive [4]. Additionally, the combination rule of the classical Dempster–Shafer theory can be implemented to fuse data from sensors, but it can lead to illogical results in the presence of highly conflicting evidence.

Therefore, we aim at a technique to propose a reasoning approach for activity recognition under uncertainty that (a) avoids the previous limitations, (b) responds in real time, (c) runs on embedded platforms, and (d) uses an evidence combination rule [4] that delivers logical results even in the presence of highly conflicting evidence.

3 Approaches to Uncertainty Handling

There is a variety of approaches for handling uncertainty in activity recognition. Since we will present a novel approach to this topic subsequently within this chapter, this section discusses the state of the art within that field.

Bayesian approach: A probabilistic distribution expressing data uncertainty was the first approach to handling the problem of imperfect data. Later, new techniques appeared that dealt with the limitations of probability theory, such as fuzzy set theory and evidential reasoning. Many event detection approaches require prior knowledge of the cross covariance of data to perform well. Unfortunately, prior knowledge can be affected by different sources of noise in the observation environment. The Bayesian inference network offers the following advantages: it incrementally estimates the probability of the truth of a hypothesis for new given observations; reasoning can be incorporated using prior knowledge about the likelihood of a hypothesis being true; and when empirical datasets are not available, it allows using subjective probability estimators to estimate the prior of hypotheses. Although Bayesian networks have these advantages, Bayesian reasoning also has some disadvantages [20] in that it suffers from the difficulty in finding prior probabilities, from complexities when there are multiple hypotheses and multiple conditionally dependent events, and from the inability to account for general uncertainty. Dynamic Bayesian networks [33] are suitable for the consideration of temporal aspects. They represent state variable changes over time. Moreover, Kalman filtering [47] is an optimal solution for estimating the moments of a probabilistic distribution that uses a series of measurements observed over time containing inaccuracies, uncertainties, and noise.

Hidden Markov models: Simple hidden Markov models (HMM) can be used to model simple events to detect complex events, but they do not support modeling temporal aspects. They offer the possibility to model temporal granularity, which is not possible with a simple HMM. Therefore, to solve this problem, layered HMMs offer this possibility.

Dynamic Bayesian networks (DBN) offer more flexibility in representing relationships between activities and subactivities, but some problems could arise when the system is detecting complex events that might be solved using tractable variational algorithms. DBN is a generalization of HMMs and CRFs. It supports modeling complex relationships between variables over time. However, this can affect the reasoning process. Tractable variational algorithms can help to eliminate this effect [7].

Fuzzy logic: Fuzzy set theory deals with vagueness of data, and evidential belief theory focuses on both uncertain and ambiguous data. However, a disadvantage of fuzzy logic is that it cannot be the main fusion method in a generalizable architectural solution to design a context-aware computing system. Moreover, fuzzy set membership function assignment and production rules are usually extremely domain- and problem-specific, making it difficult to implement the method as a general approach.

Dempster–Shafer: Dempster–Shafer theory performs well only under situations of minimal conflict or irrelevant conflict in which all sources are considered reliable [39]. Because of such limitations, new approaches have been developed, for example, the new version of DSET called the transferable belief model (TBM) [40] and DezertSmarandache theory (DSmT) [12]. The transferable belief model (TBM) theory extends DSET by DSmT, which allows the combination of all types of independent sources to be represented as belief functions, but it is specifically focused on the fusion of uncertain, highly conflicting sources of evidence. Moreover, the combination rule of the classical Dempster–Shafer theory can be implemented to fuse data from two sensors, but it can lead to illogical results in the presence of highly conflicting evidence. However, researchers in [4] proposed an evidence combination rule to provide more realistic results than those offered by the standard Dempster–Shafer combination rule. In order to perform event detection successfully, in the case of fusing sensors that do not require preliminary or additional information such as data distribution or a membership function, rough set theory is suitable [24].

Random sets and Monte Carlo simulation-based techniques: The conditional random fields technique models the conditional probability of observations for better class discrimination. A key advantage of CRFs is that they offer the possibility to include a wide variety of arbitrary nonindependent features of the input [28]. CRFs have been compared to HMMs for activity recognition. In general, they show better results than HMMs [43]. However, they need more computation time, especially if the low-level features are large. Several solutions have been suggested for optimizing the training of conditional random fields for event detection such as gradient tree boosting [13].

Furthermore, the Monte Carlo simulation-based techniques such as sequential Monte Carlo (SMC) and Markov chain Monte Carlo (MCMC) are among the most powerful approaches to approximating probabilities. Particle filters are a recursive implementation of the SMC algorithm [19]. They provide an alternative for Kalman filtering in dealing with non-Gaussian noise and nonlinearity in the system. They assign weights to the randomly chosen samples (particles) to approximate the probability density. Particle filters can be used in the framework of event detection to increase the performance of Bayesian approaches.

Ontologies and logic Based: Ontologies and logic-based event detection approaches are a tentative solution to performing complex reasoning tasks. The current frequently used ontology language is the Ontology Web Language (OWL 2), which has recently become a W3C recommendation for ontology representation [15]. Therefore, several fuzzy extensions of description logics can be found in the literature [27], and some fuzzy DL reasoners have been implemented, e.g., fuzzyDL [8] and Fire [42]. Each reasoner uses its specific fuzzy description logic (DL) language to model the fuzzy ontologies. Therefore, there is a need for a standard way to represent such information.

Logic-based approaches that use hidden Markov models, Bayesian networks, or conditional random fields typically encode only pairwise event constraints, and therefore, they take time points as primitives of their models. Consequently, many types of events are fundamentally interval-based and are not accurately modeled in terms of time points [10].

Hybrid approaches: Hybrid approaches combine components (methods) of complex event detection to gain the advantages of each approach. Some hybrid event detection approaches, e.g., the hybridization of fuzzy set theory with D-S evidence theory, have been studied frequently [48].

A combination of fuzzy set theory with rough set theory (FRST), proposed by Dubois and Prade, is another important theoretical hybridization that has appeared in the literature [14]. Application of FRST to complex event detection in visual surveillance systems has not often been investigated, since rough set theory itself is still not an established data event detection approach under uncertainty.

4 Answer Set Programming (ASP)

Answer set programming (ASP) [6, 9] is widely used in artificial intelligence (AI0). It is recognized as a powerful tool for knowledge representation and reasoning, especially due to its high expressiveness and ability to deal with incomplete knowledge.

ASP programs consist of two major parts: the knowledge base part, in which the facts are included, and the rules part, which describes how the problem should be solved. The output of ASP systems is the answer sets (models) that present the possible solutions of the encoded problem. Figure 2 shows the overall steps for solving problems using ASP.

An ASP program formulated in the language of AnsProlog (also known as A-Prolog) is a set of rules of the form

$$a_0 \leftarrow a_1, \ldots, a_m, \neg a_{m+1}, \ldots, \neg a_n, \tag{1}$$

where $1 \leq m \leq n$, and each a_i is an atom of some propositional language. Here $\neg a_i$ is a negation-as-failure literal (naf-literal). Given a rule of this form, the left- and right-hand sides are called head and body, respectively.

A rule may have either an empty head or an empty body, but not both. Rules with an empty head are called constraints; rules with an empty body are called facts.

Let X be a set of ground atoms in a given ASP program, i.e., all atoms that do not have free variables; as such, X is the Herbrand base of that ASP program. Then the body in a rule of the form (1) is satisfied by X if $\{a_{m+1}, \ldots, a_n\} \cap X = \phi$ and $\{a_1, \ldots, a_m\} \subseteq X$. A rule with a nonempty head is satisfied by X if either $a_0 \in X$ or its body is not satisfied by X. A constraint is satisfied by X if its body is not satisfied by X.

Many facts from the state of the art [5, 25, 41] made ASP one of the most powerful knowledge representation paradigms, due to its strong expressive ability to model and represent many classical problems of knowledge representation. Although defeasible information cannot simply be represented easily, ASP offers the use of default negation in the body of rules, which makes it conceivable.

Furthermore, conditions allow for instantiating variables for collections of terms within a single rule. This is particularly useful for encoding conjunctions or disjunctions over arbitrarily many ground atoms, as well as for the compact representation of aggregates. Additionally, optimization in ASP is indicated via maximization and minimization statements that can extend a basic question whose answer set can be upgraded to an optimal one.

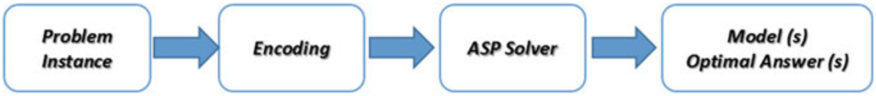


Fig. 2 Problem-solving steps using ASP

5 Reasoning Process Structure

This section describes the process structure of the proposed approach. It consists of two major phases: (1) an offline phase for analyzing and windowing the streaming data, and (2) an online phase to recognize activities using the same windowing technique.

We exploit the advantages of ASP to optimize extracted features from sensor streams. The goal of the optimization is to help in assigning weights to the online sensor streams with respect to their priorities.

Consequently, the concept is to maximize the total combined beliefs of those candidates (see Fig. 3). To evaluate the overall performance, we apply our approach to the HBMS dataset. This set consists of data from 22 sensors (switches and motion sensors).

Each sensor generates binary output only, 1 if it is activated, 0 otherwise. The dataset is annotated with five activities such as watching TV, going shopping, checking blood pressure, getting a drink, and preparing a meal. None of these activities occur simultaneously. Due to the binary nature of sensors, context values for these sensors provide simple events if dishes or cups are taken, devices are turned on or off. The lab had three virtual rooms (a living room, a kitchen, and a bedroom).

Activities were recorded over 18 days in the HBMS lab. The actors performed different activities over two hours, distributed over three activity periods per day.

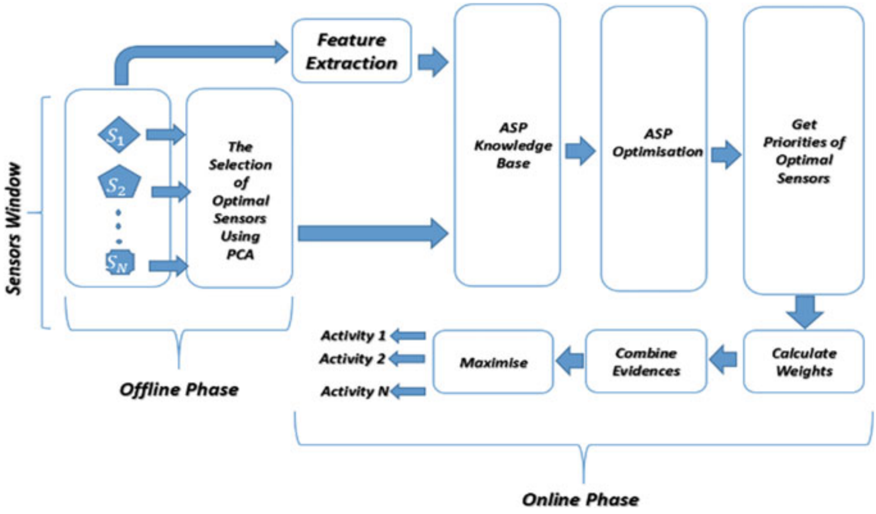


Fig. 3 Process structure of the activity recognition approach

5.1 Offline Phase

The basic concept of the offline phase is to analyze the dataset based on different features to find the optimal number of activated sensors. In order to employ efficient approach, we extracted the following features within two to three days of the given dataset: (a) the number of activations of each sensor, (b) the duration time of each sensor activation, (c) the duration time of each activity, (d) the time of performance of each activity, (e) the number of activated sensors for each activity, and (f) the location of the sensor.

The analysis is applied using information gain (IG) attribute evaluation [34]. The results showed that the number of activations for each sensor was the most relevant attribute.

Consequently, for each activity, we sorted the sensors based on their activation times and started out by choosing the first three, four, five, and six sensors in the window as optimal sensors.

Hence, using the support vector machine (SVM) based attribute ranking approach [16], we chose the window whose optimal number of sensors delivered the highest rank.

5.2 Online Phase

After the optimal sensors for each activity have been determined, the sensor data is collected in a window until the optimal sensors of one activity are activated (thus, the provided dynamic window size avoids the previously mentioned disadvantages). As soon as this happens, the following three steps are applied: (a) assignment of priority levels to each optimal sensors, (b) adjustment of sensors' belief, and (c) evidence combination of optimal sensors' beliefs.

5.2.1 ASP Optimization to Assign Priority Levels to Sensors

The assignment of priority levels is calculated based on three different features, which are categorized as follows: (1) the number of activations of each optimal sensor (which is the result of the offline phase); (2) the cost value, which is the performance of the measurement for each optimal sensor; (3) the sensor activation time. Consequently, sensors are represented in our knowledge base as follows:

```
sensor(Id).
sensor_time(SensorId,Time).
hist_importance(SensorId,ImportanceValue).
cost(SensorId,CostValue).
timing(SensorId,Duration).
```

```
user_current_time(Time).
```

The cost value (costValue) per optimal sensor is calculated as $1 - CV$, where the confidence value (CV) is the performance of the measurement for each optimal sensor. The importance value (`hist_imprtance`) is defined as the number of activations of each optimal sensor in the current window divided by the total number of activations. See Eqs. 2 and 3, where N is the total number of optimal sensors in the window, i is the index of the sensor in the window, $AllAct$ is the sum of all activations in the window, and s_k is the current sensor:

$$hist\_importance(s_k) = \frac{1}{AllAct} \sum_{i=1}^{N} f(s_{k_i}) \tag{2}$$

$$f(s_{k_i}) = \begin{cases} 1, & \text{if } s_k \; is \; ON, \\ 0, & \text{if } s_k \; is \; OFF. \end{cases} \tag{3}$$

Timing is considered with respect to the firing time of each optimal sensor. This feature is specified in the offline phase. Based on this, an optimization problem is to be solved using ASP and considering our three priority factors (in ascending order), identified by @:

```
1. maximize[sensor(X): hist_importance (X,Y)=Y @3].
2. minimize[sensor(X): cost (X,Y):
hist_importance (X,Z)=Y/Z @1].
3. maximize{sensor(X): timing (X,Y)=Y @2}.
```

Lines 1–3 contribute to optimization statements in descending order of significance. The optimization statement (line 1) gives the first priority to `hist_importance`, which should be maximized. Line 2 serves to minimize the cost of each optimal sensor, which has the last priority. Line 3 states that timing is our second priority, which should be maximized. The statements `maximize` and `minimize` are predefined optimization statements that are provided by ASP.

5.2.2 The Adjustment of Sensor Belief

After reading sensor data in the window, each sensor defines its belief (propagates) across the context values for the sensor via a mass function. The adjustment of sensors' belief is considered with respect to sensors' priority, which results from the sensor occurrence sequence in the answer set. Consequently, evidence propagation from context values is achieved using compatibility relations and evidential mapping [26, 29].

For illustration, at time t, the sensor mass functions produce "GetDrink," where θ is the frame of discernment:

$\{FridgeUsed = 1, notFridgeUsed = 0\} \rightarrow$
$\{GetDrink = 1, notGetDrink = 0\}$,
$\{CupUsed = 0, notCupUsed = 1\} \rightarrow$
$\{GetDrink = 0, notGetDrink = 0.8, \theta = 0.2\}$

"Prepare a meal," where θ is the frame of discernment:

$\{FridgeUsed = 1, notFridgeNotUsed = 0\} \rightarrow$
$\{Prepareameal = 1 notPrepareameal = 0\}$,
$\{MicrowaveUsed = 1, MicroUsed = 0\} \rightarrow$
$\{Prepareameal = 0.2, notPrepareameal = 0, \theta = 0.8\}$
$\{PlateUsed = 1, notPlatesUsed = 0\} \rightarrow$
$\{Prepareameal = 1, notPrepareameal = 0\}$,
$\{GroceriesUsed = 0, notGroceriesUsed = 1\} \rightarrow$
$\{Prepareameal = 1, notPrepareameal = 0\}$

After setting the priorities for each sensor, Eq. 4 is used to adjust the belief of each optimal sensor, where W is the weight of the sensor, Pr is the priority of the sensor, s_i is the current sensor, and Mu is the number of optimal sensors.

For example, in case of $Mu = 5$, the weights will be assigned as follows: the sensor with first priority will be weighted by 1, the sensor with second priority will be weighted by 0.80, the third by 0.60, the fourth by 0.40, and the fifth by 0.2:

$$W(s_i) = \frac{((Mu - Pr(s_i)) + 1)}{Mu}. \tag{4}$$

5.2.3 Evidence Combination

Dempster–Shafer theory can effectively represent uncertain and imprecise information. It has been widely used in the field of information fusion. But in multimodal sensor networks, there are often conflicting sensor reports due to the interference of the natural environment or other reasons.

It has been proven that classical Dempster–Shafer evidence theory cannot deal with the integration of conflict information effectively. If Dempster's combination rule is used directly to integrate evidence, with such conflicting cases, the results do not reflect reality. Many improved methods have been proposed to combine evidence.

As an example, Ali et al. [4] proposed a combination method by complementing the multiplicative strategy by an additional strategy. This method shows promising results for evidence combinations in comparison to other existing approaches.

The major components of evidence theory proposed by Dempster–Shafer are the frame of discernment θ and the basic probability assignment (BPA). The frame of discernment θ is the power set of the set of all possible mutually exclusive hypotheses (at most one of which is true), i.e., in our case, the set of all possible events (in the sense of operation sequences). BPA is a function $m : 2^{\theta} \rightarrow [0, 1]$ related to a proposition satisfying conditions (1) and (2) [38] (see Eqs. 5 and 6):

$$m(\phi) = 0, \tag{5}$$

$$\sum_{A \in \theta} m(A) = 1. \tag{6}$$

Here, A is any element of the frame of discernment, and ϕ refers to the empty set. Consequently, the whole body of evidence of one sensor is the set of all basic probability assignments greater than 0 under one frame of discernment.

The combination of multiple evidence defined in the same frame of discernment is a combination of the confidence level values based on the basic probability assignments (BPA). If there are two sensors, where each sensor has its body of evidence ms_1 and ms_2, these bodies of evidence are the corresponding BPA functions of the frame of discernment.

We have used the combination rule proposed by [4], since it provides more realistic results than the standard Dempster–Shafer rule when conflicting evidence from multiple sources is combined. Equation 7 shows how to calculate the combined probability assignment function:

$$m_{s_1} \oplus m_{s_1}(e) = \frac{1 - (1 - m_{s_1}(e)) * (1 - m_{s2}(e))}{1 + (1 - m_{s_1}(e)) * (1 - m_{s2}(e))}. \tag{7}$$

Equation 7 is used to combine all the beliefs of optimal sensors to maximize the occurrence of the best activity candidates, where m is the mass function, and e is the evidence.

6 Results Obtained

From the HBMS dataset, we extracted a subset consisting of 10-day observations including all five activities to determine the inference during the offline phase. The online phase was applied using the data from the other eight days. The proposed windowing technique was performed in both phases. In other words, the data is divided into 70% for training and 30% for testing.

