Lecture Notes in Electrical Engineering 307

Nikos Mastorakis
Valeri Mladenov *Editors*

Computational Problems in Engineering

Springer

Lecture Notes in Electrical Engineering

Volume 307

For further volumes:
http://www.springer.com/series/7818

Nikos Mastorakis • Valeri Mladenov
Editors

Computational Problems in Engineering

Springer

Editors
Nikos Mastorakis
Technical University of Sofia
Sofia
Bulgaria

Valeri Mladenov
Technical University of Sofia
Sofia
Bulgaria

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-3-319-03966-4 ISBN 978-3-319-03967-1 (eBook)
DOI 10.1007/978-3-319-03967-1
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014935238

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

B. Abou-Zalam Department of industrial electronics and control, Faculty of electronic engineering, Menofia University, Menuf, Egypt

A. G. Abro Department of Electrical Engineering, NED University of Engineering & Technology Karachi, Karachi, Pakistan

Karel Adámek Department of simulations, VUTS Liberec, a.s., Liberec XI, Czech Rep.

S. V. Agafonov Irkutsk State Agricultural Academy (ISAA), Irkutsk, Russia

Jayanthi Arasan Department of Mathematics, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia

V. Barrile DICEAM Department, Mediterranean University of Reggio Calabria, Reggio Calabria, Italy

G. Bilotta Ph.D. NT&ITA, Planning Department, IUAV University of Venice, Venice, Italy

Dz. Borovina Public Company for producing, distributing and retail of electrical energy, Sarajevo, Bosnia and Herzegovina

M. Cevri Department of Mathematics, Faculty of Science, Istanbul University, Istanbul, Turkey

K. J. Chee Intelligent Biometric Research Group, School of Electrical and Electronic, Engineering Campus, Universiti Sains Malaysia, Nibong Tebal, Penang, Malaysia

A. V. Daneev Irkutsk State Railway University (ISRU), Irkutsk, Russia

R. Farid Department of industrial electronics and control, Faculty of electronic engineering, Menofia University, Menuf, Egypt

S. R. Fleurke Radiocommunications Agency Netherlands, Groningen, The Netherlands

W. G. Früh Institute of Mechanical, Process and Energy Engineering, Heriot-Watt University, Edinburgh, UK

A. Ghaffar Luther College—Mathematics, University of Regina, Regina, SK, Canada

Eduardo González-Olivares Grupo de Ecología Matemática, Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

E. Grabovica Public Company for producing, distributing and retail of electrical energy, Sarajevo, Bosnia and Herzegovina

Jaime Huincahue-Arcos Grupo de Ecología Matemática, Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Departamento de Matemáticas, Universidad de Playa Ancha, Valparaíso, Chile

F. Hunka Faculty of Science, Department of Computer Science, University of Ostrava, Ostrava, Czech Republic

I. Husain Department of Mathematics, Luther College—University of Regina, Regina, SK, Canada

A. Ibrahim Department of industrial electronics and control, Faculty of electronic engineering, Menofia University, Menuf, Egypt

Noor Akma Ibrahim Department of Mathematics, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia

W. Kase Department of Electrical and Electronic Systems Engineering, Osaka Institute of Technology, Osaka, Japan

H. K. Kayadelen Department of Marine Engineering Operations, Yildiz Technical University, Istanbul, Turkey

K. N. Kiousis Institute of Mechanical, Process and Energy Engineering, Heriot-Watt University, Edinburgh, UK

Jan Kolář Department of simulations, VUTS Liberec, a.s., Liberec XI, Czech Rep.

Z. Kominkova Oplatkova Department of Informatics and Artificial Intelligence, Tomas Bata University in Zlin, Zlin, Czech Republic

S. Kovacevic Public Company for producing, distributing and retail of electrical energy, Sarajevo, Bosnia and Herzegovina

F. Labropulu Luther College—Mathematics, University of Regina, Regina, SK, Canada

C. Langdon Department of Mathematics, Luther College—University of Regina, Regina, SK, Canada

S. V. Lyamin Irkutsk State Railway University (ISRU), Irkutsk, Russia

E. S. G. Maciel Department of Energy Engineering, Foundation University of Great Dourados, Dourados, MS, Brazil

Z. Melis Faculty of Science, Department of Computer Science, University of Ostrava, Ostrava, Czech Republic

V.E. Messerle Combustion Problems Institute, Almaty, Kazakhstan

Institute of Thermophysics of Russian Academy of Science, Novosibirsk, Russia

J. Mohamad-Saleh School of Electrical & Electronic Engineering, Universiti Sains Malaysia, Nibong-Tebal, Penang, Malaysia

A. X. Moronis Energy Technology Department, Technological Educational Institute of Athens, Aegaleo, Greece

V. Piscopo Department of Science and Technology, Centro Direzionale—Isola C4, The University of Naples "Parthenope", Naples, Italy

M. Pluhacek Department of Informatics and Artificial Intelligence, Tomas Bata University in Zlin, Zlin, Czech Republic

Petr Půlpán Department of measuring, VUTS Liberec, a.s., Liberec XI, Czech Rep.

Martin Pustka Department of measuring, VUTS Liberec, a.s., Liberec XI, Czech Rep.

D. A. Ramli Intelligent Biometric Research Group, School of Electrical and Electronic, Engineering Campus, Universiti Sains Malaysia, Nibong Tebal, Penang, Malaysia

V. A. Rusanov Institute for System Dynamics and Control Theory (ISDCT SB RAS), Irkutsk, Russia

A. Scamardella Department of Science and Technology, Centro Direzionale—Isola C4, The University of Naples "Parthenope", Naples, Italy

J. Schwark Department of Mathematics, Luther College—University of Regina, Regina, SK, Canada

R. Senkerik Department of Informatics and Artificial Intelligence, Tomas Bata University in Zlin, Zlin, Czech Republic

N. Sulaiman School of Electrical & Electronic Engineering, Universiti Sains Malaysia, Nibong-Tebal, Penang, Malaysia

W. C. Tan Intelligent Biometric Research Group, School of Electrical and Electronic, Engineering Campus, Universiti Sains Malaysia, Nibong Tebal, Penang, Malaysia

Fauzia Taweab Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysi

Department of Statistics, University of Tripoli, Tripoli, Libya

Vyacheslav A. Trofimov Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russian Federation

Evgeny M. Trykin Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russian Federation

V.E. Tumanov Laboratory of Information Support for Research, Institute of Problems of Chemical Physics RAS, Chernogolovka, Russian Federation

Y. Ust Department of Naval Architecture and Marine Engineering, Yildiz Technical University, Istanbul, Turkey

A.B. Ustimenko Research Institute of Experimental and Theoretical Physics, Al-Farabi Kazakh National University, Almaty, Kazakhstan

D. Üstündag Department of Mathematics, Faculty of Science and Letters, Marmara University, Istanbul, Turkey

A.C. D. van Enter Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, The Netherlands

H. M. Yeap Intelligent Biometric Research Group, School of Electrical and Electronic, Engineering Campus, Universiti Sains Malaysia, Nibong Tebal, Penang, Malaysia

J. Zacek Centre of Excellence IT4Innovations , Division of the University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, Ostrava, Czech Republic

I. Zelinka Department of Computer Science, VŠB-Technical University of Ostrava, Ostrava-Poruba, Czech Republic

Chapter 1
Noise Reduction at the Fan Outlet

Karel Adámek, Jan Kolář, Petr Půlpán and Martin Pustka

Abstract The paper deals with numerical flow simulation in the fan outlet of a large painting shop. The received results of pressure fluctuations in numerical models are evaluated using both frequency analysis of pressure fluctuations and measuring and evaluation of a really operating system. From the conclusion there is defined the hypothesis of noise origin and more, it is proposed a more suitable design of the system without creation of pressure fluctuations.

Keywords Noise reduction · Fan outlet · Numerical flow simulation

1.1 Introduction

The paper deals with numerical flow simulation in the outlet system of a large painting shop. The increased noise level is spread into the surroundings. Several models were used for the identification of noise sources.

The relatively simple geometry of the large volume of about 100 m^3 consists from rectangular volumes made from thin metallic sheets with a cylindrical outlet, see following Figs. 1.1, 1.2 and 1.3. Due to the high and quick volume flow, the walls of the channel are vibrating and thundering, mostly in the central horizontal part of the observed system. The aim of the numerical flow modeling is to survey the pressure/velocity fields in the system and to define the source of the noise.

K. Adámek (✉) · J. Kolář
Department of simulations, VUTS Liberec, a.s.,
Svarovska 619, 460 01 Liberec XI, Czech Rep.
e-mail: karel.adamek@vuts.cz

J. Kolář
e-mail: jan.kolar@vuts.cz

P. Půlpán · M. Pustka
Department of measuring, VUTS Liberec, a.s.,
Svarovska 619, 460 01 Liberec XI, Czech Rep.
e-mail: petr.puplan@vuts.cz

M. Pustka
e-mail: martin.pustka@vuts.cz

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_1,

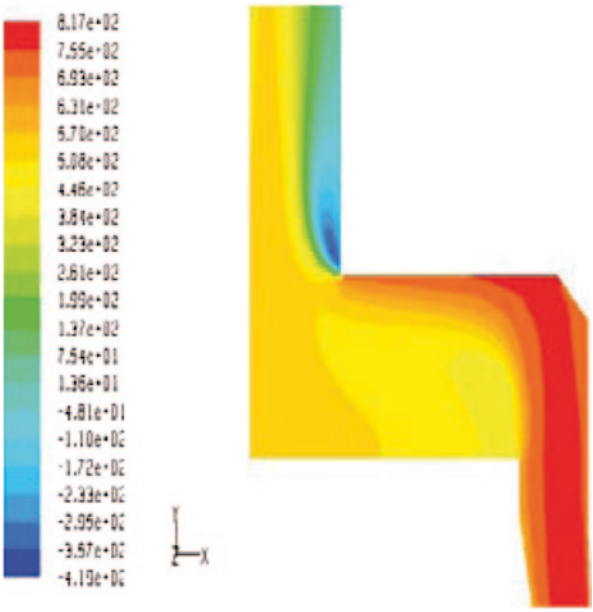


Fig. 1.1 Pressure field—steady 3D solution

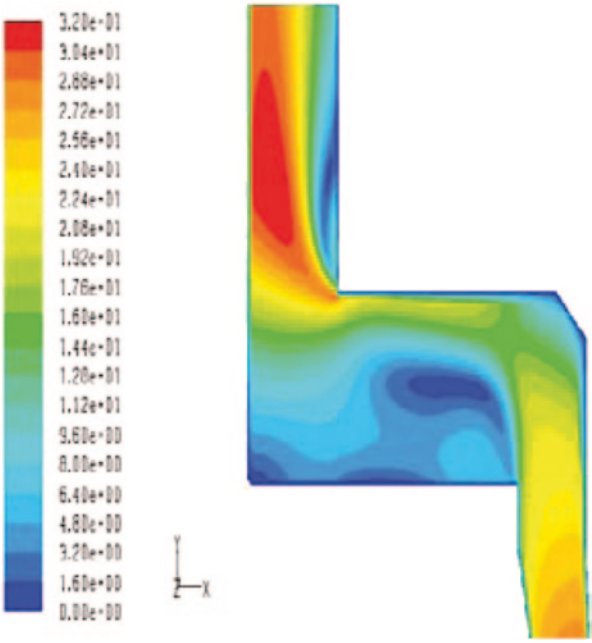


Fig. 1.2 Velocity field—steady 3D solution

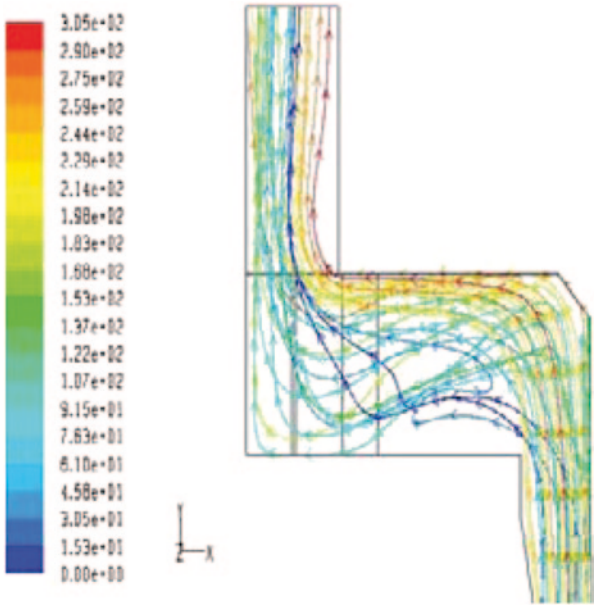


Fig. 1.3 Streamlines—steady 3D solution

1.2 Resolution

1.2.1 Numerical Model in General

The simple three-dimensional (3D) geometry is evident from the following Figs. 1.1, 1.2 and 1.3, where the longitudinal plane of symmetry can be used. The rectangular inlet is situated on the right side from below; the cylindrical outlet is situated on the left side upwards. Both parts are connected by a horizontal prismatic part. The defined pressure difference of 1,000 Pa creates the inlet velocity, which corresponds well with the real air flow of 58 m^3/s in the real size of inlet cross-section. The k-ε used standard commercial code for incompressible ideal gas and model of turbulence.