Table 1 shows the results of our experiments with respect to accuracy and F-measure. Clearly, our overall accuracy is (96.76). Figure 4 shows the overall activity distribution in the dataset.

In order to measure the runtime behavior of the answer set programming approach, we performed several tests on an embedded platform: A pITX-SP[2] 1.6 plus board manufactured by Kontron. It was equipped with a 1.6-GHz Atom Z530 and 2 GB

[2] See http://www.kontron.de/.

Table 1 Overall performance C1: watch TV; C2: go shopping; C3: Check blood pressure; C4: get a drink; C5: prepare a meal

Class	Accuracy (%)	F-Measure (%)
C1	100	1
C2	95.4	0.94
C3	96.9	0.94
C4	91.3	0.89
C5	100	0.96

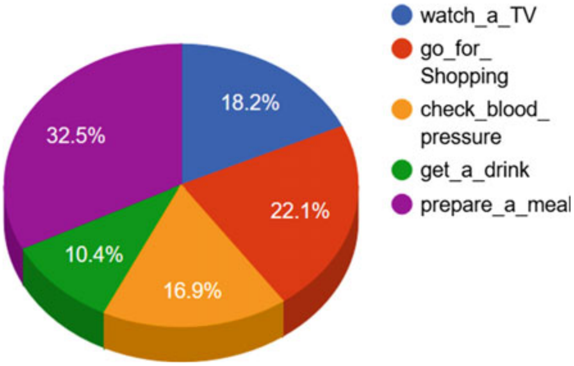


Fig. 4 Overall activity distribution as a percentage

RAM. For the evaluation, Clingo[3] was used as a solver for ASP. The average time to detect a complex event was 0.4 s.

7 Uncertainty Handling

Our approach convincingly shows that (a) it does not face the problem of the traditional Bayes's theorem for assigning the right priority probabilities, (b) it can respond in real time and run on embedded platforms, (c) it uses an evidence combination rule that can lead in the presence of highly convicting evidence to logical results, (d) ASP is an appropriate approach to dealing with incomplete knowledge and thus uncertainty. Little research has been proposed into the use of answer set programming (ASP) for reasoning under uncertainty in AAL environments.

The proposed approach can be used for any purposes simply by adjusting the knowledge base to the new context. This adjustment is not difficult, since only the facts have to be adapted but not the rules. ASP supports a number of arithmetic functions that are evaluated during grounding. Therefore, the major reasoning-under-uncertainty approaches can be implemented in ASP.

[3] See http://potassco.sourceforge.net/.

Also, different optimization problems have the same formulations to be represented as logic programs. Therefore, ASP provides this possibility using maximize and minimize statements. Additionally, the intuitive semantics of ASP programs avoid the complex representation of optimization problems that are based on other standard approaches, for instance simulated annealing, genetic algorithms, and artificial neural networks. Moreover, the syntax of logic programs offers the possibility of fast implementation of different complex problems that might be difficult to represent in any other form.

Furthermore, constraints play an important role in ASP, because adding a constraint to a logic program P affects the collection of stable models of P in a very simple way. It eliminates the stable models that violate the constraint. This feature can be applied to activity recognition by the definition of the constraints in the environment.

8 Conclusion

Activity recognition requires a detailed analysis and understanding of the domain in which the activities to be recognized occur. Within the scope of this chapter we have shown that combining logic programming (ASP) and Dempster–Shafer theory is a solid basis for implementing a powerful tool to detect complex activities. In particular, the ASP paradigm proved to be suitable for activity recognition systems due to its inherent knowledge representation and optimization capabilities. In addition, we were able to improve our technique's accuracy by assigning weights to sensor events with respect to different spatial and temporal features. Altogether, this concept allowed us to come up with a methodology that improves the handling of uncertainty. With respect to other approaches, a disadvantage of the one presented here is the fact that it needs previously collected knowledge about users and sensors. This chapter is mainly concerned with the development of effective activity recognition systems for complex event detection under uncertainty. It discusses the consideration of uncertainty in the framework of complex event detection involving multiple sensors. Moreover, we addressed diverse state-of-the-art approaches for complex event detection, the advantages and disadvantages of each technique, and a comprehensive evaluation about the performance of the methodologies for handling uncertainty. In our future work, we will test the proposed reasoning approach using other international datasets and increase the number of activities to be able to compare the proposed approach with other state-of-the-art approaches.

Acknowledgements This work was funded the by Klaus Tschira Stiftung GmbH, Heidelberg

References

1. Akdemir, U., Turaga, P., Chellappa, R.: An ontology based approach for activity recognition from video. In: Proceedings of the 16th ACM international conference on Multimedia, pp. 709–712. ACM (2008)
2. Al Machot, F., Kyamakya, K., Dieber, B., Rinner, B.: Real time complex event detection for resource-limited multimedia sensor networks. In: 2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 468–473. IEEE (2011)
3. Al Machot, F., Tasso, C., Dieber, B., Kyamakya, K., Piciarelli, C., Micheloni, C., Londero, S., Valotto, M., Omero, P., Rinner, B.: Smart resource-aware multimedia sensor network for automatic detection of complex events (2011)
4. Ali, T., Dutta, P., Boruah, H.: A new combination rule for conflict problem of dempster-shafer evidence theory. Int. J. Energy Inf. Commun. **3**(1), 35–40 (2012)
5. Baral, C.: Knowledge Representation, Reasoning and Declarative Problem Solving. Cambridge University Press, Cambridge (2003)
6. Baral, C., Gelfond, G., Son, T.C., Pontelli, E.: Using answer set programming to model multi-agent scenarios involving agents' knowledge about other's knowledge. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1, pp. 259–266. International Foundation for Autonomous Agents and Multiagent Systems (2010)
7. Beal, M.J.: Variational algorithms for approximate Bayesian inference. Ph.D. thesis, University of London (2003)
8. Bobillo, F., Straccia, U.: fuzzydl: an expressive fuzzy description logic reasoner. In: FUZZ-IEEE, pp. 923–930 (2008)
9. Brain, M., De Vos, M.: Answer set programming–a domain in need of explanation. In: Exact08: International Workshop on Explanation-aware Computing (2008)
10. Brendel, W., Fern, A., Todorovic, S.: Probabilistic event logic for interval-based event recognition. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3329–3336. IEEE (2011)
11. Chiang, S.-Y., Kan, Y.-C., Tu, Y.-C., Lin, H.-C.: Activity recognition by fuzzy logic system in wireless sensor network for physical therapy. In: Intelligent Decision Technologies, pp. 191–200. Springer (2012)
12. Dezert, J.: Foundations for a new theory of plausible and paradoxical reasoning. Inf. Secur. **9**, 13–57 (2002)
13. Dietterich, T.G., Ashenfelter, A., Bulatov, Y.: Training conditional random fields via gradient tree boosting. In: Proceedings of the twenty-first international conference on Machine learning, p. 28. ACM (2004)
14. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets*. Int. J. Gen. Syst. **17**(2–3), 191–209 (1990)
15. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: Owl 2: the next step for owl. Web Semant. Sci. Serv. Agents World Wide Web **6**(4), 309–322 (2008)
16. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
17. Jones, R.E., Connors, E.S., Endsley, M.R.: Incorporating the human analyst into the data fusion process by modeling situation awareness using fuzzy cognitive maps. In: 12th International Conference on Information Fusion, 2009. FUSION'09, pp. 1265–1271. IEEE (2009)
18. Katz, S.: Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. J. Am. Geriatr. Soc. **31**(12), 721–727 (1983)
19. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: a review of the state-of-the-art. Inf. Fusion **14**(1), 28–44 (2013)
20. Klein, L.A.: Sensor and Data Fusion Concepts and Applications. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham (1999)
21. Klein, L.A.: Sensor and Data Fusion: A Tool for Information Assessment and Decision Making, vol. 324. Spie Press, Bellingham (2004)

22. Kofod-Petersen, A., Cassens, J.: Using activity theory to model context awareness. In: Modeling and Retrieval of Context, pp. 1–17. Springer (2006)
23. Kohlas, J., Monney, P.-A.: A Mathematical Theory of Hints: An Approach to the Dempster-Shafer Theory of Evidence, vol. 425. Springer Science & Business Media, Berlin (2013)
24. Liang-zhou, C., Wen-kang, S., Yong, D., Zhen-fu, Z.: A new fusion approach based on distance of evidences. J. Zhejiang Univ. Sci. A **6**(5), 476–482 (2005)
25. Lifschitz, V.: Answer set programming and plan generation. Artif. Intell. **138**(1), 39–54 (2002)
26. Lowrance, J.D., Garvey, T.D., Strat, T.M.: A framework for evidential-reasoning systems. In: Classic Works of the Dempster-Shafer Theory of Belief Functions, pp. 419–434. Springer (2008)
27. Lukasiewicz, T., Straccia, U.: Managing uncertainty and vagueness in description logics for the semantic web. Web Semant. Sci. Serv. Agents World Wide Web **6**(4), 291–308 (2008)
28. McCallum, A.: Efficiently inducing features of conditional random fields. In: Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, pp. 403–410. Morgan Kaufmann Publishers Inc. (2002)
29. McKeever, S.: Recognising situations using extended Dempster-Shafer theory. Ph.D. thesis, National University of Ireland, Dublin (2011)
30. Meditskos, G., Dasiopoulou, S., Kompatsiaris, I.: Metaq: a knowledge-driven framework for context-aware activity recognition combining sparql and owl 2 activity patterns. Pervasive Mob. Comput. (2015)
31. Menschner, P., Prinz, A., Koene, P., Köbler, F., Altmann, M., Krcmar, H., Leimeister, J.M.: Reaching into patients homes-participatory designed aal services. Electron. Mark. **21**(1), 63–76 (2011)
32. Michael, J., Grießer, A., Strobl, T., Mayr, H.C.: Cognitive modeling and support for ambient assistance. In: Information Systems: Methods, Models, and Applications, pp. 96–107. Springer (2012)
33. Murphy, K.P.: Dynamic bayesian networks: representation, inference and learning. Ph.D. thesis, University of California (2002)
34. Novakovic, J.: Using information gain attribute evaluation to classify sonar targets. In: 17th Telecommunications Forum TELFOR, pp. 24–26 (2009)
35. Okeyo, G., Chen, L., Wang, H., Sterritt, R.: Dynamic sensor data segmentation for real-time knowledge-driven activity recognition. Pervasive Mob. Comput. **10**, 155–172 (2014)
36. Piyathilaka, L., Kodagoda, S.: Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features. In: 2013 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 567–572. IEEE (2013)
37. Pruteanu-Malinici, I., Carin, L.: Infinite hidden markov models for unusual-event detection in video. IEEE Trans. Image Process. **17**(5), 811–822 (2008)
38. Rice, J.: Mathematical Statistics and Data Analysis. Cengage learning, Boston (2006)
39. Sentz, K., Ferson, S.: Combination of Evidence in Dempster-Shafer Theory, vol. 4015. Citeseer (2002)
40. Smets, P.: The combination of evidence in the transferable belief model. IEEE Trans. Pattern Anal. Mach. Intell. **12**(5), 447–458 (1990)
41. Soininen, T.: An Approach to Knowledge Representation and Reasoning for Product Configuration Tasks. Finnish Academies of Technology, Espoo (2000)
42. Stoilos, G., Simou, N., Stamou, G., Kollias, S.: Uncertainty and the semantic web. Intell. Syst. IEEE **21**(5), 84–87 (2006)
43. Vail, D.L., Veloso, M.M., Lafferty, J.D.: Conditional random fields for activity recognition. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, pp. 235. ACM (2007)
44. Van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: Proceedings of the 10th international conference on Ubiquitous computing, pp. 1–9. ACM (2008)
45. Walczak, B., Massart, D.: Rough sets theory. Chemometr. Intell. Lab. Syst. **47**(1), 1–16 (1999)

46. Wang, X., Ji, Q.: Incorporating contextual knowledge to dynamic bayesian networks for event recognition. In: Pattern Recognition (ICPR), 2012 21st International Conference on, pp. 3378–3381. IEEE (2012)
47. Welch, G., Bishop, G.: An introduction to the kalman filter (1995)
48. Yen, J.: Generalizing the dempster–shafer theory to fuzzy sets. In: Classic Works of the Dempster-Shafer Theory of Belief Functions, pp. 529–554. Springer (2008)
49. Zhang, W., Chen, F., Xu, W., Cao, Z.: Decomposition in hidden markov models for activity recognition. In: Multimedia Content Analysis and Mining, pp. 232–241. Springer (2007)

Estimation of Infection Force of Hepatitis C Virus Among Drug Users in France

Selain Kasereka, Yann Le Strat and Lucie Léon

Abstract The spread of diseases is a dynamic and complex phenomenon. In the world, they are due to misery and poverty. To understand epidemiological systems is essential for governments. Since modeling simplifies reality, it is an excellent method. The hepatitis C virus (HCV) infection is common worldwide, and injection drug use remains the major mode of transmission of the disease, especially because of equipment sharing. Consequently, it is crucial to monitor the HCV transmission dynamics over time and to assess the effect of harm reduction measures. The aim of this work is to estimate the force of infection of hepatitis C from two national cross-sectional epidemiological surveys conducted in 2004 and 2011 by the French Institute for Public Health Surveillance and its partners in a drug user population in France. HCV prevalence was estimated according to age and calendar time through fractional polynomials adjusted or not to the HIV serological status and to injected drug users or oral drug users in general. The force of infection was modeled according to an SIS (susceptible–infected–susceptible) compartmental model using ordinary differential equations (ODE) and as a function of the derivative of the prevalence function depending on age, time, HIV serological status, and having injected at least once in their life, from 2000 to 2020. Our model was applied on real and simulated surveys using *R* and *Stata* software. The results show that HCV prevalence and the force of infection are linked to age and time, and are very high for drug users who injected at least once in their life and who are simultaneously HCV and HIV infected. Based on this model, we estimated that HCV incidence will continue to decline over the following years. Currently in France, there is no cohort study of the HCV among drug users. The only way to estimate HCV incidence in the French population is to use

S. Kasereka (✉)
University of Kinshasa, Kinshasa, Democratic Republic of the Congo
e-mail: selain.kasereka@unikin.ac.cd

Y. Le Strat · L. Léon
Santé publique France, French National Public Health Agency, Saint-Maurice, Ile de France, France
e-mail: yann.lestrat@santepubliquefrance.fr

L. Léon
e-mail: lucie.leon@santepubliquefrance.fr

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_15

the only existing two national cross-sectional surveys. Our work provides guidance for researchers to compare several cross-sectional epidemiological surveys among drug users and proposes an alternative method to estimate the force of infection among drug users from cross-sectional surveys in the absence of a cohort.

1 Introduction

Mathematical modeling and computer-based simulation of epidemics/endemics are a key to achieving a better understanding of the spreading of infectious diseases in order to eradicate them in a population. As a rule, the idea is to try to understand the flow of people who become infected at a given time t. These steps become even more realistic when we use data resulting from epidemiological surveys. An epidemiological survey is aimed at discovering facts, the improvement of knowledge, or the resolution of doubts and problems. Classically, in epidemiology, we make a distinction between descriptive studies that examine the frequency and the distribution of health parameters and/or of risk factors in populations and the etiological studies that consist in comparing some groups of subjects in order to highlight an association between an exposition and a pathology.

In these etiological studies, we find the cohort surveys (these are of great interest but are very costly and difficult to implement on a large scale). The transversal surveys, those dealt with in the framework of our research, are characterized by their easy organization and their more sensible cost but suffer from the difficulty in establishing a temporal relationship between the risk factor and the pathology. The case-control surveys consist in comparing the preceding exposition among the sick people (the cases) with the exposition of control samples. They are particularly interesting for the study of rare pathologies.

When we use data resulting from several transversal surveys, the issue of the interest indicator estimate according to time arises regularly. That estimate is not easy to find, for its calculation requires one to take into account the survey plans used to constitute the samples, the size of surveys, and the population composition, which can be different from one survey to another.

To take up this challenge, we have organized this work in two distinctive parts. In the first part, we have generated a population of drug users (DU) during a seven-year period in which the hepatitis C virus spreads and diffuses according to an SIS (susceptible–infected–susceptible) compartmental model. In the second part, the hepatitis C virus (HCV) prevalence was estimated according to age and calendar time through fractional polynomials adjusted or not to the HIV serological status and to injected drug users or oral drug users in general. The force of infection was modeled using ordinary differential equations (ODE) and as a function of the derivative of the prevalence depending on age, time, HIV serological status, and having injected at least once in their life, from 2000 to 2020. We have worked on real and simulated surveys at the same time.

This work is organized in five part. We introduce general notions on hepatitis C in the first part before dealing with some generalities on the transversal epidemiological surveys in the second part. The third and fourth parts present respectively the materials and methods used and the results obtained before their discussion in the fifth part.

2 Generalities on Hepatitis C

2.1 Introduction

Hepatitis C is a contagious liver disease that results from an infection by the virus of hepatitis C. It manifests itself with a seriousness that can range from a benign form lasting a few weeks to a serious illness that will become permanent. The HCV is one of the viruses that infect the liver most frequently. Infection by the HCV is a major problem of public health around the world. The countries having a high chronic infection rate are Egypt (15%), Pakistan (4.8%), and China (3.2%). Every year, 3–4 million people are infected by the HCV around the world. About 150 million individuals are chronic carriers and run the risk that their hepatitis will evolve toward liver cirrhosis and/or liver cancer. Every year, around the world 350,000 to 500,000 people die of hepatic pathologies related to hepatitis C [20]. In France, about 600,000 people were carriers of the HCV (1% of the population) in 2004, with more than one-third of them not knowing about their status. In the literature, it is estimated that between 200,000 and 300,000 people have been infected by blood transfusions [20].

The transmission of the HCV is carried out most of the time by exposure to infected blood, for instance during blood transfusion, blood products of a contaminated graft, injections carried out with contaminated syringes or wounds by a needle prick in a treatment environment, use of injectable drugs, or at birth by a mother infected with hepatitis C.

The natural history of infection by the HCV proceeds in three steps:

1. The contamination by the HCV ends in an acute hepatitis (very recent infection), not visible most of the time. About 80% of individuals infected by the HCV are asymptomatic, that is, they manifest no symptoms. Among people manifesting symptoms, we observe fever, fatigue, loss of appetite, nausea, vomiting, abdominal pain, dark coloration of urine, grayish coloration of excrement, articular pain and/or icterus (jaundice of the skin and the white of the eye). Most infected individuals remain chronic carriers of this virus.
2. The persistence of the viral infection leads to the appearance of chronic hepatitis lesions and the development of a fibrosis that can lead to cirrhosis several decades after the date of contamination.
3. Clinical complications (hepatocellular carcinoma, notably) occur essentially at the stage of cirrhosis and are responsible for mortality linked to this infection.