1.2.2 Steady 3D Solution—Results

On both pressure and velocity fields, see Figs. 1.1 and 1.2, it is visible that in the system with the rectangular changes of both cross-sections and flow direction, there are present intensive pressure and velocity gradients, the large areas of flow separation with backflows etc., which could be the reason of the pressure forces, causing the vibrations of the relative thin structure of air channels. Another view on the uneven flow field, there are the streamlines in Fig. 1.3. Generally said, the cross-section of the model volume is not fully filled by air flow.

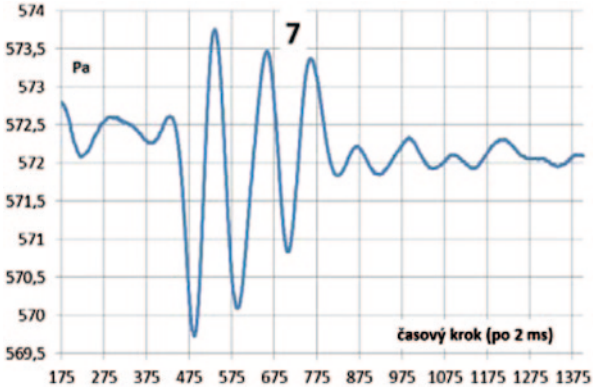


Fig. 1.4 Detected pressure fluctuations (time step of 2 ms)

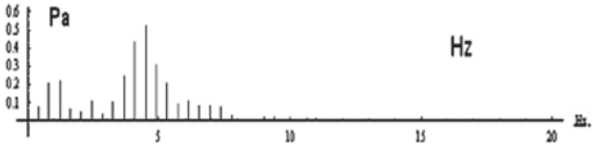


Fig. 1.5 Frequency analysis of recorded pressure fluctuations

The idealized evaluation of the results leads to significant force effects—the inlet velocity of 22 m/s represents the momentum of about 1,500 N and the impact power of such flow reaches a value of about 30 kW respectively, idealized as full impact effect. Of course, real flow is not fully stopped, so such excessive values are the theoretically possible maximum, only.

1.2.3 Unsteady Solution—Harmonic Analysis

From the following unsteady simulation, some pressure fluctuations were detected in the selected points. Figure 1.4 shows pressure fluctuations, recorded in the center on the wall of the horizontal part of the system (the geometry see the previous Figs. 1.1, 1.2 and 1.3).

The result of the frequency analysis of the recorded pressure fluctuations is presented in Fig. 1.5. It is visible the highest amplitude of 0.5 Pa approx. at the frequency of about 5 Hz, but it is not any audible frequency. Simply calculated, on the large wall surface of 12 m^2, made from a thin metallic sheet, there acts the total pressure force of 6.5 N approx. Excessively said, it looks like when on this sheet surface a 0.6 kg hammer is falling 5 times per second. It should be a really intense noise! Of course, due to real stiffening by the channel frame, the real force effect would be smaller.

The possible deformation of such a thin rectangular sheet, loaded by uniform pressure, is shown in Fig. 1.6—boundary conditions simplified as two sides free, two sides fixed and at constant pressure value.

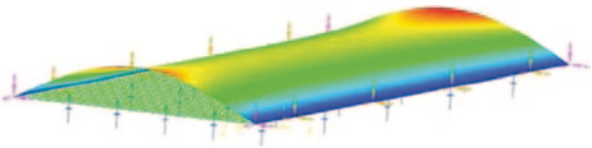


Fig. 1.6 Deformation of the thin sheet by constant pressure

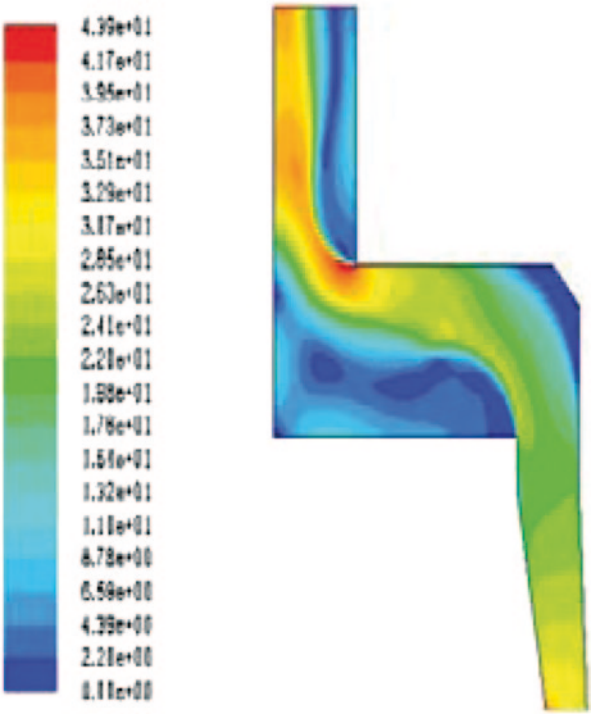


Fig. 1.7 Velocity field in the 2D model, fine mesh

From the above presented results, the hypothesis was determined that such primary low frequencies (pressure fluctuations) of the flow, separated from walls of the rectangular channel, could cause any secondary audible vibrations of the thin metallic sheets, which could be further amplified in the following outlet volume as in any trumpet.

1.2.4 Planar (2D) Model with Fine Mesh

To be sure that the above observed pressure fluctuations are not caused—maybe—by the relatively coarse mesh, used in this large model, the next model was created in 2D, only, but with the refined mesh and with boundary layers, too. The results are very similar so that in Fig. 1.7 similar to Fig. 1.2, it is presented the velocity field, only, as an example. Thus the reason of the detected pressure fluctuations

and subsequently of generated sound frequencies, too, could be the flow separation from the walls in the sharp changes of the flow direction.

1.2.5 Simulation of the Rotating Fan Rotor Influence

The rotating blade wheel of the exhaust fan creates other pressure fluctuations in the observed system. The generation of such fluctuations was simply simulated as a flap, rotating in the inlet cross-section of the previous model. As an illustration, in Fig. 1.8 there is the pressure field around such rotating flap in any random position.

The recorded pressure fluctuations are presented in Fig. 1.9 and the relevant result of the frequency analysis in Fig. 1.10. The maximum amplitude at the frequency of 4.5 Hz approx. remains the same as above (see Fig. 1.5); the next local amplitude maximum at the frequency of about 43 Hz corresponds with the fourth harmonic frequency of the rotating blade wheel.

Of course, the real rotating blade wheel could be simulated, too, but for the representation of periodical inlet excitations of the flow field the elementary rotating flap is sufficient.

1.2.6 Field Measurements

The results of the field measurements on the site [1] are very similar to the above presented results of numerical flow simulations. Measured in the air flow, the frequencies of 2.6—3.5—4.5 Hz were detected, depending on the actual position of the measuring point. The measured values are very similar to the above mentioned simulated values. It is hardly to get any better coincidence because the exact fixation of the pressure sensor in the strong air flow is difficult. And more, the next detected frequencies are very expressive harmonic multiples of the basic frequency 11.5 Hz of the rotating blade wheel.

As an illustration, only, Fig. 1.11 presents the result of harmonic analysis in one position of the pressure detector, where the first amplitude maximum at the frequency of 4.5 Hz corresponds with the basic pressure fluctuations found by numeric simulation and the second amplitude maximum at the frequency of 92.2 Hz is the eighth harmonic of the fan rotation frequency etc. Other local amplitude maximums are situated at frequencies approx. 45—66—88 Hz, corresponding to other harmonic multiples of the basic rotational frequency of 11 Hz.

1.3 New Design

It is clear that for suppressing the above identified pressure fluctuations it should be to improve the flow field in the system. In other words, it is necessary to design and to use a better shape of the exhaust channel, which complies better with the natural

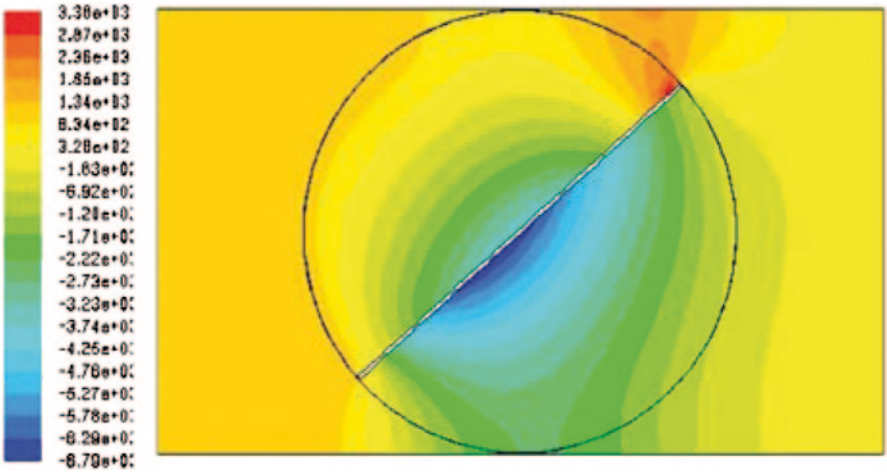


Fig. 1.8 Pressure field around the rotating flap at the inlet

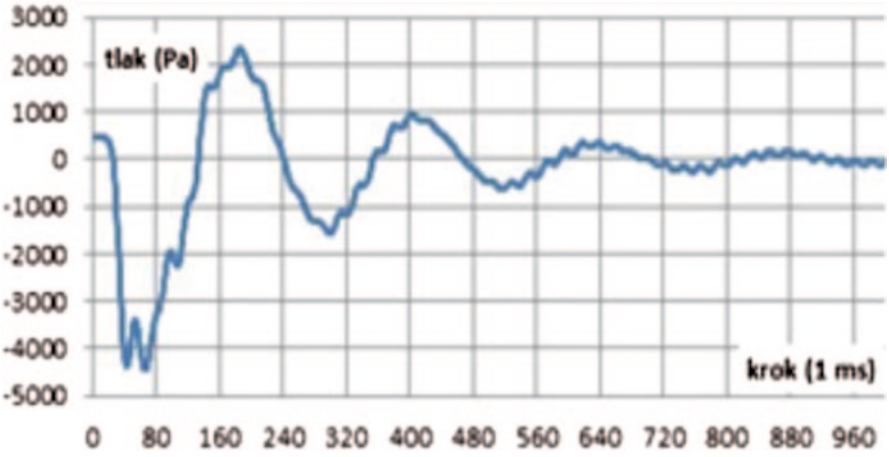


Fig. 1.9 Pressure fluctuations after the rotating flap (time step of 1 ms)

image of the simulated flow field. In the former channel of rectangular both cross-section and changes of the flow direction, the flow fills only a part of the whole cross-section, due to the large areas of the flow separation from the wall just behind each of rectangular bends.

Using a smaller and circular cross-section, designed after results of above presented simulations, the flow field becomes smoother, practically without flow separation—there are not any sharp changes of the flow direction. For comparison with Figs. 1.1, 1.2 and 1.3, here, there are presented analogous Figs. 1.12, 1.13 and 1.14—the pressure field with typical maximum at the outer diameter of the channel bend and minimum at the inner diameter of each bend, the velocity field

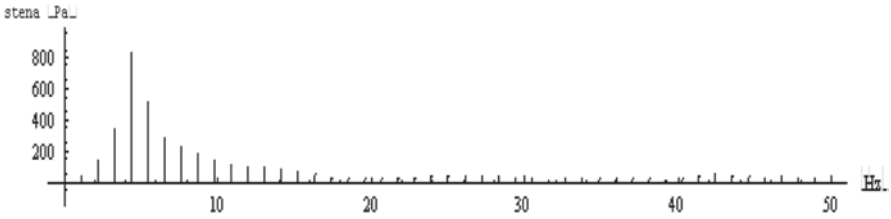


Fig. 1.10 Frequency analysis of pressure fluctuations after the rotating flap

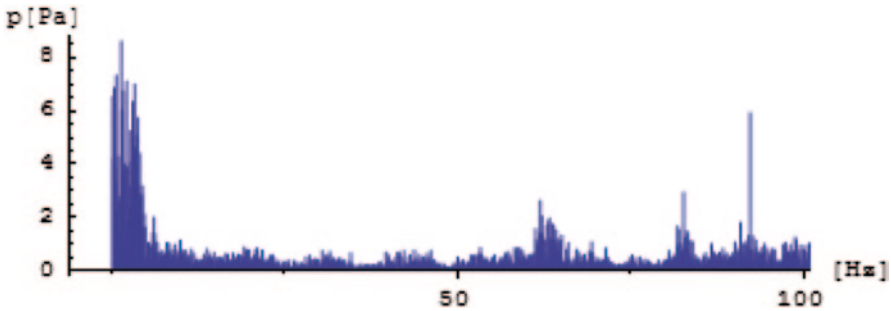
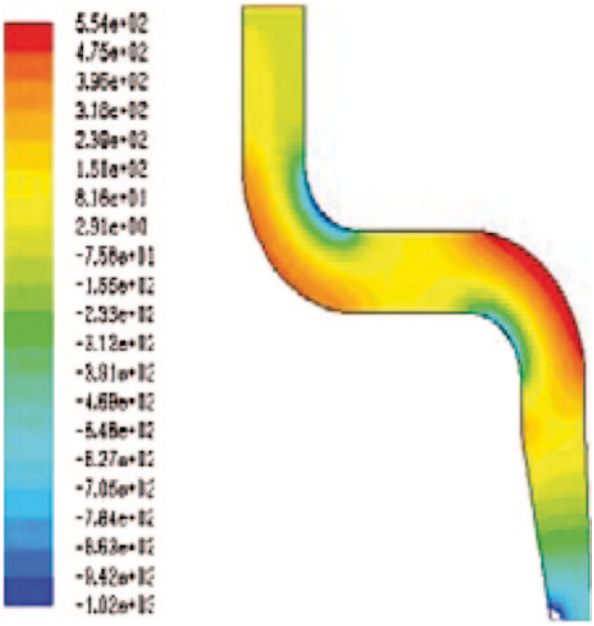