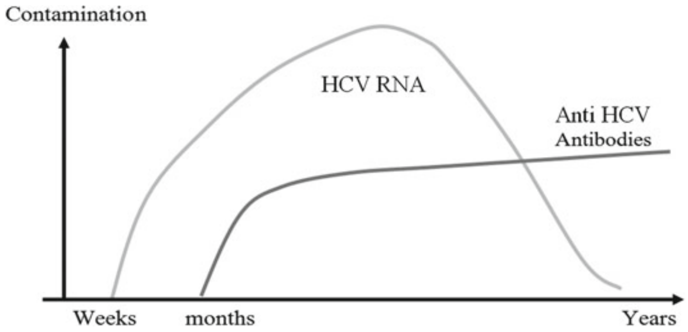


Fig. 1 Evolution of markers (HCV RNA and anti-HCV antibodies) in case of very recent infection [9]

2.2 Very Recent Infection

As illustrated in Fig. 1, the diagnosis of a very recent infection begins with the detection in the serum of a marker, the HCV RNA, 7–21 days after the date of contamination. The rise of serumal transaminases, often 10 times above the superior limit of the average, arises beyond the 15th day, often beyond 4 weeks. Clinical symptoms (often an icterus), if observed, appear 2–12 weeks after the date of contamination and disappear quickly. The anti-HCV antibodies appear in the serum 20–150 days after contamination. Seroconversion is important for the diagnosis of a very recent infection, the RNA positivity not allowing one to distinguish a very recent infection from a chronic one. Healing from a very recent infection is defined by the spontaneous disappearance of the detection of the HCV RNA in the serum. In this case, the HCV RNA becomes undetectable, after 19 months in one case or two, and remains in that state the rest of the time. Conversely, the persistent positivity of the HCV RNA in the serum reveals the evolution toward a chronic infection [1].

2.3 Chronic Infection

Chronic infection is attested by the presence of HCV RNA in the serum long after the onset of infection, as shown in Fig. 2. The HCV RNA remains constantly detectable throughout the evolution. Sick people having a positive serology (detection of anti-HCV antibodies) and a search of HCV RNAs constantly negative are probably healed. The absence of HCV RNAs detectable in the liver reveals a complete viral elimination.

Table 1 presents, depending on the detection or not of the two biological markers (HCV RNA and anti-HCV antibodies), four different situations: chronic infection, very recent infection, healing, and absence of contamination (the individual has never been in contact with the virus).

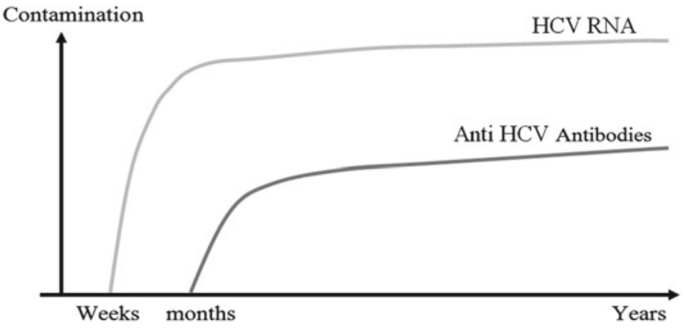


Fig. 2 Evolution of markers (HCV RNA and anti-HCV antibodies) (antibodies anti-HCV) in case of chronic infection [9]

Table 1 Classification according to detection of the RNA markers of HCV and anti-HCV antibodies

	Detection of the HCV RNA	Nondetection of the HCV RNA
Detection of anti-HCV antibodies	Chronic infection	Healing (spontaneous or after treatment)
Nondetection of anti-HCV antibodies	Very recent infection	Noncontaminated

About 75–85% of newly infected people contract the chronic illness and among chronic carriers, 60–70% will suffer from a chronic hepatic disease; 5–20%, will be affected by cirrhosis, and 1–5% will die of liver cancer. For 25% of liver cancer cases, the underlying cause of the cancer is hepatitis C [9].

2.4 *Diagnosis*

Common methods of antibody detection are unable to tell the difference between chronic infections and recent infections. The presence of antibodies against the hepatitis C virus indicates that an individual is or has been infected. We resort to the RIBA test (recombinant immunoblot assay) and viral RNA detection to confirm the diagnosis. We pinpoint the diagnosis of chronic infection when anti-HCV antibodies are present in the blood for more than six months. As for recent infections, the diagnosis must be confirmed by a further test. We often use specialized tests to assess whether patients are affected by a hepatic disease such as cirrhosis or liver cancer.

2.5 Treatment

The treatment of people infected by the HCV stretches from 24 to 48 weeks depending on the genotypes:

- In case of genotype 2 or 3: 24 weeks with a good reaction to the treatment (70% success).
- In case of genotype 1: 48 weeks with a good reaction to the treatment (less than 1 person out of 2).

The treatment of infected people ensures a viral nonreplication of the virus [5]. Six months after treatment, patients are tested once again to see whether viremia is absent. Let us note that older patients recover less frequently from hepatitis C [9].

The healing of infected people can occur spontaneously: the infected person eliminates the virus and recovers spontaneously from infection (in other words, without treatment). This fact can be due to the genes of that person. It is what we will call seroreversion in the following pages. These cases are very rare [15]. Healing can also be posttherapeutic [10]: here the subject does not exhibit the virus after a 24–48 week treatment.

2.6 Transmission of Hepatitis C Among Drug User Injectors

The risk of HCV transmission is more important in case of a deep wound, a prick with a hollow needle, and notably a needle that was used in intravenous or intraarterial injection. The HCV is transmitted through blood (transfusion, use of drug injection, organ transplant). Drug injection users are a high-risk population, particularly for the transmission of hepatitis C [2]. A borrowed syringe and shared preparation equipment are major factors of risk in the transmission of this virus. Needles and syringes have the highest potential contamination by the HCV because of the direct contact with blood during intravenous injection.

A contaminated subject produces HCV RNA, which becomes detectable after about 6–10 days. We will have to wait a few weeks (about 60 days) for the anti-HCV antibodies produced by that person be detected. To detect infection, a screening test is essential. It detects the presence anti-HCV antibodies. If this test is positive, a viral charge test is carried out in order to quantify the RNA rate in the blood [9].

3 Transversal Epidemiological Surveys

In broad outline, two principal types of epidemiological studies are generally distinguished and most often carried out: descriptive surveys and etiological (analytic) studies. In most cases, cross-sectional surveys are derived from observation. Contrary to therapeutic essays or the studies assessing the actions of public health, the distribution of exposure in the studied population is not controlled by the investigator and is also not randomized. Descriptive surveys are principally aimed at providing sanitary statistics in the populations. They study the frequency and the distribution of health indicators or risk factors and their variation over time, geographic zones, and population groups. From these observations, they make it possible to state hypotheses on the risk factors of diseases.

Etiological surveys, which concern the framework of our work, examine and analyze the relationships between exposure to one or several risk factors and state of health. They are always comparative (two different groups are compared, in the presence of either the disease or the risk factor). A risk factor is a characteristic associated with a higher probability of contracting the disease. These surveys can be implemented to check and make precise hypotheses formulated from the results of descriptive studies or other types of studies (animal or toxicological, etc.), concerning the relationships between exposure factor and diseases. They are sometimes realized on an exploratory basis to identify the effects of an exposure factor or the risk factors of a disease. In our research, we are interested in transversal epidemiological surveys.

In transversal surveys, a sample is taken from the population without necessarily being selected on the exposure or disease. This type of surveys allows that information can be collected on the disease and exposure at the same moment. It is common in transversal surveys that the information is also collected on past exposure or other past health events. When in a transversal survey information on exposure is collected, one speaks of a retrospective transversal survey.

Transversal surveys have the advantage of enabling one to collect, at the same time, information on different factors of exposure and on different health events. Their implementation is relatively easy, and their cost is less significant than for other surveys. These surveys have, however, some methodological drawbacks. They are subject, as for the cohorts, to biases of participation, since they are based on voluntary participation. They are also subject to biases of memory, classically described for case-control studies, when exposure is reconstituted in a retrospective way.

4 Materials and Methods

4.1 Materials

4.1.1 Population of Study

Drug users (DU) are the population most widely affected by hepatitis C because of their high-risk practices, particularly the sharing of materials (syringes, needles, etc.). In France, drug consumption is the principal mode of contamination. Every year, among new contaminations, 70% are related to drugs [16, 17, 19]. For DU, contaminations occur early, at the first injections: from the very moment of the introduction of the injection, users take risks. They have little accurate information [19]. Drug consumption is illegal in France, which places the high-risk population in situations of precariousness and which causes them to hide from other people, making them often inaccessible to investigating officers.

4.1.2 Coquelicot 2004 and 2011 Surveys

The Coquelicot survey is a multicentric survey carried out in five towns (Lille, Strasbourg, Paris, Bordeaux, and Marseille) and two departments (Seine-et-Marne and Seine-Saint-Denis). A random sample of DU was constituted in almost all of the specialized services for DU in these towns and departments, according to a two-degree poll. A list of all the services open on a half-day, constituting in this way the sampling frame, has been constructed in order to pull random the services and half-days, according to a simple poll random.

In the specialized services for DU in towns and departments cited below, investigation officers have enlisted in a random way the first DU who shows up. Others DU have been questioned according to a sampling interval adapted to the size of the specialized service for DU, in order to prevent the investigator to choose the DU to investigate, because this one can introduce selection bias. The generalized method of weight sharing (MGPP) has made it possible to take into account the heterogeneity of the DU as regards their attendance of the specialized services [11, 12].

Coquelicot 2004

In 2004, 1462 people who had used injectable drugs and/or drugs by inhalation at least once in their lifetime took part in the Coquelicot survey. The objective of this survey was to value the seroprevalence of HIV and HCV among DU and to describe their high-risk behavior and practices. The participants agreed to complete a questionnaire on socio-compartmental aspects, and biological samples were collected by self-sampling of blood at the finger level for 79% of them [11].

Drug users are primarily men (74%), and their average age is 35.6 years for men and 34.5 years for women. They are often inactive (at the time of the survey, 65% declared themselves unemployed) and experience precarious life conditions (only 45% have a stable accommodation and 19% live in the street or are squatters).

Among the 1462 participants, 10.8% were HIV positive and 59.8% were HCV positive, 10.2% were simultaneously HCV and HIV infected. While HIV seroprevalence is almost zero among DU of less than 30 years, it is 28% for HCV among these same people (less than 30 years) and reaches 71% for those 40 years and older. The HIV seroprevalence varies according to towns (1% in Lille, 10.9% in Paris, and 31.5% in Marseille). On the other hand, there is no significant difference for the HCV [11].

Among those surveyed, 71% had received an opiate-based treatment of substitution in the previous six months (57% with Subutex and 36% with methadone). The principal illicit psychoactive products consumed by the DU (during the last month) were crack (30%), cocaine (27%), heroine (20%), and ecstasy (12%). Those less than 30 years old reported more frequent consumption of hallucinogenic substances and had resorted more often to cocaine (40%), ecstasy (26%), amphetamines (14%), LSD (12%), and other hallucinogens (11%). Intravenous injection was practiced by 70% of DU who participated in the survey, at an average age of 20.4 years [11].

Coquelicot 2011

This second edition of the Coquelicot survey was realized in 2011 among a sample of 1568 DU taken in 122 specialized services. Almost all the specialized services for DU in towns and departments contacted agreed to take part in the survey. Altogether, 25% of DU were recruited in services belonging to some CAARUD (Welcoming and Caring Centers for the Reduction of Risk among DU), 70% in services belonging to CSAPA (Caring and prevention health centers in addiction studies), 1.5% in housing centers, and 3.7% in other types of associated structures.

The rate of participation in the survey was 75%, and among the respondents, 92% agreed to self-sampling of blood. For the analysis of seroprevalence, a total of 1418 subjects were registered. The nonrespondent sample is similar to the sample of respondents in terms of age and gender [12].

The DU population is essentially masculine (79% men) and 39 years old on average (16% of DU are less than 30 years old). More than two thirds (70%) declare their level of education was a secondary school, 6% declare a primary level, and 24% attained a level higher than the baccalaureate. More than three-fourths (79%) of the DU did not work during the period of the survey. More than half of the DU live in insecure situations: they do not live in their own house, at their spouse's house, or at their parents' house. Among them, 18% face very great insecurity, for they are squatters or live in the street. The majority of DU (57%) have been in prison [12].

4.1.3 Simulated Data

Based on real data from Coquelicot 2004 and 2011, we have generated a population having the same distribution as the real data. This population named *pop0* had a size of 21300 individuals, and the global HCV prevalence was 62%. Consequently, the generated population constituted our simulated population for 2004.

Once population *pop0* was established, we made it evolve by simulations, taking into account several factors explained below, according to an SIS-type compartmental model (introduced later in our work, where the HCV spreads and diffuses during a seven-year period) and considering the varying size of the population. A new population *pop1* of size 21996 individuals was obtained seven years later, with a global prevalence of 53% (Table 2).

Table 2 Different variables of the population studied

Variables	Labels
id	Identifier of each individual in the population
vhc	HCV status for each individual (yes/no)
age	Age of each individual
vih	HIV status for each individual (yes/no)
injvie	Injector status for each individual (yes/no)

4.2 Methods

4.2.1 Generation of Population

To generate the initial DU population, named *pop0*, of size n and having the same characteristics as those of the survey in terms of age distribution, HCV status, HIV status, the fact of being injector or having high-risk practices, we used the results of the survey of Coquelicot 2004. The following steps were taken:

- The individuals were duplicated according to their poll weight multiplied by 50; a number arbitrarily chosen for having a very large population.
- A random number was assigned to each individual.
- Then the individuals were arranged in order according to this random number;
- Lastly, the first n individuals were selected to constitute our newly generated population *pop0*.

The DU population benefits from proposed services at different moments of the day by different welcoming structures devoted to their use. Based on the design of the Coquelicot survey, we suppose that:

- 80 structures have proposed 10 services per each open half-day.
- The structures were opened from Monday to Friday, that is, 5 days a week.
- The survey lasted 8 weeks.
- The number of attendance of specialized services of drug users follows a negative binomial distribution of mean μ = 3 and variance θ =10.

We associate a service to each individual during a given half-day in a given structure. One thousand samples (2000 individuals in each simple) were generated according to a three-degree screening called *place-moments*:

- At the first degree: 20 structures drew lots according to a simple random poll.
- At the second degree: 25 half-days drew lots according to a simple random poll in each structure sampled.
- At the third degree: 4 services drew lots according to a simple random poll in each half-day sampled.

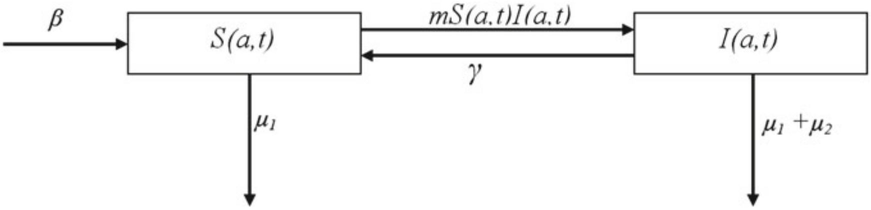


Fig. 3 SIS model describing the transmission of the HCV

4.2.2 Compartmental Model for the HCV

The spreading of an infectious agent in a population is a dynamic phenomenon: the numbers of infected individuals and uninfected individuals evolve in time, according to contacts. Such a phenomenon can be studied by separating the individuals according to their status in the compartments [3, 4, 6, 13]. We call susceptible the fraction of the population that is uninfected but can potentially come in contact with a virus, and we denote by S the compartment containing those individuals. At an instant t, compartment S has $S(t)$ individuals. In the same way, we call infected the individuals infected by the virus. The compartment containing this fraction of the population will be denoted by I and will contain $I(t)$ individuals at time t. Individuals can then move from one state to another and change compartments. This phenomenon can then be modeled by differential equations and its behavior determined through the numerical solution of these equations [7, 8, 13].

In a diagram, an SIS model can be represented as shown in Fig. 3.

In the figure:

- β: rate of new DU;
- t: time;
- γ: rate of seroreversion;
- μ_1: rate of mortality (not related to infection);
- μ_2: rate of mortality related to infection;
- m: rate of contact;
- S: fraction of susceptible people;
- I: fraction of infectious people;
- a: age (in years).

It should be noted that several models can be used to model the spreading of hepatitis C. for instance, a compartment named E containing people in the period of incubation could be added, leading to an SEIS model.

The hypotheses to be considered for the SIS model in our case are the following:

1. At the beginning, two groups of individuals: susceptible people S and infectious people I (who are infected and can transmit the disease). A person is considered infectious if he manifests anti-HCV antibodies.

2. As time goes by, the group of infectious people can be added to by the group of susceptible people S according to a force of infection $mI(a, t)$, m being the contact rate, a and t representing respectively the age and the time.
3. The infectious people I can become susceptible again according to the rate of seroreversion γ.
4. New people can be added to the group S at a rate β.
5. In group S, people can die (natural causes) at a rate μ_1.
6. In group I, people can die (natural causes) at the rate μ_1 or experience death linked to the infection by the HCV at a rate μ_2.

This model can be represented by the following system of differential equations:

$$\begin{cases} \dfrac{dS(a,t)}{d(a,t)} = -mI(a,t)S(a,t) - \mu_1 S(a,t) + \beta S(a,t) + \gamma I(a,t), \\ \\ \dfrac{dI(a,t)}{d(a,t)} = mI(a,t)S(a,t) - (\mu_1 + \mu_2)I(a,t) - \gamma I(a,t). \end{cases} \tag{1}$$

4.2.3 Model Parameters

The study period t considered in this work was seven years with an annual time step. That period was inspired by two epidemiological surveys (Coquelicot 2004 and 2011), from which we took the data used in our research. The parameter values were essentially taken from the literature. Seroreversion γ was set at 0.001 [15]. The mortality rate μ_1 (not linked to infection) among DU was set at 1.5% [16]. A recent article [14] considers the all-cause mortality rate among HCV infected drug users equal to 1.85% (among the ARN+, 1.75% and among the ARN−, 2.05%). The 7.5% death rate among the ARN+ was due to liver disease, against 2.3% among the ARN−. The death rate was 2.35% person-years among injector drug users (IDU) [17]. Given that (μ_1 + μ_2) represents the all-cause mortality rate among the HCV infected DU, and that μ_1 is set at 1.5%, μ_2 was obtained by the difference: 1.85–1.5%. The proportion of people joining the DU population (β) was set at 3% per year based on the Coquelicot survey data. We have considered that the DU age is from 18 to 34 years old. The average age of new IDU was 26 years [18].

4.2.4 Estimation of Infection Force (Force of Infection)

Modeling of infection force (IFO) according to age: we have estimated the IFO by expressing it according to the prevalence derivative using fractional polynomials [21]. Two link functions were used: *complementary log-log* and *logit*.