Fig. 1.11 Field measurement—record of frequency analysis

Fig. 1.12 Pressure field—smooth shape



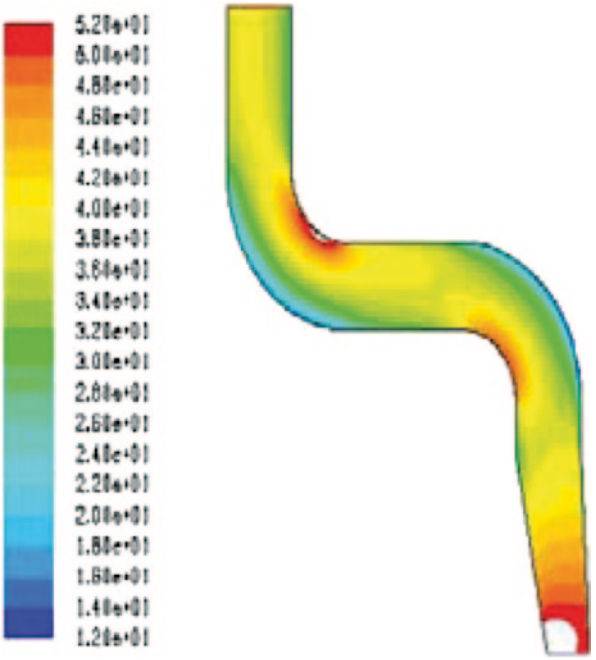


Fig. 1.13 Velocity field—smooth shape

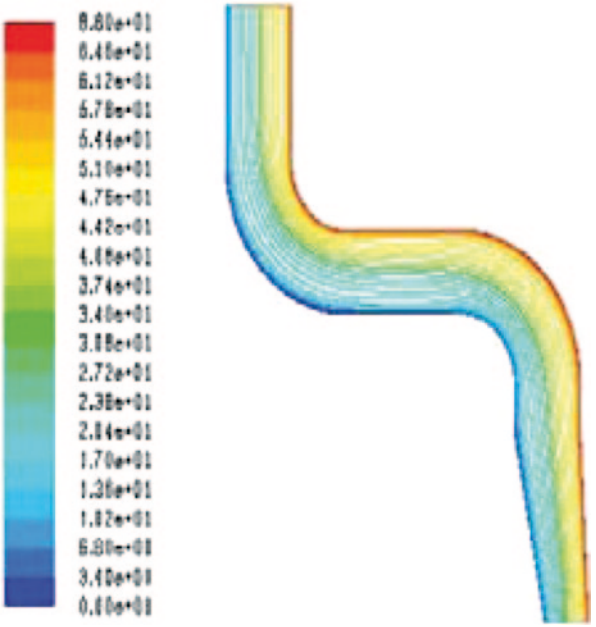


Fig. 1.14 Streamlines—smooth shape

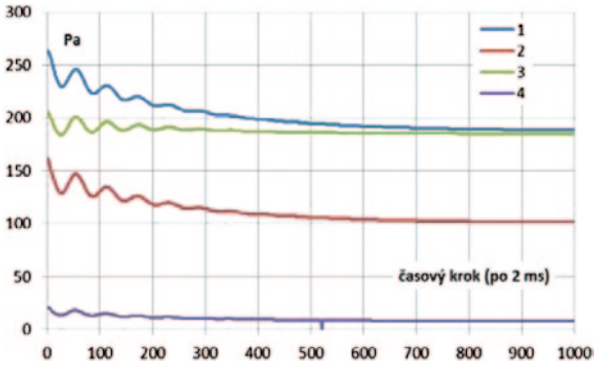


Fig. 1.15 Pressure fluctuations—smooth shape (time step of 2 ms)

with typically inverse values after the Bernoulli's equation (maximum value at the minimum radius and minimum value at the maximum radius) and the streamlines, differed by color, well following the channel shape without separation.

After the start-up period of simulation, the observed pressure fluctuations are going practically to zero, see Fig. 1.15. So it is possible to expect the flow field without pressure fluctuations, given by the above mentioned strong flow separation.

1.4 Conclusion

The used standard method of numerical flow simulation, here together with verification of results by frequency analysis of recorded pressure fluctuations, gives a suitable guide how to suppress the noise level spreading from the outlet of a fan in the surroundings.

The results of the flow simulation show the possible primary reason of the pressure fluctuations in the air flow, probably subsequently modified into vibrations of thin metallic structure, which are the secondary source of the noise, spread into the surroundings. Pressure fluctuations of the air flow, recorded in numerical models, are verified by real in situ measurements. From the following data analysis it is evident a good coincidence of both methods.

On the basis of the results of numerical flow modeling, there are proposed some arrangements of the geometry—rounded transitions of cross-sections instead of former rectangular ones. The resulting flow field does not show so large pressure fluctuations, which are usually the source of the increased noise level, particularly of the induced noise, generated by the interaction of such disturbed flow with thin metallic walls.

The presented study conserves the actual shape of the system. On the basis of next survey it is clear that significant shape modifications could be made, too, as for instance after Fig. 1.16. An absolutely straight (vertical) duct, equipped by a noise silencer at the outlet end, is designed as a labyrinth and/or louvers and made as walls resistant to vibrations. Due to the radial (horizontal) flowing in such a

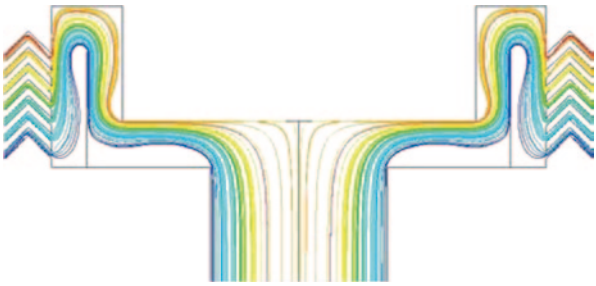


Fig. 1.16 Labyrinth as a noise silencer

silencer, the velocity value is decreasing so that the outlet value is a fraction, only, of the inlet one.

Such solution is not only ecologic—reduced noise level, but economic, too—simple shape. Generally said, an ecologic solution without economic effect is not the right solution.

In general, firstly, it should be used a pure technical solution, presented in this paper. The used better shape could suppress or remove the actual reason of the increased noise level. And secondly, only, it should be used any additional noise insulation of the existing system, in principle characterized by some operational defects, leading to increased noise level, as presented above.

References

1. Pustka M, Půlpán P (2012) Noise emissions reductions at the outlet from painting shop. Report VÚTS Liberec, unpublished

Chapter 2
Performance Evaluation of Gibbs Sampling for Bayesian Extracting Sinusoids

M. Cevri and D. Üstündag

Abstract This chapter involves problems of estimating parameters of sinusoids from white noisy data by using Gibbs sampling (GS) in a Bayesian inferential framework which allows us to incorporate prior knowledge about the nature of sinusoidal data into the model. Modifications of its algorithm is tested on data generated from synthetic signals and its performance is compared with conventional estimators such as Maximum Likelihood (ML) and Discrete Fourier Transform (DFT) under a variety of signal to noise ratio (SNR) conditions and different lengths of data sampling (N), regarding to Cramér–Rao lower bound (CRLB) that is a limit on the best possible performance achievable by an unbiased estimator given a dataset. All simulation results show its effectiveness in frequency and amplitude estimation of noisy sinusoids.

Keywords Bayesian inference · Parameter estimation · Gibbs sampling · Cramér–Rao lower bound and Power spectral density

2.1 Introduction

The sinusoidal frequency model embedded in noise is extensively important because of its wide applicability in many areas of science and engineering such as, modeling and manipulation of time-series from speech, audio to radar, seismology, nuclear magnetic resonance, communication problems and underwater acoustics [28].

We therefore address here a problem of estimating parameters of noisy sinusoids within a Bayesian inferential framework that provides a rigorous mathematical foundation for making inferences about them and a basis for quantifying uncertainties in their estimates. Under an assumption that a number of sinusoids is known a priori,

M. Cevri (✉)
Faculty of Science, Department of Mathematics, Istanbul University, Istanbul, Turkey
e-mail: cevri@istanbul.edu.tr

D. Üstündag
Faculty of Science and Letters, Department of Mathematics, Marmara University, Istanbul, Turkey
e-mail: dustundag@marmara.edu.tr

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,* Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_2,

several algorithms have already been applied to spectral analysis and parameter estimation problems, such as least-square fitting [33], maximum likelihood(ML) [25], discrete Fourier transform (DFT) [29, 8], and periodogram [27]. After Jayness' work [21], researchers in different fields of science have given much attention to the relationship between Bayesian inference and parameter estimation. Bretthorst and the others [4, 16, 6, 11, 12, 1, 38, 36, 37, 39] have done excellent works in this area for the last 16 years.

In Bayesian framework, it is necessary to evaluate high dimensional integrals that can be difficult and complex to tackle with. In order to solve these problems, different stochastic sampling algorithms have already been suggested and implemented by the different researches. Therefore, we introduce here one of the stochastic algorithms called Gibbs sampling [11, 12, 7] for recovering sinusoids from noisy data and compare its performance with classical estimators, regarding to Cramér–Rao lower bound (CRLB), that is widely used in statistical signal processing as a benchmark to evaluate unbiased estimators given a dataset [30]. For this purpose, a series of simulation studies with a variation in levels of noise and length of data sampling for a single sinusoid is set up.

The outline of this chapter is as follows. In Sect. 2.2, the harmonic signal models are introduced. In Sect. 2.3, we briefly outline Bayesian data analysis and summarize Gibbs sampling estimator in Sect. 2.4. Cramer–Rao lower bound (CRLB) is introduced in Sect. 2.5. Computer simulation results are given in Sect. 2.6 to evaluate the performance of the Gibbs sampling estimator by comparing with that of classical estimators in different conditions. Finally, conclusions from these simulations are drawn.

2.2 Harmonic Signal Model

In many experiments, a discrete data set $\mathbf{D} = \{d_1, d_2, ..., d_N\}^T$ denoted as an output of a physical system that we want to be modeled is sampled from an unknown function $y(t)$ at discrete times $\{t_1, ..., t_N\}^T$:

$$\begin{aligned} d_i &= y(t_i) \\ &= f(t_i; \theta) + e_i, \quad (i = 1, \ldots, N), \end{aligned} \tag{2.1}$$

where θ is a vector containing parameters that characterize behavior of physical system $f(t;\theta)$ and that are usually unknown. The term e_i is assumed to be drawn from a known random process. The choice of the model function $f(t,\theta)$ depends on the specific application, but we will consider here a superposition of k sinusoids:

$$f(t,\theta) = \sum_{j=1}^{k} a_{c_j} \cos(t\omega_j) + a_{s_j} \sin(t\omega_j), \tag{2.2}$$

where $\{a_{c_j}, a_{s_j}\} \in \mathbb{R}^{2k}$ and $\omega_j \in (0, \pi)$ are amplitudes and angular frequencies, respectively. Hence, Eq. (2.1) can be written in the matrix-vector form:

$$\mathbf{D} = \mathbf{G}\mathbf{a} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{D}$ is $(N \times 1)$ matrix of data points and $\mathbf{e}$ is $(N \times 1)$ matrix of independent identically distributed Gaussian noise samples. $\mathbf{G}$ is $(N \times 2k)$ matrix whose each column is a basis function evaluated at each point of time series. The linear coefficient $\mathbf{a}$ is a $(2k \times 1)$ matrix whose components are arranged in order of coefficients of cosine and sine terms $\{a_{c_1}, a_{s_1}, ..., a_{c_k}, a_{s_k}\}$. Then, the goal of data analysis is usually to infer $\theta = \{(a_{c_j}, a_{s_j}, \omega_j)\}_{j=1}^{k}$ from $\mathbf{D}$ and it is a non-linear optimization, due to frequencies. In signal processing literature, numerous approaches are based on frequentists statistics whereas only a few of them based on Bayesian statistics.

2.3 Bayesian Data Analysis

Bayesian inference can be provided from the product rule of probability calculus which can be originated rigorously starting with the formulation of a small number of desiderata required to define a rational theory of inference as first enunciated by Cox [9], with a more complete treatment given by Jaynes [22]. This formulation directs to the ordinary rules of probability calculus and indicates that every allowed (consistent) theory for inference must be mathematically equivalent to probability theory, or else inconsistent.