Starting from differential equations (1), we define the IFO by $\lambda(a, t) = mI(a, t)$. System (1) becomes then

$$\begin{cases} \dfrac{dS(a,t)}{d(a,t)} = -\lambda(a,t)S(a,t) - \mu_1 S(a,t) + \beta S(a,t) + \gamma I(a,t), \\ \\ \dfrac{dI(a,t)}{d(a,t)} = \lambda(a,t)S(a,t) - (\mu_1 + \mu_2)I(a,t) - \gamma I(a,t). \end{cases} \tag{2}$$

From (2) we have

$$\begin{cases} \dfrac{\partial}{\partial a}S(a,t) + \dfrac{\partial}{\partial t}S(a,t) = -\lambda(a,t)S(a,t) - \mu_1 S(a,t) + \beta S(a,t) + \gamma I(a,t) \\ \\ \dfrac{\partial}{\partial a}I(a,t) + \dfrac{\partial}{\partial t}I(a,t) = \lambda(a,t)S(a,t) - (\mu_1 + \mu_2 + \gamma)I(a,t). \end{cases} \tag{3}$$

From (3) we obtain

$$\lambda(a,t) = \frac{\beta S(a,t) - \dfrac{\partial}{\partial a}S(a,t) - \dfrac{\partial}{\partial t}S(a,t) - \mu_1 S(a,t) + \gamma I(a,t)}{S(a,t)} \tag{4}$$

$$= \frac{(\beta - \mu_1)S(a,t) - \dfrac{\partial}{\partial a}S(a,t) - \dfrac{\partial}{\partial t}S(a,t) + \gamma I(a,t)}{S(a,t)}. \tag{5}$$

By definition, $S(a,t) = 1 - I(a,t)$.

From (5) we deduce

$$\lambda(a,t) = \frac{(\beta - \mu_1)(1 - I(a,t)) - \dfrac{\partial}{\partial a}(1 - I(a,t)) - \dfrac{\partial}{\partial t}(1 - I(a,t)) + \gamma I(a,t)}{1 - I(a,t)} \tag{6}$$

$$= \frac{\dfrac{\partial}{\partial a}I(a,t) + \dfrac{\partial}{\partial t}I(a,t) + (\beta - \mu_1) - (\beta - \mu_1 - \gamma)I(a,t)}{1 - I(a,t)}. \tag{7}$$

By modeling prevalence with a *cloglog* link, we have the following relation:

$$\begin{aligned} &\log(-\log(1 - P(Y = 1|a,t))) = \eta(a) + ct + \alpha \\ &\Rightarrow \log(1 - P(Y = 1|a,t)) = -\exp[\eta(a) + ct + \alpha] \\ &\Rightarrow 1 - P(Y = 1|a,t) = -\exp[-\exp(\eta(a) + ct + \alpha)] \\ &\Rightarrow P(Y = 1|a,t) = 1 - \exp[-\exp(\eta(a) + ct + \alpha)], \end{aligned}$$

where $\eta(a)$ is the *age fractional polynomial*. It is the regression coefficient related to time, and α is a constant (*intercept*).

The calculation of the prevalence derivative shows that

$$
\begin{aligned}
&\frac{\partial}{\partial a} P(Y + 1|a, t)\\
&= -\frac{\partial}{\partial a}(-\exp(\eta(a) + ct + \alpha)) \times \exp[-\exp(\eta(a) + ct + \alpha)]\\
&= \frac{\partial}{\partial a}(\eta(a) + ct + \alpha) \times exp(\eta(a) + ct + \alpha) \times \exp[-\exp(\eta(a) + ct + \alpha)]\\
&= \eta^{'}(a) \times \exp(\eta(a) + ct + \alpha) \times \exp[-\exp(\eta(a) + ct + \alpha)].
\end{aligned}
$$

At time t, we denote by $\pi(a)$ the prevalence according to age. The expression $\lambda(a)$ is then given by

$$
\lambda(a) = \frac{-\eta^{'}(a)\log(1 - \pi(a))(1 - \pi(a)) + (\beta - \mu_1) - (\beta - \mu_1 - \gamma)\pi(a)}{1 - \pi(a)}. \quad (8)
$$

By modeling prevalence with a *logit* link, we have

$$
\begin{aligned}
&\log\left[\frac{P(Y = 1|a, t)}{1 - P(Y = 1|a, t)}\right] = \eta(a) + ct + \alpha\\
&\Rightarrow \frac{P(Y = 1|a, t)}{1 - P(Y = 1|a, t)} = \exp(\eta(a) + ct + \alpha)\\
&\Rightarrow P(Y = 1|a, t) = \frac{\exp(\eta(a) + ct + \alpha)}{1 + \exp(\eta(a) + ct + \alpha)}.
\end{aligned}
$$

The calculation of the prevalence derivative gives the following relation:

$$
\frac{\partial P(Y = 1|a, t)}{\partial a} = \frac{\eta^{'}(a)\exp(\eta(a) + ct + \alpha)}{[1 + \exp(\eta(a) + ct + \alpha]^2}. \quad (9)
$$

The force of infection, denoted by $\lambda(a)$, according to age is expressed by

$$
\lambda(a) = \frac{-\eta^{'}(a) \times \pi(a) \times (1 - \pi(a)) + (\beta - \mu_1) - (\beta - \mu_1 - \gamma) \times \pi(a)}{1 - \pi(a)}. \quad (10)
$$

4.2.5 Choice of Model

Fractional polynomials were applied to model the relationship between prevalence and age using two link functions (*cloglog*, *logit*) and supposing that the explained variable follows a binomial distribution.

To select the best model, we have used the information criterion of Akaike. The smallest value of AIC (Akaike information criterion) obtained for the *logit* model is equal to 25,841.84 for the model, which takes into account age and survey year, 20,134.82 for the model including age, survey year, and injector status (yes/no), and 22,698.33 for the model including age, survey year, and HIV seropositive status (yes/no) as illustrated in Table 3.

5 Results

5.1 Introduction

The results obtained for prevalence and incidence of hepatitis C among DU in France are presented in this section.

We organize the presentation of results obtained as follows:

1. Results obtained using simulated data;
2. Results obtained using real data;
3. Results obtained using the 2000 generated samples.

A comparative study between results from simulated (generated) data and those from real data will be presented below. The results obtained using the simulated samples made it possible for us to validate our model. In all the figures, the horizontal axis represents age.

5.2 Simulated Data

We have used simulated data from the Coquelicot surveys. We have presented below the results obtained. In Fig. 4, the prevalence and HCV IFO are presented according to age and survey year. Figures 5 and 6 present them according to age, survey year, and injector status. Finally, Figs. 7 and 8 take into account age, survey year, and HIV seropositive status (yes/no). In all figures, the size of the points is proportional to the number of individuals in the population at each age. We took into account the survey weights for each individual (Fig. 4).

Table 3 Criteria of choice of the best model with different link functions (*logit* et *cloglog*)

Link function	Df	-Log likelihood	Deviance	AIC
Model according to age and survey year				
logit	4	12916.92	25833.84	25841.84
cloglog	4	13044.94	26089.88	26097.88
Model according to age, survey year and injector status (yes/no)				
logit	5	10062.41	20124.825	20134.82
cloglog	4	10372.30	20562.97	20572.97
Model according to age, survey year and HIV seropositivity (yes/no)				
logit	5	11344.163	22688.33	22698.33
cloglog	5	11409.84	22819.68	22829.68

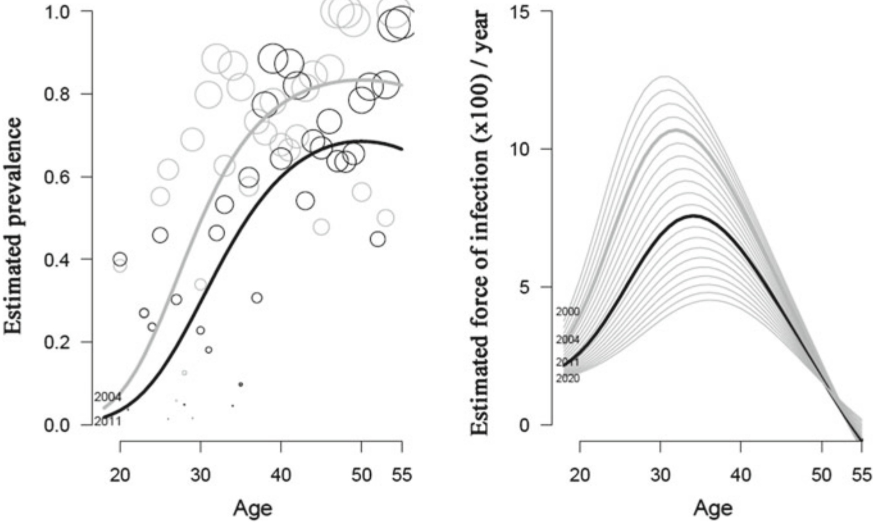


Fig. 4 *Left* estimate of hepatitis C prevalence among total population of drug users (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

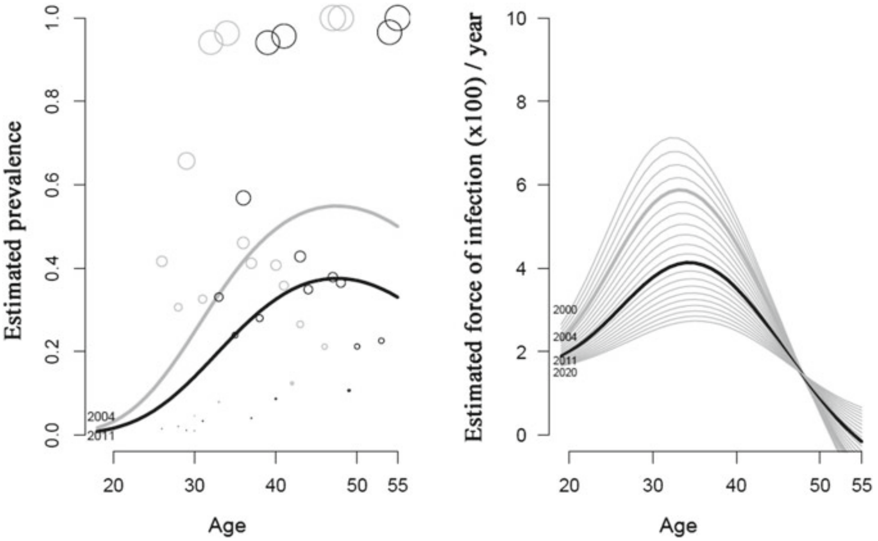


Fig. 5 *Left* estimate of hepatitis C prevalence among DU who have never injected (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

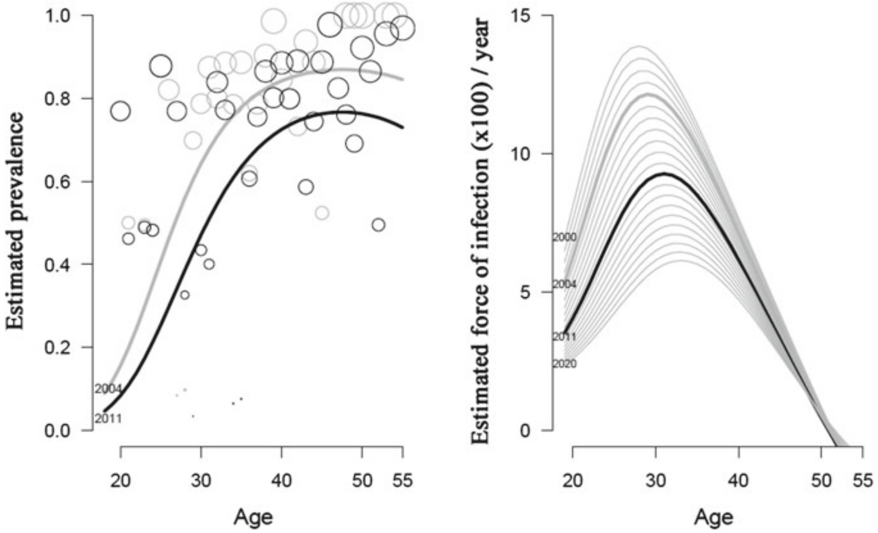


Fig. 6 *Left* estimate of hepatitis C prevalence among DU who have injected at least once (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

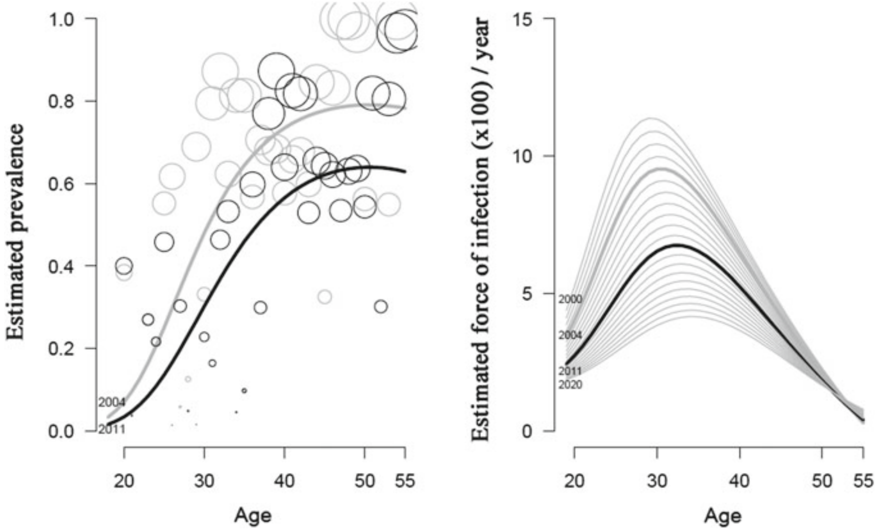


Fig. 7 *Left* estimate of hepatitis C prevalence among non-HIV-infected DU (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

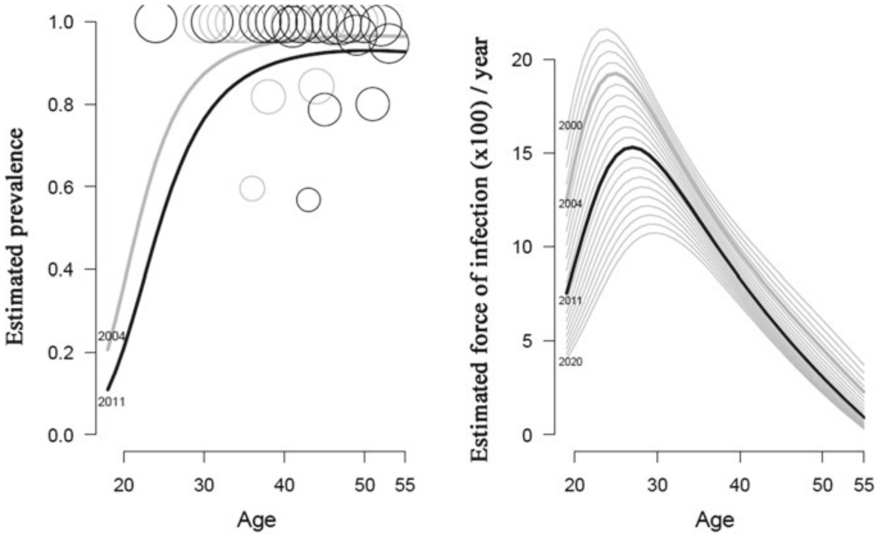


Fig. 8 *Left* estimate of hepatitis C prevalence among DU who are HCV and HIV simultaneously infected (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

5.3 Real Data

Using real data, prevalence was estimated according to age and survey year. The IFO obtained from this prevalence is illustrated in Fig. 9. Figures 10 and 11 show prevalence and IFO estimated according to age, survey year, and injector status. Finally, Figs. 12 and 13 present prevalence and IFO according to age, survey year, and HIV seropositive status (yes/no). In all these figures, the size of points is proportional to the number of individuals in the population at each age.

5.4 Model Validation

To validate our model, we have drawn 2000 random samples of the simulated population (1000 samples for 2004 and 1000 for 2011). Each sample was made up of 2000 individuals. We have estimated prevalence and IFO in these samples while taking into account the survey weight. We present respectively the estimated prevalence and IFO for the model taking into account age and survey year (Fig. 14); age, survey year, and injector status (Figs. 15 and 16); age, survey year, and HIV seropositive status (yes/no) (Figs. 17 and 18).

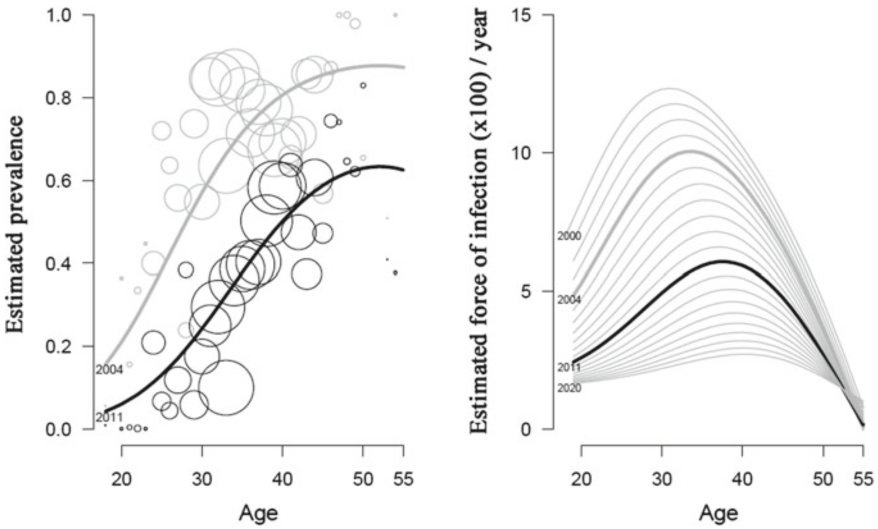


Fig. 9 *Left* estimate of hepatitis C prevalence among total population of drug users (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

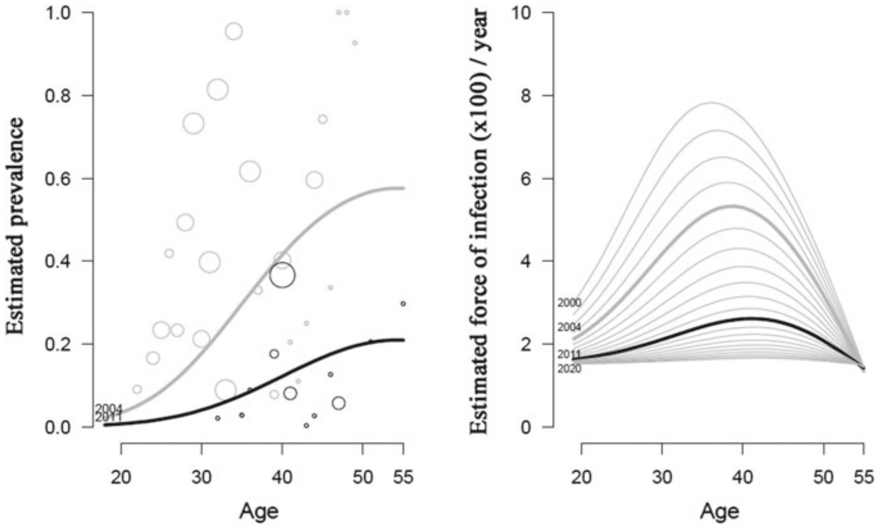


Fig. 10 *Left* estimate of hepatitis C prevalence among DU who have never injected during their life (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

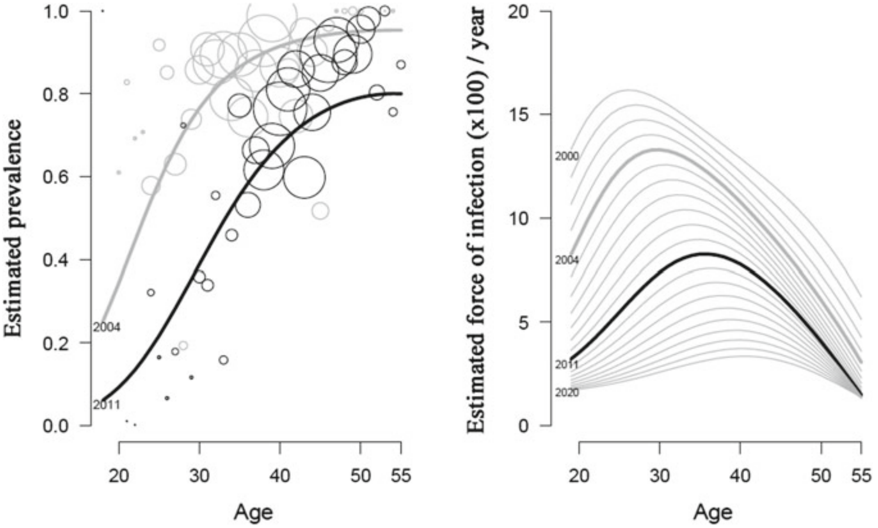


Fig. 11 *Left* estimate of hepatitis C prevalence among DU who have already injected at least once in their life (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

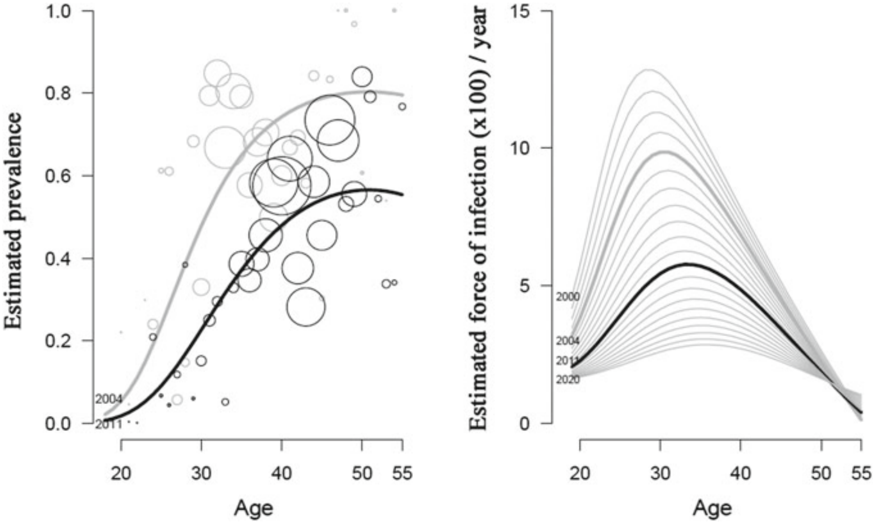


Fig. 12 *Left* estimate of hepatitis C prevalence among non-HIV-infected DU (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

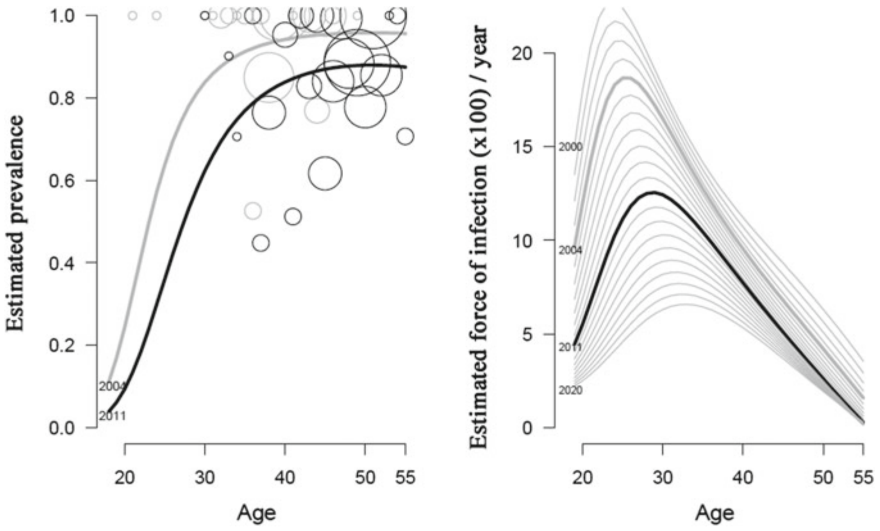


Fig. 13 *Left* estimate of hepatitis C prevalence among DU who are simultaneously HCV and HIV infected (2004: *gray* and 2011: *black*). *Right* estimate of IFO from 2000 to 2020

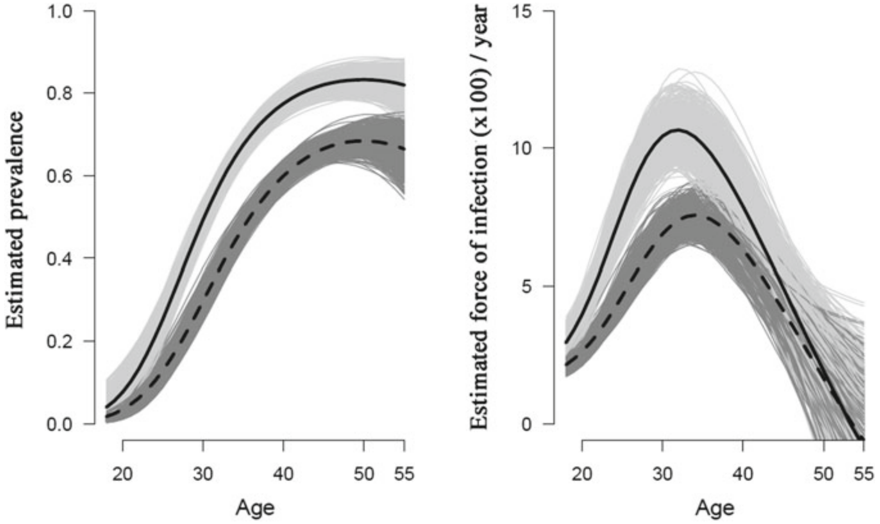


Fig. 14 *Left* estimate of prevalence in the total population for 2000 samples (2004: *light gray* and 2011: *dark gray*). The *black solid* and *dotted lines* represent respectively prevalence for 2004 and 2011. *Right* IFO

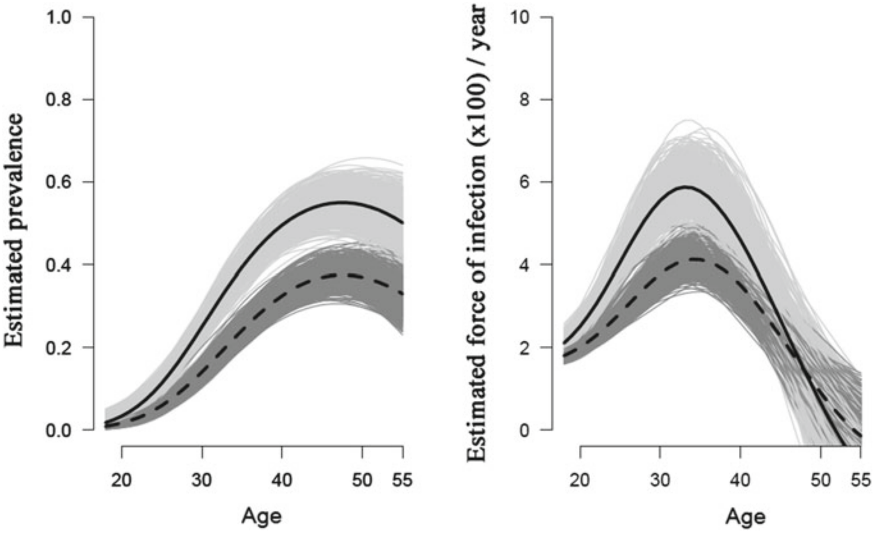


Fig. 15 *Left* estimate of prevalence according to age and survey year among DU who have never injected during their life (2004: *light gray* and 2011: *dark gray*). The *black solid* and *dashed lines* represent respectively prevalence for 2004 and 2011. *Right* the force of infection

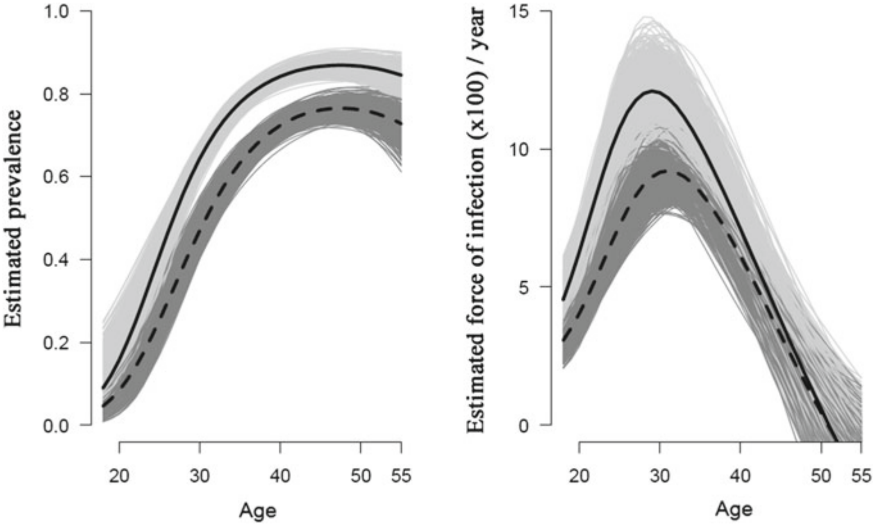


Fig. 16 *Left* estimate of prevalence according to age, survey year among DU who have already injected at least once in their life (2004: *light gray* and 2011: *dark grey*). The *black solid* and *dashed lines* represent respectively prevalence for 2004 and 2011. *Right* the force of infection

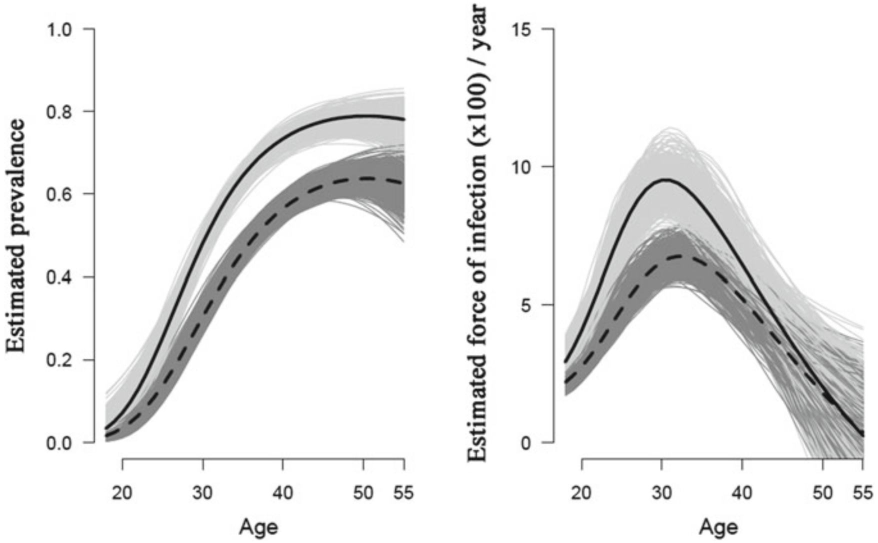


Fig. 17 *Left* estimate of prevalence according to age, survey year among non-HIV-infected DU (2004: *light gray* and 2011: *dark gray*). The *black solid* and *dashed lines* represent respectively prevalence for 2004 and 2011. *Right* the force of infection

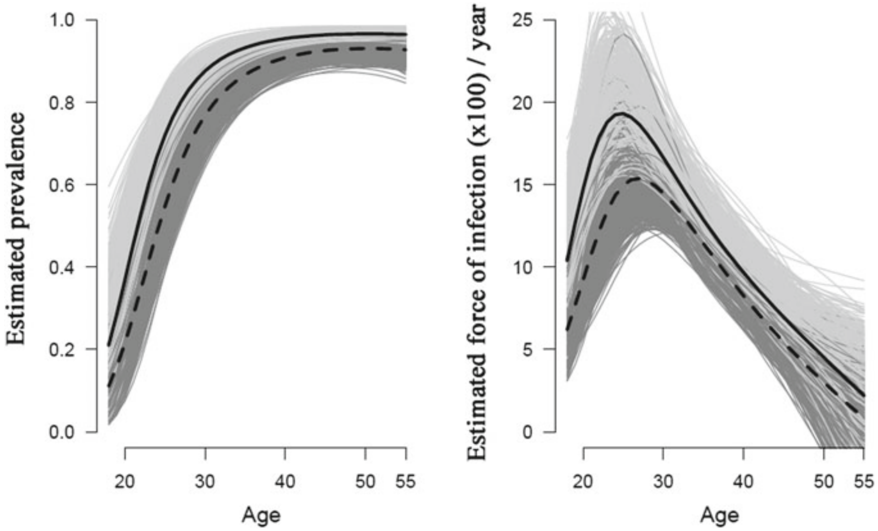


Fig. 18 *Left* estimate of prevalence according to age, survey year among DU drug users who are simultaneously HCV and HIV infected (2004: *light gray* and 2011: *dark gray*). The *black solid* and *dashed lines* represent respectively prevalence for 2004 and 2011. *Right* the force of infection

6 Discussion of Results and Concluding Remarks

6.1 General Discussion

Our estimates show that prevalence and IFO depend on age and time. IFO was estimated as a function of the derivative of prevalence between 2000 and 2020. The results obtained show that prevalence increases over age and evolves very quickly for the DU who are at least 30 years old. In the total population, the prevalence peak is reached around the age of 50 for 2004 and 2011. IFO increases among DU and reaches its peak around the age of 35 for the total population. A very high prevalence and IFO were observed among DU who are simultaneously HCV and HIV infected; prevalence reaches almost 90% for the DU aged from 35 to 55 years for 2011 versus 65% for the infected DU by HCV but not infected by HIV. For this population, the highest IFO is observed among the DU aged from 25 to 30 years old. The DU who have injected at least once in their lives have a high prevalence, more than 80% between the age of 40 and 50in 2004 and more than 75% in 2011 versus less than 60% in 2004 and less than 40% in 2011 for the DU who have never injected in their lifetime. Our estimates reveal a decrease in the IFO to less than 2% in 2020 for the DU of less than 40 years versus about 10% in 2004 for the same population.

The results obtained on simulated data tally with those collected from the real data for the different populations (total population, having injected at least once in their lives, and those who are simultaneously HCV and HIV infected). The simulations carried out with our model on the randomly drawn samples (2,000 samples) indicate a certain robustness of our estimates.

Our estimates show that prevalence and the IFO of the HCV among the DU in France will continue to decline over the following years for the categories of DU we have studied.

6.2 Concluding Remarks

Efforts to improve the measures of risk reduction of the contamination of hepatitis C among the DU in France have been shown to be effective. The considerable fall of prevalence rate and IFO in the coming years among the DU, as our study shows, is good news for French public health.

In France, as in several countries, the measures of risk reduction continue to be improved by programs for sharing syringes and needles, HCV treatment, and the voluntary screening for the HCV. This would contribute more to the decrease of HCV prevalence and HCV infection force in this particular population. One of the limits of our work is that we have worked only on anti-HCV antibodies. The data on the RNAs for the survey carried out in 2004 were unavailable.

In the context of perspectives, the model can be improved by considering rates of mortality (μ_1 and μ_2) as a function of age and HIV status. Another improvement could result from considering in our model the new DU who might be susceptible or infectious, for in our case they were susceptible when they joined the DU population (based on the Coquelicot data). We can also take into account, in the model, individuals who cease to be DU.

Currently in France there is no cohort study of the HCV among drug users. The only way to estimate HCV incidence in the French population is to use the only two national cross-sectional surveys that we have exploited.

The task carried out in this chapter is a contribution to science in the sense that it provides guidance for researchers to compare several cross-sectional epidemiological surveys among drug users and proposes an alternative method to estimate the force of infection among drug users from cross-sectional surveys in the absence of a cohort.

Our model is sufficiently general and can be easily used with transversal survey data of the HCV among the DU in other countries without major changes.