By using Bayes' rule [2, 3, 18], the context of the current problem can be expressed as follows:

$$p(\boldsymbol{\theta} \mid \mathbf{D}, I) = \frac{p(\boldsymbol{\theta})\, p(\mathbf{D} \mid \boldsymbol{\theta}, I)}{p(\mathbf{D})}, \tag{2.4}$$

where $p(\boldsymbol{\theta})$ is the prior probability density function (PDF) of the parameter vector $\boldsymbol{\theta}$ that encapsulates our state of knowledge of the parameters before observing $\mathbf{D}$; $p(\mathbf{D} \mid \boldsymbol{\theta}, I)$ is called the likelihood function when considered as a function of $\boldsymbol{\theta}$, but it is known as the sampling distribution when considered as a function of $\mathbf{D}$. $p(\mathbf{D})$ is denoted as an evidence or the marginal likelihood and $p(\boldsymbol{\theta} \mid \mathbf{D}, I)$ is the posterior PDF of the parameters $\boldsymbol{\theta}$ of interest, which summarizes the last information about it:

$$p(\boldsymbol{\theta} \mid \mathbf{D}, I) \propto p(\boldsymbol{\theta})\, p(\mathbf{D} \mid \boldsymbol{\theta}, I). \tag{2.5}$$

It is noted that for parameter estimation, the evidence $p(\mathbf{D})$ is $\boldsymbol{\theta}$-independent because of constant and simply plays role of a normalization factor. To proceed further in the specification of the posterior PDF, we now need to assign functional forms

for $p(\boldsymbol{\theta})$ and $p(\mathbf{D}\,|\,\boldsymbol{\theta}, I)$. After computing $p(\boldsymbol{\theta}\,|\mathbf{D}, I)$, the problem turns out to search a vector $\boldsymbol{\theta}$ that satisfies

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ p(\boldsymbol{\theta}\,|\mathbf{D}, I) \right\}, \tag{2.6}$$

where Θ is a parameter space.

2.4 Gibbs Sampling

In order to avoid computing the multivariate maximization problem described in Sect. 2.3, an alternative way is the one, proposed by Dou and Hogdson [11, 12], which combines Gibbs sampling (GS) with Bayesian inference theory. Gibbs sampling is an iterative Monte Carlo sampling process [14, 26, 20] and a special case of Metropolis–Hastings sampling [27, 19] wherein the random value is always accepted. It was also used by Geman and Geman [15] in image restoration. Statisticians [35, 13] began to utilize the method for Bayesian computations. It is based on supposing that the target distribution is a posterior probability distribution but, it can be applied to any target distribution, when their full conditional probability distributions are available. We extend here its derivation for multiple frequency signals and briefly summarize it below, but refer to the papers [12, 1] for more detail information.

For linear parameters $\mathbf{a}$ in Eq. (2.3), when σ^2 is known and there is no any specific information about $\{\mathbf{a}_c, \mathbf{a}_s\}$ prior to the observation $\mathbf{D}$, then Eq. (2.5) turns out to be the following form:

$$p(\mathbf{a}_c, \mathbf{a}_s \,|\mathbf{D}, \sigma^2, \omega, I) \propto p(\mathbf{D}\,|\omega, \mathbf{a}_c, \mathbf{a}_s, \sigma^2, I), \tag{2.7}$$

where $p(\mathbf{a}_c, \mathbf{a}_s) \propto$ constant as an uninformative uniform prior PDF for $\{\mathbf{a}_c, \mathbf{a}_s\}$. The marginal posterior distribution of $\mathbf{a}$ given $\boldsymbol{\omega}$ and $\mathbf{D}$ becomes a multivariate normal distribution $\mathcal{N}_m(\hat{\mathbf{a}}, \sigma^2(\mathbf{G}^T\mathbf{G})^{-1})$ [11, 12]:

$$p(\mathbf{a}\,|\,\mathbf{D}, \omega, \sigma^2) = \frac{|\,\mathbf{G}^T\mathbf{G}\,|^{1/2}}{(\sqrt{2\pi}\sigma)^m} e^{-\frac{1}{2\sigma^2}(\mathbf{a}-\hat{\mathbf{a}})^T\mathbf{G}^T\mathbf{G}(\mathbf{a}-\hat{\mathbf{a}})} \tag{2.8}$$

where $\hat{\mathbf{a}}$ is best estimate for $\mathbf{a}$ and $p(\mathbf{a}\,|\,\mathbf{D}, \sigma^2)$ is maximized at $\hat{\mathbf{a}}$. When the variance σ^2 is unknown, by using Jeffreys prior

$$p(\sigma^2) = \frac{1}{\sigma^2} \tag{2.9}$$

and integrating this joint posterior PDF in (2.7) with respect to σ^2,

$$p(\mathbf{a}_c, \mathbf{a}_s \,|\mathbf{D}, \omega, I) \propto \int_0^\infty p(\mathbf{D}|\omega, \mathbf{a}_c, \mathbf{a}_s, \sigma^2)\, p(\sigma^2) d\sigma^2$$
$$\propto \int_0^\infty \frac{p(\mathbf{D}|\omega, \mathbf{a}_c, \mathbf{a}_s, \sigma^2)}{\sigma^2} d\sigma^2, \tag{2.10}$$

the marginal posterior distribution of $\mathbf{a}$ given $\boldsymbol{\omega}$ and $\mathbf{D}$ in Eq. (2.8) turns into the multivariate Student's t distribution $t_m(\hat{\mathbf{a}}, s^2(\mathbf{G}^T\mathbf{G})^{-1}, \nu)$ [11, 12]:

$$p(\mathbf{a} \mid \mathbf{D}, \omega) = \frac{\Gamma[(\nu+m)/2]\,|\,\mathbf{G}^T\mathbf{G}\,|^{1/2}\; s^{-m}}{\left[\Gamma(1/2)^m \Gamma(\nu/2)(\sqrt{\nu})^m\right.} \left[1 + \frac{(\mathbf{a}-\hat{\mathbf{a}})^T \mathbf{G}^T\mathbf{G}(\mathbf{a}-\hat{\mathbf{a}})}{\nu s^2}\right]^{-(\nu+m)/2} \tag{2.11}$$

where $\nu = N - m$ is degrees of freedom and $s^2 = \frac{1}{\nu}(\mathbf{D}-\hat{\mathbf{D}})^T(\mathbf{D}-\hat{\mathbf{D}})$ is sampling variance.