References

1. Aaron, S., McMahon, J.M., Milano, D., Torres, L., Clatts, M., Tortu, S., Mildvan, D., Simm, M.: Intranasal transmission of hepatitis C virus: virological and clinical evidence. Clin. Infect. Dis. **47**(7), 931–934 (2008)
2. Backmund, M., Reimer, J., Meyer, K., Gerlach, J.T., Zachoval, R.: Hepatitis C virus infection and injection drug users: prevention, risk factors, and treatment. Clin. Infect. Dis. **40**(5), S330–S335 (2005)
3. Ball, F.G.: Dynamic population epidemic models. Math. Biosci. **107**(2), 299–324 (1991)
4. Bonabeau, E.: Agent-based modeling: methods and techniques for simulating human systems. Proc. Natl. Acad. Sci. U.S.A **99**(3), 7280–7287 (2002)
5. Bourliere, M., Khaloun, A., Wartelle-Bladou, C., Oules, V., Portal, I., Benali, S., Adhoute, X., Castellani, P.: Chronic hepatitis C: treatments of the future. Clin. Res. Hepatol Gastroenterol **35**(2), S84–S95 (2011)
6. Chokki, A., et al., Kasereka, S., Kasoro, N.: Un modèle centré agents interactifs pour simuler la propagation de la grippe aviaire. Annales de la Faculté des Sciences de l'Université de Kinshasa **1**, 59–68 (2014)
7. Corson, S., Greenhalgh, D., Hutchinson, S.: Mathematically modelling the spread of hepatitis C in injecting drug users. Math. Med. Biol. **29**(3), 205–230 (2012)
8. Corson, S., Greenhalgh, D., Taylor, A., Palmateer, N., Goldberg, D., Hutchinson, S.: Modelling the prevalence of HCV among people who inject drugs: an investigation into the risks associated with injecting paraphernalia sharing. Drug Alcohol Depend. **133**(1), 172–179 (2013)
9. Dény, P., Roulot, D.: Virus de lh́epatite C. Elsevier, Paris (2003)
10. Hellard, M.E., Jenkinson, R., Higgs, P., Stoove, M.A., Sacks-Davis, R., Gold, J., Hickman, M., Vickerman, P., Martin, N.K.: Modelling antiviral treatment to prevent hepatitis C infection among people who inject drugs in Victoria, Australia. Med J Aust. **196**(10), 638–641 (2012)
11. Jauffret-Roustide, M., Le Strat, Y., Couturier, E., Thierry, D., Rondy, M., Quaglia, M., Razafandratsima, N., Emmanuelli, J., Guibert, G., Barin, F., Desenclos, J.C.: A national cross-sectional study among drug-users in France: epidemiology of HCV and highlight on practical and statistical aspects of the design. BMC Infect. Dis. **9**, 1–13 (2009)
12. Jauffret-Roustide, M., Pillonel, J., Weill-Barillet, L., Léon, L., Le Strat, Y., Brunet, S., Benoit, T., Chauvin, C., Lebreton, M., Barin, F.: Estimation of HIV and hepatitis C prevalence among drug users in France, first results of the survey anrs-coquelicot 2011 survey. BEH **509**, 39–40 (2013)

13. Kasereka, S., Kasoro, N., Chokki, A.P.: A hybrid model for modeling the spread of epidemics: theory and simulation, ISKO-Maghreb: concepts and Tools for knowledge Management (ISKO-Maghreb). In: 2014 4th International Symposium, pp. 1–7. Algiers (2014). doi:10.1109/ISKO-Maghreb.7033457
14. Kielland, K.B., Skaug, K., Amundsen, E.J., Dalgard, O.: All-cause and liver-related mortality in hepatitis C infected drug users followed for 33 years: a controlled study. J. Hepatol. **58**, 31–37 (2013)
15. Le Page, A.K., Robertson, P., Rawlinson, W.D.: Discordant hepatitis C serological testing in Australia and the implications for organ transplant programs. J. Clin. Virol. **57**, 19–23 (2013)
16. Martin, N.K., Vickerman, P., Grebely, J., Hellard, M., Hutchinson, S.J., Lima, V.D., Foster, G.R., Dillon, J.F., Goldberg, D.J., Dore, G.J., Hickman, M.: Hepatitis C virus treatment for prevention among people who inject drugs: Modeling treatment scale-up in the age of direct-acting antivirals. Hepatology **58**(5), REFE, 1598–1609 (2013)
17. Mathers, B., Degenhardt, L., Bucello, C., Lemon, J., Wiessing, L., Hick-man, M.: Mortality among people who inject drugs: a systematic review and meta-analysis. Bull World Health Organ **91**, 102–123 (2013)
18. Matser, A., Urbanus, A., Kretzschmar, M., Xiridou, M., Buster, M., Coutinho, R., Prins, M.: The effect of hepatitis C treatment and human immunodeficiency virus (HIV) co-infection on the disease burden of hepatitis C among injecting drug users in Amsterdam. Addiction **107**(6), 614–623 (2011)
19. Meffre, C., Le Strat, Y., Delarocque-Astagneau, E., Antona, D., Desenclos, J.: Prévalence des hépatites B et C en France en 2004. Institut de veille sanitaire **57**, 1–14 (2006)
20. OMS. *Hépatite C*, aide-mémoire numéro 64, Avril. http://www.who.int/mediacentre/factsheets/fs164/fr/ (2014)
21. Royston, P., Ambler, G., Sauerbrei, W.: The use of fractional polynomials to model continuous risk variables in epydemiology. Int. J. Epidemiol. **28**, 964–974 (1999)

Neurocomputing-Based Matrix Inversion: A Critical Review of the Related State of the Art

Vahid Tavakkoli, Jean Chamberlain Chedjou and Kyandoghere Kyamakya

Abstract Solving matrix inversion is very useful in many areas of science (e.g., in physics and engineering, such as chemical processes, robotics, electronic circuits, engineered materials, and other natural sciences). Various methods exist to solve matrix inversion problems. Most of them are very good algorithms, which, however, have the drawback of being efficient only when implemented on single-processor systems. Therefore, those algorithms are very inefficient when implemented on multiprocessor platforms; thus, they lack sufficient parallelizability. The main root of the problem lies in the nature of the algorithms, since they were originally designed for implementations on single-processor systems. Some novel concepts involving neurocomputing, however, have the potential for more efficiency in multicore environments. This chapter provides a comprehensive overview of both traditional and neurocomputing-based methods for solving the matrix inversion problem (i.e., analytical, heuristics, dynamical-system-based methods, etc.). These methods are compared based on some important criteria including convergence, parallelizability/scalability, accuracy, and applicability to time-varying matrices. Finally, we propose a new concept based on neurocomputing for solving the matrix inversion problem. The main advantage of this concept is the possibility of efficiently satisfying all the important criteria.

V. Tavakkoli (✉) · J.C. Chedjou · K. Kyamakya
Institute of Smart Systems Technologies, Alpen Adria University of Klagenfurt,
Universitätsstraße 65-67, 9020 Klagenfurt, Austria
e-mail: vtavakko@edu.aau.at

J.C. Chedjou
e-mail: jean.chedjou@aau.at

K. Kyamakya
e-mail: kyandoghere.kyamakya@aau.at

K. Kyamakya et al. (eds.), *Recent Advances in Nonlinear Dynamics and Synchronization*, Studies in Systems, Decision and Control 109,
DOI 10.1007/978-3-319-58996-1_16

1 Introduction

Matrix inversion is extensively used in linear algebra (e.g., for solving linear equations). Although matrix inversion is referred to in very old books, significant attention has been devoted to it (by scientists) mainly since the seventeenth century. The interest devoted to matrix inversion has led to the development of various methods, concepts, and algorithms for solving linear equations.

Solving matrix inversion is very useful in physics, engineering, and other natural sciences [1]. Solving real-time matrix inversion is part of mathematics and control theory. Several applications of matrix inversion are found in communication [2], machine learning [3], and robotics [4, 5]. To fulfill their roles in many applications, different methods have been developed to achieve fast convergence and higher accuracy of calculation. Among those methods, some famous ones are elimination of variables, Gaussian elimination (also known as row reduction), LU decomposition, Newton's method, eigendecomposition, and Cholesky decomposition, just to name a few. Generally, one can categorize matrix inversion methods into two different groups: recursive (or iterative) methods and direct methods [6, 7].

The first group encompasses methods like Gauss–Seidel and gradient descent. The initial condition (or starting point) is provided, and each step uses the previous value to calculate the new value. At each iteration, the solution approximation becomes better until the desired accuracy is attained [8, 9].

On the other hand, direct methods like Cholesky and Gaussian elimination typically compute the solution in a finite number of iterations. These methods can lead to exact solutions if there is no rounding error [10].

This chapter provides a full investigation of iterative and direct methods for matrix inversion. The prior history of the topic is related to a review of some important properties and characteristics of matrices, which could motivate the choice of appropriate or efficient methods (among the two main groups identified) for matrix inversion. Indeed, a suitable method for matrix inversion can be selected based on the sparsity level (i.e., number of zero elements in the matrix) of the matrix under consideration. Further criteria of relevance for such a method selection can also be requirements related to memory consumption and speed [11].

The remainder of the chapter is organized as follows. In Sect. 2 we present different types of matrices and discuss their properties with regard to the matrix inversion issue. Section 3 presents a critical review of both direct and recursive/iterative methods for matrix inversion. We also present our method developed as a new contribution to the matrix inversion problem. The pros and cons of the methods of matrix inversion are discussed through the benchmarking procedure. Some concluding remarks are formulated in Sect. 4 in order to clearly highlight the scientific achievement in this chapter.

2 Matrix Definition and Types

In mathematics and computer science, a matrix is defined as a set of numbers laid out in tabular form (i.e., in rows and columns). Each element of the matrix can contain different types of numbers, and they can also be expressed in multiple dimensions.

For illustration purposes, a 3 × 3 matrix is denoted by A:

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 5 & 4 \\ 4 & 6 & 5 \end{bmatrix}.$$

A **vector** is a 1-dimensional matrix that has either one row or one column (denoted by B).

$$B = \begin{bmatrix} 3 & 4 & 5 \end{bmatrix} \quad or \quad B = \begin{bmatrix} 5 \\ 8 \\ 1 \end{bmatrix}.$$

A **scalar** is a matrix with one row and one column.

A **square** matrix is a matrix with the same number of columns as rows (see C); otherwise it is a **nonsquare** matrix (see D):

$$C = \begin{bmatrix} 1 & 3 \\ 6 & 4 \end{bmatrix} \quad , \quad D = \begin{bmatrix} 7 & 1 & 2 \\ 8 & 9 & 5 \end{bmatrix}.$$

A **symmetric** matrix is a square matrix (see E) with all elements described by the following rule:

$$\forall\, i,\ j\ \in\ \mathbb{N}\ ,\ x_{i,j} = x_{j,i}$$

:

$$E = \begin{bmatrix} 2 & 4 & 3 \\ 4 & 5 & 6 \\ 3 & 6 & 1 \end{bmatrix}.$$

A **diagonal** matrix (see F) is a symmetric matrix in which all elements except those on the diagonal are zero:

$$F = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

An **identity** matrix (see I) is a diagonal matrix in which all elements on the diagonal are 1:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The **determinant** of a square matrix is a numerical value denoted by det (A). In case of a 2 × 2 matrix, we can calculate the determinant directly using the following formula:

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

Similarly, we can calculate the determinant of a 3 × 3 matrix using the following formula:

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}.$$

Each 2 × 2 determinant is called a **minor** determinant. This formula can be extended to calculate other determinants (e.g., 4 × 4 , 5 × 5 , $n \times n$).

The determinant of an $n \times n$ matrix can be calculated using the Leibniz formula or Laplace expansion [12].

Using the Leibniz formula, the determinant of a matrix A is expressed as a permutation of its elements. The Leibniz formula is expressed as follows:

$$\det(A) = \sum_{\sigma \in S_n} sgn\ (\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)},$$

where sgn is a function that returns +1 and −1 for even and odd permutation; S_n is the permutation group (e.g., a 3 × 3 matrix has six permutations). With this method, calculating the determinant of a matrix A requires O(n!) operations. The determinant of a matrix can be calculated with the LU decomposition method, which requires a maximum of $O(n^3)$ operations, as follows:

$$\det\ (A)\ = \det\ (L)\ \cdot \det(U)$$

The determinants of the matrices L and U are easy to calculate, since all elements below or above the diagonal are zero. Therefore, this determinant is obtained as the product of all diagonal elements.

Using the Laplace expansion, the determinant of a matrix A is expressed as follows:

$$\det(A) = \sum_{i'=1}^{n} a_{i',j} |C_{i',j}|\ \ ,\ C_{i,j} = -1^{i+j} \cdot M_{i,j},\ adj\ (A) = C^T.$$

Here M is a **minor** of A, C is called a **cofactor** of A, and the **adjugate** of the matrix A is the transpose of the matrix C. The matrix M is obtained by removing row i and column j from the matrix A. Thus each minor has $(n - 1) \times (n - 1)$ elements. For illustration, the matrix A below has nine minors:

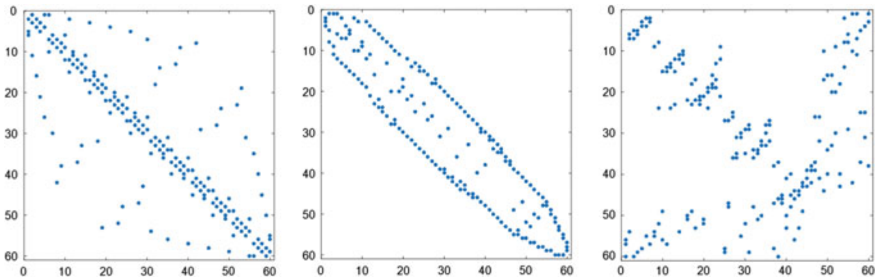


Fig. 1 Sparse matrix examples

$$A = \begin{bmatrix} 1\ 2\ 3 \\ 4\ 5\ 6 \\ 7\ 8\ 9 \end{bmatrix} \rightarrow \begin{vmatrix} 5\ 6 \\ 8\ 9 \end{vmatrix}, \begin{vmatrix} 4\ 6 \\ 7\ 9 \end{vmatrix}, \begin{vmatrix} 4\ 5 \\ 7\ 8 \end{vmatrix}, \begin{vmatrix} 1\ 2 \\ 7\ 8 \end{vmatrix}, \begin{vmatrix} 1\ 3 \\ 7\ 9 \end{vmatrix}, \begin{vmatrix} 2\ 3 \\ 8\ 9 \end{vmatrix}, \begin{vmatrix} 1\ 2 \\ 4\ 5 \end{vmatrix}, \begin{vmatrix} 1\ 3 \\ 4\ 6 \end{vmatrix}, \begin{vmatrix} 2\ 3 \\ 4\ 5 \end{vmatrix}.$$

This method is not used for large matrix determinant calculations because of its complexity of $O(n!)$.

In computer science, it is very important to reduce the amount of space needed for storing the matrix data in a computing system. Therefore, the matrix sparsity (density) issue has been studied. Sparsity is the ratio of the number of zero elements to the number of nonzero elements. Sparsity shows how much information exists in a given matrix. Thus, a sparse matrix is a matrix in which most of the elements are zero (see examples in Fig. 1). Such a situation is often encountered in many contexts such as graph problems and partial differential equations. The solving and storing of sparse matrices in an efficient way has been intensively addressed (see [13–20]).

One can categorize the different sparse matrix storage schemes based on efficient access or on efficient modification. For the first group, targeting efficient access, one can name the methods CSR (compressed spare row) and CSC (compressed spare column). And for the second group, targeting efficient access, one can name the methods LIL (list of lists), DOK (dictionary of keys), and COO (coordinated list).

The CSR and CSC schemes are similar except that in the first one, values are related to their respective row and column index. The matrix is saved in three different one-dimensional arrays. The first array contains the actual values of the nonzero elements. The second array contains the number of elements in each row (CSR) or column (CSC), and the last array has the actual column (CSR) or row (CSC). Therefore, each value has the following format: (value, column/row index, pointer to row/column).

LIL is a very simple scheme that creates one list for each row, and each list contains column indices and respective values. For improving performance, the list will be always automatically sorted.

DOK is created from keys made of row and column indices. The values of those keys are stored as pairs in a dictionary. In this way, one can find all elements in the dictionary; otherwise, the value(s) is (are) zero. This method significantly reduces the amount of memory required for saving the data. However, the drawback of the method is related to the long time it takes to find the value corresponding to a given key.

The COO scheme, unlike the DOK, does not create pairs but rather creates tuples of row, column, and value, which are saved in a list. The list can be indexed either by row or by column; in this way, one expects better access to the elements of a matrix.

3 Solving the Matrix Inversion Problem

Various traditional methods exist to solve the matrix inversion problem. These methods can be used either directly to solve matrix inversion or to solve a system of linear equations. In both cases, one proceeds similarly, and the result is then expressed as a matrix or a vector.

For example, let us consider the case of a system of linear equations. If a matrix A is invertible, its inverse is expressed as the matrix A^{-1}. To generalize this problem, we can express the system of equations to be solved in order to determine X. In this case, $A \in R^{n\times m}$ and the matrix X is its general inverse A^g:

$$A\,A^{-1} = I \quad or\ AXA = A \implies X = A^g. \tag{1}$$

In general, calculating the inverse of a nonsquare matrix $(n \times m)$ is impossible. It is generally required to create a new matrix of size $R^{n\times n}$ or $R^{m\times m}$. The methods used for nonsquare matrix inversion are called **pseudoinverse** methods [21].

3.1 Direct Methods for Matrix Inversion

This part explains the classical and commonly used direct methods for matrix inversion. The main property of those methods requires a finite number of iterations to reach a solution. The methods can lead to solutions if there is no rounding error.

- Cramer's rules

In this method, the number of equations is equal to the number of variables. This leads to a square matrix. The solution is valid if the system of equations has a unique solution.

Consider the following system of equations expressed in compact form (or matrix form):

$$AX = B,$$

where $A \in R^{n\ \times n}$, B and $X \in R^{n\ \times 1}$, and A has a non-zero determinant. The solution is vector $X{=}(x_1, x_2, \ldots, x_n)^T$ such that each element of X can be expressed as follows:

$$x_i = \frac{\det(A_i)}{\det(A)}, \tag{2}$$

where the matrix A_i is obtained by replacing the ith column of the matrix A with a column of zeros. It is clear that the problem has no solution if the determinant of A is zero.

It is possible to extend this method (see Eq. (2)) to realize a full matrix inversion. This is done using Eq. (3):

$$A^{-1} = \frac{1}{\det(A)} Adj(A). \tag{3}$$

Previously, this method appeared inefficient for computing, since it has a complexity of order $O(n!)$. But it has been proven that for large-scale matrices, it is possible to reach an algorithm complexity in the range experienced by LU decomposition [22] (the LU decomposition method will be explained later). In order to increase the efficiency of the method, the Chio condensation technique is used, which reduces the matrix size from order n to $n - 1$ when the determinants are calculated [23]. This technique significantly reduces the computational effort and makes the Cramer's rules a very effective algorithm when compared to alternative methods.

It is also possible to provide simple solutions for 2 × 2, 3 × 3, and 4 × 4 matrices. Block matrices can also be solved using this method, as explained in the next subsection.

- Blockwise inversion

A matrix can be inverted using blocks. An analytical method to invert a matrix using blocks is expressed as follows:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}. \tag{4}$$

This method is incorporated in some matrix inversion algorithms, and by combining it with efficient matrix multiplication algorithms like the Coppersmith–Winograd algorithm, it is possible to reach a complexity of order $\boldsymbol{O}(n^{2.3728639})$ [24].

- Gaussian elimination (also known as row reduction)

Gaussian elimination is an algorithm to solve systems of linear equations. This algorithm can also be used in problem solving for matrix inversion, determinant calculation, and matrix rank calculation. Using this method, series of row operations are performed, and the final result creates an upper triangular matrix in "unique reduced row echelon" form.

The following operations are allowed to be executed on rows:

1. Multiply a row by a nonzero constant
2. Add one row to another
3. Interchange two rows

This method is also used for matrix inversion. For matrix inversion, places the matrix A and the identity matrix I together (the matrix A on the left side and the identity matrix on the right side) and tries to use an appropriate sequence of the above-listed row operations to create an identity matrix in the left side. The inverse of the matrix A will be thereby created in the right side (see the three block matrices below):

$$A = \begin{bmatrix} 1\,3\,3 \\ 3\,4\,4 \\ 1\,2\,5 \end{bmatrix}$$

$$A\,|I\, = \left[\begin{array}{c|c} 1\,3\,3 & 1\,0\,0 \\ 3\,4\,4 & 0\,1\,0 \\ 1\,2\,5 & 0\,0\,1 \end{array}\right]$$

$$A\,|I\, = \left[\begin{array}{c|ccc} 1\,0\,0 & -\frac{4}{5} & \frac{3}{5} & 0 \\ 0\,1\,0 & \frac{11}{15} & \frac{1}{6} & -\frac{1}{3}\,. \\ 0\,0\,1 & \frac{2}{15} & -\frac{1}{15} & \frac{1}{3} \end{array}\right]$$

The complexity of this algorithm can be estimated as $O(n^3)$, since the number of divisions required in the solution is $\frac{n(n+1)}{2}$, and the numbers of multiplications and subtractions are roughly $3\mathrm{n}^3 + 5\mathrm{n}^{\,2} + 2\mathrm{n}$.

This method becomes unstable when a pivot element is zero. Therefore, while using this method, one should be particularly attentive to the possibility of instability due to errors [25, 26].

- QR factorization algorithms

In this method, the matrix A is factorized into an orthogonal vector Q and an upper triangular matrix R:

$$A{=}\begin{bmatrix} q_1 \; q_2 \; \ldots \; q_n \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ 0 & R_{22} & \cdots & R_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{nn} \end{bmatrix}$$

In this case, one can determine the inverse of the matrix A using Eq. (5). It is impossible to invert A if a single diagonal element of the matrix R is zero (this is because the determinant of the matrix R would become zero, and R would not be invertible) [27, 28]:

$$A^{-1}{=}(Q\ R)^{-1}{=}R^{-1}\ Q^T. \tag{5}$$

When the matrix R is of triangular form, finding the inverse of R is simple (the inverse of a triangular matrix can be calculated rapidly using forward substitution). This method is, however, too expensive to be used for matrix inversion.

For solving QR factorization, one can use a series of classical algorithms: Schmidt's algorithm [29], modified Gram–Schmidt [29], Householder transformations [29]. The last-named algorithm is used in various applications for solving the QR factorization problem with a relatively good convergence potential.

In the Householder transformation scheme, the algorithm uses reflections. A reflection across the plane orthogonal to a unit normal vector ν is shown in Eq. (6). The algorithm selects a vector υ such that H reflects it onto a coordinate axis; H is orthogonal to the vector:

$$H = I - 2\frac{\upsilon\,\upsilon^T}{\upsilon^T\,\upsilon}. \tag{6}$$

After finding reflections of the vector onto a different axis, one can find the matrix R by removing all projections. For example, suppose the matrix A is of size 4 × 4 and we use the Householder algorithm. In each step, after removing projections, we obtain the following status:

$$A = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \overset{P1}{\Longrightarrow} \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix} \overset{P2}{\Longrightarrow} \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix} \overset{P3}{\Longrightarrow} \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix} = R$$

We provide here MATLAB code for finding the Q and R matrices using the Householder transformation:

```
function [Q,R] = HouseHolder(A)

    [m, n] = size(A);

    Q=eye(n);

    R=A;

    I = eye(n);

    for j=1:n-1

        x=R(j:m,j);

        v=-sign(x(1))*norm(x)*eye(n-j+1,1)-x;

        if norm(v)>0

         v=v/norm(v);

         P=I; P(j:m,j:n)=P(j:m,j:n)-2*v*v';

         R=P*R;

         Q=Q*P;
```

```
            end
        end
    end
```

- LU decomposition

Let A be a square matrix. The LU decomposition of the matrix A creates two matrices, denoted by L and R. The matrix L is a lower triangular matrix, and U is an upper triangular matrix. The above matrices have the following relationship to each other:

$$A{=}LU. \tag{7}$$

The LU decomposition can be viewed as Gaussian elimination in matrix form. Many algorithms exist that satisfy the previous formula. This formula leads to Eq. (8):

$$a_{i,j} = \sum_{p=0}^{min(i,j)} l_{i,p}u_{p,j}. \tag{8}$$

The matrix L can be obtained based on Doolittle's LU decomposition:

$$l_{i,j} = \frac{a_{i,j} - \sum_{p=1}^{j-1} l_{i,p}u_{p,j}}{u_{i,j}} \quad if\ i > j \quad and\ u_{i,j} = a_{i,j} - \sum_{p=1}^{i-1} l_{i,p}u_{p,j}\ \ if\ \ j \geq 1. \tag{9}$$

This formula requires that one find the first row of the LU matrix, and then, in the same manner, continue to solve the remaining part.

In Crout's LU decomposition, one uses the following formula:

$$l_{i,j}{=}a_{i,j}{-}\sum_{p=1}^{\mathrm{i}-1} l_{i,p}u_{p,j} \quad if\ i{\geq}j \quad and\ u_{i,j}{=}\frac{a_{i,j}{-}\sum_{p=1}^{\mathrm{i}-1} l_{i,p}u_{p,j}}{l_{j,j}} \quad if\ j{>}i. \tag{10}$$

In this case, one can begin from the first row and column elements as the starting points of the algorithm and then continue to solve all values based on Eq. (9).

Some algorithms exist in which the matrix decomposition is based on LDU. The algorithms based on LDU provide a diagonal matrix D at the middle of the LU matrix. In this case, the previous algorithm will be used to eliminate the subdiagonal portion of A during each computational stage in order to form the matrix D (see Tinney and Hart [30]). According to the previous algorithm, the LU decomposition fails if a single pivot element (a_{ii}) is zero.

In many practical applications, it is possible that one pivot number becomes zero, and this leads to an instability of the solution. The origin of this problem is in Gaussian elimination, which is numerically unstable even if there are nonzero pivot elements in the matrix. This problem also leads to other problems such as the fact the approximate representation of numbers in a computer could create instability in the Gaussian elimination. Therefore, specific care should be taken in using this algorithm.

Like Gaussian elimination, this method has a complexity of order $O(n^3)$. For solving large sparse matrices, a special algorithm has been developed using a minimum number of elements, which is demonstrated in sparse L and U matrices [31].

It is also possible to combine this method (LU decomposition) with blockwise matrix inversion and create a method called block LU. This last-named method is very famous for parallel computing matrix inverses, since it is possible to separate the problem into different blocks and compute each part separately [32–34].

- Cayley–Hamilton

The potential of this method for matrix inversion has been proven. If a matrix A has size $n \times n$ and I has the same dimension as $\boldsymbol{A}$, then the characteristic of $\boldsymbol{A}$ is defined as follows:

$$p(\lambda) = \det(\lambda I - A). \tag{11}$$

Equation (11) is a monic polynomial in the variable λ. According to the Cayley–Hamilton theorem, substituting the matrix A for λ results in the zero polynomial:

$$p(A) = 0. \tag{12}$$

Using Eq. (12), the following expression is derived:

$$p(A) = \sum_{i=0}^{n} C_i A^i = 0 \overset{replacing\ constant\ and\ multiply\ with\ A^{-1}}{\Longrightarrow} A^{-1} = \frac{-1}{C_0} \sum_{i=0}^{n-1} C_i A^i. \tag{13}$$

The Caley–Hamilton expression in Eq. (13) offers a straightforward possibility for matrix inversion. However, the method is efficient only for matrices of small dimensions, since it requires an analytical calculation of the determinant of the matrix [35].

- Cholesky decomposition

The Cholesky decomposition of a positive definite matrix A is defined as follows:

$$A = LL^T,$$

where L is lower triangular with positive diagonal elements. It is possible to use this decomposition for matrix inversion or for solving systems of linear equations. None of the elements on the diagonal should be zero (a zero on the diagonal makes the matrix A noninvertible):

$$A = LL^T \ \Rightarrow A^{-1} = L^{-T} L^{-1}.$$

Obviously, the Cholesky decomposition leads to matrix inversion in terms of L^{-1} and L^{-T}. Although solving a lower triangular matrix using Cramer's rule is easy, it increases the complexity to order $O(n^3)$[36].

This method can be expressed in a different model called LDL. The main difference between this method and Cholesky is the introduction of the diagonal matrix D. Therefore, the matrix A can be expressed as follows:

$$A{=}LDL^T.$$

This expression can be rewritten as follows:

$$A{=}LDL^T = LD^{\frac{1}{2}}D^{\frac{1}{2}}L^T = (LD^{\frac{1}{2}})(LD^{\frac{1}{2}})^T$$

By introducing the matrix ***S***, which is a diagonal matrix containing the squared value of D, one can derive the following relationship between Cholesky and the LDL method:

$$S = D^2 \Rightarrow L^{LDL} = \; L^{Cholesky}S^{-1}.$$

The LDL method can be implemented efficiently for solvers [37]. This method shows very good performance as the sparsity of the matrix is increased [38].

3.2 Recursive/Iterative Methods

This section explores methods/concepts for matrix inversion inspired from the "Dynamic-systems" perspective. The main advantage of these methods is the possibility of monitoring and controlling the convergence to exact solutions analytically. Thus for a specific problem under investigation, analytical expressions/formulas are derived (in terms of the system parameters) under which the fast convergence to exact solutions is insured.

- Newton's method

This is a generalized method also used for solving matrix inversion problems. Newton's method requires a good seed (e.g., A^T) as initial condition. Newton's method is expressed mathematically in the following discrete formula:

$$X_{n+1} = 2X_n - X_nAX_n \quad , \; X_0 = A^T. \tag{14}$$

This method was first introduced by Victor and Reif in 1985 [39]. The corresponding mathematical model is expressed as follows:

$$R_{n+1} = (I - AX_{n+1}).$$

Using A^T as seed, R_0 is expressed as follows:

$$R_0 = \left(I - AA^T\right) \ \ and\ 0 < \left|\left(I - AA^T\right)\right| < 1.$$

Equation (14) can be used to determine R_{n+1}, and further, R_n is obtained as follows:

$$R_n = (I - AX_n) = (\ I - A\ X_{n-1}\ (I + R_{n-1})) = (I - AX_{n-1})\ R_{n-1} = {R_{n-1}}^2.$$

Finally,

$$R_n = (R_0)^{2n},$$

and hence

$$\lim_{n\to\infty} R_n \ \ = \lim_{n\to\infty} (R_0)^{2n} \ \ = 0.$$

Since $X_n{=}A^{-1}\ (I{-}R_n)$, the long-term evolution of the variable X is calculated as follows:

$$\lim_{n\to\infty} X_n \ \ = \lim_{n\to\infty} A^{-1}(I - {(R_0)}^{2n}) \ \ = A^{-1}.$$

This method can be implemented on parallel systems efficiently. Newton's method leads to complexity of order $O(log^2 n)$ [40].

- Neumann series

A given matrix A has the following property:

$$\lim_{n\to\infty} (I{-}A)^n = 0. \tag{15}$$

It is possible to express the matrix inversion problem using Eq. (16). This equation is derived from the geometric sum of different $(I{-}A)^n$ terms. Therefore, the solution of this equation always converges, provided the precondition (Eq.(15)) is satisfied:

$$A^{-1}{=} \sum_{n=0}^{\infty} (I{-}A)^n. \tag{16}$$

Equation (16) can provide a very fast approximation when compared to the "block-wise inverse" scheme [40, 41].

- P-adic algorithm

This method is used to invert a matrix all of whose entries are integers. If A is a square integer matrix of size n ($A \in M_n(\mathbb{Z})$), Dixon provided an algorithm [43] for solving $Ax{=}I$ using (Eq.(17)):

$$Ax_m \equiv b\ mod\ p^m, \tag{17}$$

where p is a prime number and that is coprime to $\det(A)$. Here $||A||$ represents the maximum absolute value of the components of A. The starting point for the solution, $Ax_0 \equiv I\ mod\ p$, is defined, and x_0 can be obtained by Gaussian elimination over the field $\mathbb{Z}/p$. Other values can be obtained using this starting point (i.e., for $m = 1, 2, \ldots$).

Let β be the maximum of $||A||$ and $||b||$. Dixon showed that if $m \geq 2n(\log\beta + \frac{1}{2}\log n)\frac{1}{\log p}$, then the solution x of $Ax=b$ can be found from x_m using an extended Euclidean algorithm [43]. The complexity of this algorithm is less than that of alternative algorithms for computing x_m.

It is possible to find the solution of $AX=I$ by obtaining the solution of the first formula for b and converting it to the solution of $Ax=I$ [43]. This algorithm has a complexity of order $O(n^4(\log n)^2)$.

- The gradient method

This method involves a first-order optimization technique for finding the minimum point of a function. This technique takes steps in the direction of the negative gradient. The method is based on observation of where the multivariate function $F(X)$ decreases the fastest if we choose the negative gradient of $F(X)$ at the point a: $-\nabla F(a)$:

$$b = a - \gamma\nabla F(a) \quad , \forall\gamma \in \mathbb{R} \quad and\, \gamma > 0\,, \tag{18}$$

where γ is the step size for updating decision neurons. For small γ, $F(a) \geq F(b)$; with this remark, we can extend our observations by Eq. (18):

$$x_{n+1} = x_n - \gamma\nabla F(x_n) \quad . \tag{19}$$

By iterating Eq. (19), $F(X)$ will converge to the minimum point of the function F.

Gradient descent can also be described as the Euler method for solving ordinary differential equations:

$$\dot{x}(t) = -\gamma\nabla F(x(t)). \tag{20}$$

Gradient descent can be used to solve linear equations. The problem reformulated as a quadratic minimization of a function is further solved using the gradient descent method:

$$F(x) = \frac{1}{2}||MX - I||^2 . \tag{21}$$

Then based on Eq. (21), we have

$$\nabla F(x) = M^T(MX - I) \quad . \tag{22}$$

Substituting Eq. (22) into (20) yields the following dynamic model [44]:

$$M\dot{X}(t) = -\gamma M^T MX + \gamma M^T I\,. \tag{23}$$

The exact analytical solution of equation (23) is expressed as follows:

$$X(t) = (X(0) - M^{-1})\, e^{-\gamma M^T M.t} + M^{-1}. \tag{24}$$

The first derivative of X(t), denoted by $\dot{\mathrm{X}}$ (t), is obtained as follows:

$$\dot{X}\,(t) = -\gamma\left(X\,(0) - M^{-1}\right)(M^T M + \dot{M}^T M t)e^{-\gamma M^T M.t}. \tag{25}$$

Equations (23) and (24) converge to the same solution described by the point M^{-1}:

$$\forall\gamma, a_{i,j}\in\mathbb{R}\ \ and\ \gamma > 0\ \ then\ \ \gamma M M^T > 0.$$

This gradient-based dynamics [32, 44, 45]. It is a pioneering concept developed by Tank–Hopfield on continuous-time neural networks. This concept was further extended to recurrent neural networks (also called dynamic neural networks (DNN)). DNN are designed by exploiting norm-based energy functions [46, 47].

- Zhang dynamics

Another method exists to create a dynamic system that converges to the exact or optimal solution [48, 49]. For using this method, the error function is defined in the following way:

$$E\,(t) = M(t)X\,(t) - I. \tag{26}$$

Thus the error will increase with the divergence of X from the exact solution. A convergence towards zero of the error is an insight for achieving the exact solution. Equation (26) changes over time such that the result will be the same as $AX = I$. This requires a monitoring of the error function in order to come closer to the solution. Therefore, we can define the derivative of this function as follows:

$$\dot{E}\,(t) = -\gamma E\,(t)\,. \tag{27}$$

The solution of equation (27) can be expressed as Eq. (26). This equation shows the exponential decrease of the error function. This forces convergence to the exact solution using the dynamic system (DS) perspective. This DS perspective consists in deriving a new mathematical differential equation corresponding to the problem under investigation:

$$E\,(t) = C\,e^{-\gamma t}. \tag{28}$$

Substituting $E\,(t)$ and $\dot{E}\,(t)$ in Eq. (28) leads to the following new dynamical system:

$$\mathrm{M}\dot{\mathrm{X}} + \dot{\mathrm{M}}\mathrm{X} = -(\mathrm{MX} - \mathrm{I}) \tag{29}$$

The solution of equation (29) is expressed as follows:

$$X\,(t) = Ce^{-\gamma t} + M^{-1}\,. \tag{30}$$

Taking the first derivative of Eq. (30) yields

$$\dot{X}\,(t) = -C\gamma e^{-\gamma t}\,. \tag{31}$$

Combining Eq. (30) with (31) leads to Eq. (27). It is observed that by increasing t to infinity, our dynamical system converges to the solution of the original equation $AX{=}I$.

- Chen dynamics

This method is a combination of the two previously described methods. It assumes that the matrix M is unchanging over time and thus the derivative of M is zero. Multiplying Eq. (29) by M and adding it to Eq. (23) leads to the mathematical model of a new dynamical system as reported in [50] (see also Eq.(32)):

$$\begin{gathered} M\dot{X}(t) = -\gamma MM^T \ \ (MX(t) - I) \\ M\dot{X}(t) = -\gamma\left(MM^T + I\right)(MX - I)\,. \end{gathered} \tag{32}$$

Equation (32) is solved to obtain the general solution, expressed as follows:

$$X(t) = -C\ M^{-1}e^{-(MM^T+I)t}V + M^{-1}\,. \tag{33}$$

Applying the first derivative to both sides of Eq. (33) leads to the following dynamical system:

$$\dot{X}(t) = C\ \left(M^T + M^{-1}\right)\ e^{-(MM^T+I)t}\,. \tag{34}$$

Combining Eq. (33) with (34) leads to Eq. (32). If $M^TM > 0$, we can state and confirm that this method has a better convergence rate than the first and second methods mentioned above.

- Tavakkoli dynamics

According to Chen [50], his model converges to the solution of the equation $AX = I$ for any initial value. Let us assume that one adds an additional $\left(MM^T\right)^2 + MM^T$ to Zhang's model. By adding this factor to the right-hand side of that equation, we have

$$M\dot{X}(t) = -\gamma\left(\left(MM^T\right)^2 + MM^T + I\right)(MX - I) - \dot{M}X. \tag{35}$$

The solution of this equation can be expressed as follows:

$$X(t) = -C.M(t)^{-1}e^{-\gamma\int_0^t((MM^T)^2+MM^T+I)dz} + M(t)^{-1}. \tag{36}$$

In Eq. (35), the newly added terms $\left(MM^T\right)^2 + MM^T$ provide a faster rate of convergence. The main reason for this improved convergence rate is the positive value of the integral, and it provides an additional factor to the previous time-varying model. Therefore, by adding more coefficients to the right-hand side of Eq. (35), we can obtain the following equation:

$$M\dot{X}(t) = -\gamma\left(\sum_{i=0}^{n}\left(MM^T\right)^n\right)(MX - I) - \dot{M}X. \tag{37}$$

Also, according to Zhang, the model in Eq. (37) can be extended by introducing a monotonically increasing function F, where $F(0) = 0$, to obtain the following:

$$M\dot{X}(t) = -\gamma\left(\sum_{i=0}^{n}\left(MM^T\right)^n\right)F(MX - I) - \dot{M}X \tag{38}$$

Theorem 1 *For a given nonsingular matrix $M \in \mathbb{R}^{n\times n}$ and state matrix $X(t) \in \mathbb{R}^{n\times 1}$, starting from any IVP $X(0) \in \mathbb{R}^{n\times 1}$ and monotonically increasing function F, where $F(0) = 0$, Eq. (38) will achieve global convergence to $X^*(t) = M^{-1}(t)$.*

Proof Let us define $E(t) = X(t) - X^*(t)$ for the error value during the process of finding the solution. If this equation is multiplied by M, one can find $M(t)E(t) = M(t)X(t) - M(t)X^*(t)$ or $M(t)E(t) = M(t)X(t) - I$. Thus, taking the derivative of the error function will lead to $M\dot{E}(t) + \dot{M}E(t) = M\dot{X}(t) - \dot{M}X(t)$. By replacing this in Eq. (38), we obtain the following formula:

$$M\dot{E}(t) = -\gamma\left(\sum_{i=0}^{n}\left(MM^T\right)^i\right)F(ME(t)) - \dot{M}\,E(t). \tag{39}$$

Let us define the Lyapunov function $\varepsilon(t) = \|ME(t)\|$, which is always a positive function. The derivative of this function can be obtained as follows:

$$\dot{\varepsilon}(t) = E(t)^T M^T M.\frac{dE(t)}{dt} + E(t)^T M^T \frac{dM(t)}{dt}E(t). \tag{40}$$

Substituting Eq. (39) into (40) leads to

$$\dot{\varepsilon}(t) = -\gamma E(t)^T M^T (\sum_{i=0}^{n}\left(M^T M\right)^i)F(ME(t)).$$

Hence

$$\dot{\varepsilon}(t) = -\gamma E(t)^T M^T (\sum_{i=0}^{n}\left(M^T M\right)^i)MF(ME(t))$$

$$\leq -\gamma\eta\, E(t)^T M^T F(ME(t)) \leq 0.$$

In the last equation, $E(t)^T M^T F(ME(t))$ is always positive, because if $ME(t)$ becomes negative, then $F(ME(t))$ also becomes negative and conversely.

Thus, it appears that $\dot{\varepsilon}(t)$ is always negative, and $\dot{\varepsilon}(t) = 0$ if and only if $X(t) = X(t)^*$ is satisfied. Therefore, our differential equation converges globally to a point, which is an equilibrium point for this function. By choosing different kinds of functions F, one can create the required dynamic property to implement this model. For example, sigmoid, linear, and arccosine functions are suitable for use

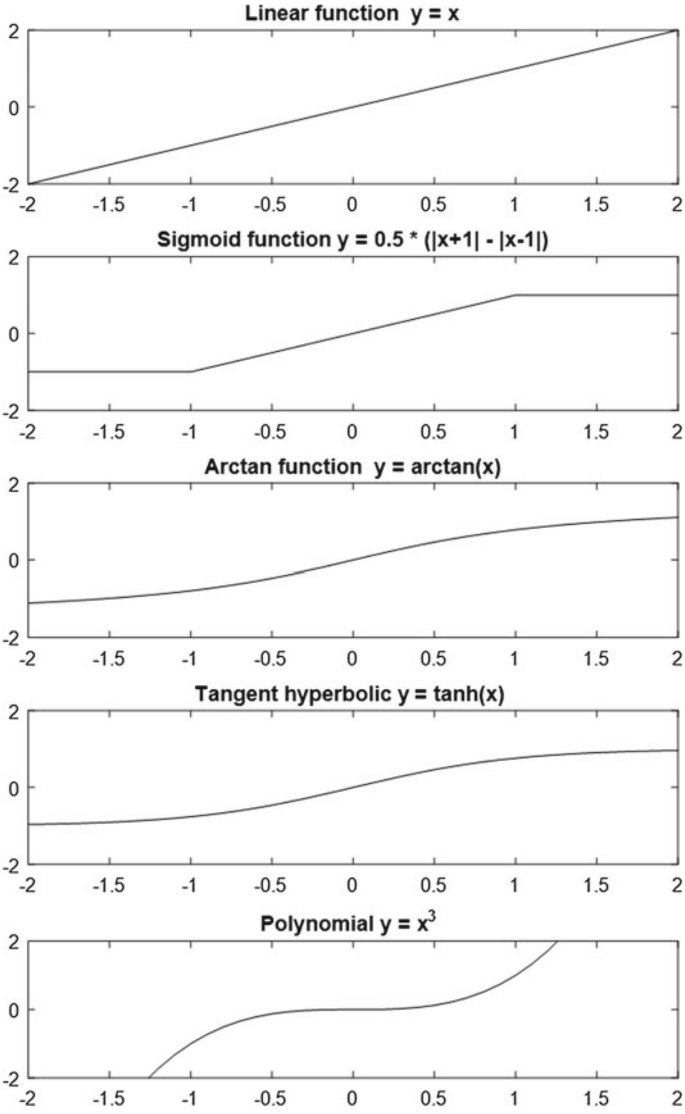


Fig. 2 Sample monotonic functions that can be used for solving time varying matrix inversion

in Eq. (37), since all of them are monotonically increasing functions satisfying the condition $F(0) = 0$. These functions are shown in Figs. 2 and 3.

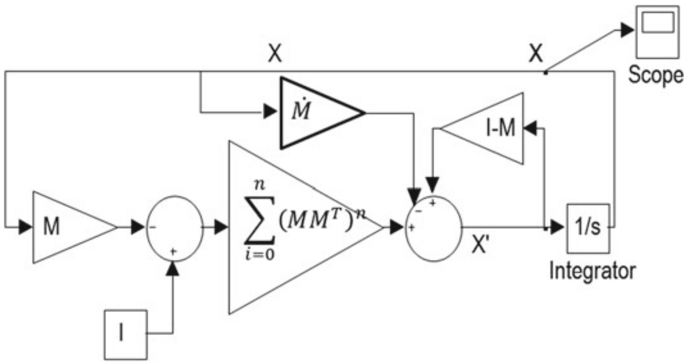


Fig. 3 The RNN Block diagram of Eq. (37)

4 Concluding Remarks

Matrix inversion is an efficient technique for solving high-order linear systems. However, the methods used for matrix inversion are prone to several challenges (e.g., accuracy, memory, consumption, robustness, stability, convergence, processing time), which have not yet been satisfactorily addressed by the relevant state of the art. This justifies the thorough critical review of the state of the art on matrix inversion carried out in this chapter. Several methods for matrix inversion have been identified, and their related pros and cons have been discussed. Some recommendations have been made for tackling the challenges at stake.

Regarding the memory issue faced in performing matrix inversion, one should note that it is possible to decrease the amount of memory use by compressing matrices. However, several methods for matrix compression work only on matrices with a high density of zeros (sparsity). Fortunately, in many fields of engineering, various problems can be identified that can be modeled in matrix form with sparse matrices (e.g., solving ordinary and/or partial differential equations). This can justify the tremendous attention that has been devoted in recent decades to the development of algorithms for sparse matrix operations. These algorithms decrease memory use while increasing computational efficiency in performing matrix inversion.

The critical review on matrix inversion carried out in this chapter has revealed that the best analytical method for matrix inversion generally encompasses algorithms for matrix multiplication (e.g., Coppersmith–Winograd algorithm) to reach a complexity of order $\boldsymbol{O}(n^{2.3728639})$. On the other hand, some theoretical methods (e.g., LU decomposition) are very efficient in solving matrix inversion; however, they are generally unstable, and this instability could lead to unreliable solutions. Analytical and heuristic methods for matrix inversion are generally robust despite their low reliability. This limitation is due to numerical rounding errors, which could lead to a divergence from the main target solution to an infeasible area.

Methods based on the dynamical system perspective (e.g., Dynamin neural networks, recurrent neural networks, cellular neural networks) appear to be the most appropriate for matrix inversion, since the challenges related to matrix inversion can be handled through monitoring of specific parameters of the dynamical system (expressed in the form of differential equations). In general, the methods inspired by the dynamical system perspective are based on error correction and energy reduction. This offers the possibility of carrying out a global stability analysis (see Lyapunov's theorem) in order to derive appropriate conditions for stability and convergence. Lyapunov's theorem also reveals the parameters of the dynamical system that can be monitored to improve accuracy in dealing with matrix inversion.

A significant contribution worth mentioning in this chapter is the new concept developed (see the Tavakkoli dynamics) for matrix inversion. This concept, inspired by the dynamical system perspective, has the main advantage of ensuring stability, accuracy, convergence, and speed of convergence through monitoring of specific parameters of the dynamical system. These parameters are coefficients of the resulting ODEs and/or PDEs. In essence, the concept developed is adaptive and has the possibility to correct itself during execution, and it can be proved that under certain conditions, the system will converge to the exact solution or will converge to an approximate solution (i.e., the closest solution to the exact solution in case the matrix is not invertible). This solution is obtained through a perturbation analysis. In fact, the parameter settings of the system corresponding to a zero determinant can be perturbed to obtain a variational form of the original system (i.e., the unperturbed system). Thus , the solution of the variational system converges to the closest solution.

Overall, the functioning principle of the methods for matrix inversion inspired by the dynamical system perspective is based on optimization. Thus the main problem of those methods is their convergence rate. Fortunately, the concept developed offers the possibility of increasing the convergence rate through monitoring of a specific parameter, and this offers the possibility of achieving convergence within a short time. The possibility of achieving convergence in a short time through monitoring of a specific parameter is a strong point (or good feature) of the new concept developed, since the other methods based on the dynamical system perspective do not offer this possibility. Another important feature of the new concept developed is that for solving differential systems of large matrices on von Neumann architecture, we can also take advantage of sparsity. In this context, sparsity provides the possibility of compressing a matrix's information. This operation (i.e., matrix compression) is efficiently performed by the methods inspired by the dynamical systems perspective. Thus the new system developed can be viewed as a significant contribution to the enrichment of the state of the art on matrix inversion.

References

1. Song, W., Wang, Y.: Locating multiple optimal solutions of nonlinear equation systems based on multiobjective optimization. IEEE Trans. Evol. Comput. **19**(3), 414–431 (2015)
2. Wang, Y., Leib, H.: Sphere decoding for MIMO systems with newton iterative matrix inversion. IEEE Commun. **17**(2), 389–392 (2013)
3. Gu, B., Sheng, V.: Feasibility and finite convergence analysis for accurate on-line v-support vector machine. IEEE Trans. Neural Netw. Learn. Syst. **24**(8), 1304–1315 (2013)
4. Zhang, Y., Ge, S.: Design and analysis of a general recurrent neural network model for time-varying matrix inversion. IEEE Trans. Neural Netw. **16**(6), 1477–1490 (2005)
5. Guo, D., Zhang, Y.: Zhang neural network, Getz-Marsden dynamic system, and discrete-time algorithms for time-varying matrix inversion with application to robots' kinematic control. Neurocomputing **97**, 22–32 (2012)
6. Ma, L., Dickson, K., McAllister, J., McCanny, J.: QR decomposition-based matrix inversion for high performance embedded MIMO receivers. IEEE Trans. Signal Process. **59**(4), 1858–1867 (2011)
7. Zhang, Y., Chen, K., Tan, H.: Performance analysis of gradient neural network exploited for online time-varying matrix inversion. IEEE Trans. Autom. Control **54**(8), 1940–1945 (2009)
8. Chen, Y., Yi, C., Qiao, D.: Improved neural solution for the Lyapunov matrix equation based on gradient search. Inf. Process. Lett. **113**(22–24), 876–881 (2013)
9. Yi, C., Chen, Y., Lu, Z.: Improved gradient-based neural networks for online solution of Lyapunov matrix equation. Inf. Process. Lett. **111**(16), 780–786 (2011)
10. Wilkinson, J.: Error analysis of direct methods of matrix inversion. J. ACM **8**(3), 281–330 (1961)
11. Straßburga, J., Alexandrovb, V.N.: Facilitating analysis of Monte Carlo dense matrix inversion algorithm scaling behaviour through simulation. J. Comput. Sci. **4**(6), 473–479 (2013)
12. Almalki, S., Alzahrani, S., Alabdullatif, A.: New parallel algorithms for finding determinants of NxN matrices. In: 2013 World Congress on Computer and Information Technology, Sousse (2013)
13. Zhanga, J., Wana, J., Lia, F., Maod, J., Zhuanga, L., Yuana, J., Liua, E., Yua, Z.: Efficient sparse matrix-vector multiplication using cache oblivious extension quadtree storage format. Future Gener. Comput. Syst. **54**, 490–500 (2016)
14. Korica, S., Guptac, A.: Sparse matrix factorization in the implicit finite element method on petascale architecture. Comput. Methods Appl. Mech. Eng. **302**, 281–292 (2016)
15. Wang, S., Peng, J., Liu, W.: Discriminative separable nonnegative matrix factorization by structured sparse regularization. Signal Process. **120**, 620–626 (2016)
16. Bickela, K., Wick, B.D.: A study of the matrix Carleson embedding theorem with applications to sparse operators. J. Math. Anal. Appl. **435**(1), 229–243 (2016)
17. Liu, W., Vinter, B.: A framework for general sparse matrix-matrix multiplication on GPUs and heterogeneous processors. J. Parallel Distrib. Comput. **85**, 47–61 (2015)
18. Pelta, D.M., Bisseling, R.H.: An exact algorithm for sparse matrix bipartitioning. J. Parallel Distrib. Comput. **85**, 79–90 (2015)
19. Aprovitolaa, A., D'Ambraa, P., Denarob, F.M., Serafinoc, D.d., Filippone, S.: SParC-LES: Enabling large eddy simulations with parallel sparse matrix computation tools. Comput. Math. Appl. **70**(11), 2688–2700 (2015)
20. Feng, Y., Xiao, J., Zhou, K., Zhuang, Y.: A locally weighted sparse graph regularized non-negative matrix factorization method. Neurocomputing **169**, 68–76 (2015)
21. Olaru, A., Olaru, S., Mihai, N.: Application of a new Iterative pseudo-inverse Jacobian Neural Network Matrix technique for controlling geckodrive DC motors of manipulators. In: 3rd RSI International Conference on Robotics and Mechatronics (ICROM), Teheran (2015)
22. Habgood, K., Arel, I.: A condensation-based application of Cramer's rule for solving large-scale. J. Discret. Algorithms **10**, 98–109 (2012)
23. Salihu, A.: New method to calculate determinants of nxn ($n \geq 3$) matrix, by reducing determinants to 2nd order. Int. J. Algebra **6**(19), 913–917 (2012)

24. Gall, F.L.: Powers of tensors and fast matrix multiplication. In: Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation, New York (2014)
25. Pana, V.Y., Qianb, G., Yan, X.: Random multipliers numerically stabilize Gaussian and block Gaussian elimination: proofs and an extension to low-rank approximation. Linear Algebra Appl. **481**, 202–234 (2015)
26. Dumasa, J., Gautierb, T., Pernet, C., Rochb, J., Sultan, Z.: Recursion based parallelization of exact dense linear algebra routines for Gaussian elimination. Parallel Comput. **57**, 235–249 (2015)
27. Burnik, K.: A structure-preserving QR factorization for centrosymmetric real matrices. Linear Algebra Appl. **484**, 356–378 (2015)
28. Su, Q.: The convergence of multi-shift QR algorithm for symmetric matrices. Appl. Math. Comput. **222**, 343–355 (2013)
29. Wilson, J.B.: Optimal algorithms of GramSchmidt type. Linear Algebra Appl. **438**, 4573–4583 (2013)
30. Tinney, W., Hart, C.: Power flow solution by newton's method. IEEE Transactions on Power Apparatus and Systems (1986)
31. Gilbert, J.R., Peierls, T.: Sparse partial pivoting in time proportional to arithmetic operations. SIAM J. Sci. Stat. Comput. **9**(5), 862–874 (1988)
32. Serre, F., Püschel, M.: Generalizing block LU factorization: a lower-upper-lower block triangular decomposition with minimal off-diagonal ranks. Linear Algebra Appl. **509**, 114–142 (2016)
33. Martinez-Fernandeza, I., Wozniakb, M., Garcia-Castilloa, L., Paszynskib, M.: Mesh-based multi-frontal solver with reuse of partial LU factorizations for antenna array. J. Comput. Sci. (2016)
34. Dewildea, P., Eidelmanb, Y., Haimovici, I.: LU factorization for matrices in quasiseparable form via orthogonal transformations. Linear Algebra Appl. **502**, 5–40 (2016)
35. Fenga, L., Tanb, H., Zhao, K.: A generalized Cayley-Hamilton theorem. Linear Algebra Appl. **436**, 2440–2445 (2012)
36. Fenga, S., Lianc, H., Xue, L.: A new nested Cholesky decomposition and estimation for the covariance matrix of bivariate longitudinal data. Comput. Stat. Data Anal. **102**, 98–109 (2016)
37. Rennich, S.C., Stosicb,D., Davis, T.A.: Accelerating sparse Cholesky factorization on GPUs. Parallel Comput
38. Langa, N., Menab, H., Saaka, J.: On the benefits of the LDLT factorization for large-scale differential matrix equation solvers. Linear Algebra Appl. **480**, 44–71 (2015)
39. Pan, V., Reif, J.: Efficient parallel solution of linear systems. In: STOC '85 Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing, Rhode Island (1985)
40. Pan, V., Schreiber, R.: An improved newton iteration for the generalized inverse of a matrix, with applications. SIAM J. Sci. Stat. Comput. **12**(5), 1109–1130 (1990)
41. Zhu, D., Li, B., Liang, P.: On the matrix inversion approximation based on neumann series in massive MIMO systems. In: IEEE ICC 2015, London (2015)
42. Qian, J., Stefanov, P., Uhlmann, G., Zhao, H.: An efficient neumann series-based algorithm for thermoacoustic and photoacoustic tomography with variable sound speed. Imaging Sci. **4**(3), 850–883 (2011)
43. Haramoto, H., Matsumoto, M.: A p-adic algorithm for computing the inverse of integer matrices. J. Comput. Appl. Math. **225**(1), 320–322 (2009)
44. Wang, J.: A recurrent neural network for real-time matrix inversion. Appl. Math. Comput. **55**, 89–100 (1993)
45. Zhang, Y., Chen, K., Tan, H.-Z.: Performance analysis of gradient neural network exploited for online time-varying matrix inversion. IEEE Trans. Autom. Control **54**, 1940–1945 (2009)
46. Zhang, Y.: Towards piecewise-linear primal neural networks for optimization and redundant robotics. In: Proceedings of IEEE International Conference on Networking, Sensing and, Control (2006)
47. Song, J., Yam, Y.: Complex recurrent neural network for computing the inverse and pseudoinverse of the complex matrix. Appl. Math. Comput. **93**, 195–205 (1998)

48. Zhang, Y., Ge, D.S.S.: Design and analysis of a general recurrent neural network model for time-varying matrix inversion. IEEE Trans. Neural Netw. **16**, 1477–1490 (2005)
49. Zhang, Y., Li, Z., Li, K.: Complex-valued Zhang neural network for online complex-valued time-varying matrix inversion. Appl. Math. Comput. **217**, 10066–10073 (2011)
50. Chen, K.: Recurrent implicit dynamics for online matrix inversion. Appl. Math. Comput. **219**, 10218–10224 (2013)