Suppose that a_{c_j} is the only unknown parameter among the others $\{\mathbf{a}_{c_{-j}}, \mathbf{a}_s, \boldsymbol{\omega}\}$ where $\mathbf{a}_{c_{-j}} = \{a_{c_1}, ..., a_{c_{j-1}}, a_{c_{j+1}}, ..., a_{c_k}\}$. Under the assumption of known distribution of the noise, Eq. (2.8) for the conditional PDF of a_{c_j} given that $\mathbf{a}_{c_{-j}}, \mathbf{a}_s, \boldsymbol{\omega}, \mathbf{D}$ and σ^2 have already been known becomes a univariate Gaussian distribution:

$$p(a_{c_j} \left| \mathbf{a}_{c_{-j}}, \mathbf{a}_s, \omega, \mathbf{D}, \sigma^2 \right.) \propto \mathcal{N}\left(\hat{a}_{c_j}, \sigma^2 (\mathbf{X}_{a_{c_j}}{}^T \mathbf{X}_{a_{c_j}})^{-1}\right), \tag{2.12}$$

where

$$\hat{a}_{c_j} = \frac{\hat{\mathbf{D}}^{(1)} \mathbf{X}_{a_{c_j}}}{\mathbf{X}_{a_{c_j}}^T \mathbf{X}_{a_{c_j}}}, \quad \mathbf{X}_{a_{c_j}} = \begin{bmatrix} \cos(\omega_j t_1) \\ \vdots \\ \cos(\omega_j t_N) \end{bmatrix} \tag{2.13}$$

and

$$\hat{\mathbf{D}}^{(1)} = \left\{ d_1^{(1)}, d_2^{(1)}, ..., d_N^{(1)} \right\} \tag{2.14}$$

whose components are defined by $d_i - \sum_{l=1}^{k} a_{c_l} \cos(\omega_l t_i)\delta_{lj} + a_{s_l} \sin(\omega_l t_i)$, $(i = 1, 2, 3..., N)$. The $\delta_{lj} = \begin{cases} 1 & l \neq j \\ 0 & l = j \end{cases}$ helps to eliminate the contribution, which comes from the cosine term of the j th sinusoid. When σ^2 is unknown, Eq. (2.12) becomes a univariate Student's t distribution:

$$p(a_{c_j} \left| \mathbf{a}_{c_{-j}}, \mathbf{a}_s, \omega, \mathbf{D}, \sigma^2 \right.) \propto \mathfrak{t}\left(\hat{a}_{c_j}, s^2_{a_{c_j}} (\mathbf{X}_{a_{c_j}}{}^T \mathbf{X}_{a_{c_j}})^{-1}, N-1\right), \tag{2.15}$$

with

$$s^2_{a_{c_j}} = \frac{1}{N-1}(\hat{\mathbf{D}}^{(1)} - \hat{a}_{c_j}\mathbf{X}_{a_{c_j}})^T(\hat{\mathbf{D}}^{(1)} - \hat{a}_{c_j}\mathbf{X}_{a_{c_j}}) \tag{2.16}$$

When $\left\{\mathbf{a}_c, \mathbf{a}_{s_{-j}}, \boldsymbol{\omega}\right\}$ is given, in a similar way, the conditional PDF of a_{s_j} given that $\mathbf{a}_c, \mathbf{a}_{s_{-j}}, \boldsymbol{\omega}, \mathbf{D}$ and σ^2 have already been known is

$$p(a_{s_j}\left|\mathbf{a}_c, \mathbf{a}_{s_{-j}}, \omega, \mathbf{D}, \sigma^2\right.) \propto \mathcal{N}\left(\hat{a}_{s_j}, \sigma^2(\mathbf{X}_{a_{s_j}}{}^T\mathbf{X}_{a_{s_j}})^{-1}\right), \tag{2.17}$$

where

$$\hat{a}_{s_j} = \frac{\hat{\mathbf{D}}^{(2)}\mathbf{X}_{a_{s_j}}}{\mathbf{X}^T_{a_{s_j}}\mathbf{X}_{a_{s_j}}}, \quad \mathbf{X}_{a_{s_j}} = \begin{bmatrix} \sin(\omega_j t_1) \\ \vdots \\ \sin(\omega_j t_N) \end{bmatrix} \tag{2.18}$$

and

$$\hat{\mathbf{D}}^{(2)} = \left\{d_1^{(2)}, d_2^{(2)}, ..., d_N^{(2)}\right\}, \tag{2.19}$$

whose components are defined by $\hat{d}_i^{(2)} = d_i - \sum_{l=1}^{k} a_{c_l}\cos(\omega_l t_i) + a_{s_l}\sin(\omega_l t_i)\delta_{lj}$, $(i = 1, ..., N)$.

When σ^2 is unknown, Eq. (2.17) turns out to be

$$p(a_{s_j}\left|\mathbf{a}_{s_{-j}}, \mathbf{a}_c, \omega, \mathbf{D}, \sigma^2\right.) \propto t(\hat{a}_{s_j}, s^2_{a_{s_j}}(\mathbf{X}_{a_{s_j}}{}^T\mathbf{X}_{a_{s_j}})^{-1}, N-1) \tag{2.20}$$

with

$$s^2_{a_{s_j}} = \frac{1}{N-1}(\hat{\mathbf{D}}^{(2)} - \hat{a}_{s_j}\mathbf{X}_{a_{s_j}})^T(\hat{\mathbf{D}}^{(2)} - \hat{a}_{s_j}\mathbf{X}_{a_{s_j}}). \tag{2.21}$$

To be able to use the theory of GS for the nonlinear parameter $\boldsymbol{\omega}$, we need to introduce some reasonable approximations to linearize the nonlinear model function $f(t_i, \boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ under the condition of the known amplitudes $\{\mathbf{a}_c, \mathbf{a}_s\}$. This can be done by expanding it around $\hat{\boldsymbol{\omega}}$ in a region where the posterior PDF is concentrated:

$$\begin{aligned} f(t_i, \hat{\boldsymbol{\omega}}) \cong & \sum_{l=1}^{k} a_{c_l}\cos(\hat{\omega}_l t_i) + a_{s_l}\sin(\hat{\omega}_l t_i) \\ & + \left(-a_{c_j} t_i \sin(\hat{\omega}_j t_i) + a_{s_j} t_i \cos(\hat{\omega}_j t_i)\right)(\omega_j - \hat{\omega}_j), \end{aligned} \tag{2.22}$$

where $\hat{\omega}_j = \arg\min_{\hat{\omega}\in\boldsymbol{\omega}} \sum_{i=1}^{N} (d_i - f(t_i, \boldsymbol{\omega}))^2$ and $\hat{\boldsymbol{\omega}} = \{\omega_1, ..., \omega_{j-1}, \hat{\omega}_j, \omega_{j+1}, ..., \omega_k\}$. Thus, the conditional PDF of ω_j given that $\boldsymbol{\omega}_{-j}, \mathbf{a}_c, \mathbf{a}_s, \mathbf{D}$ and σ^2 have already been known is a univariate Gaussian distribution:

$$p(\omega_j \,|\, \boldsymbol{\omega}_{-j}, \mathbf{a}_c, \mathbf{a}_s, \mathbf{D}, \sigma^2) \propto \mathcal{N}(\hat{\omega}_j, \sigma^2 (\mathbf{X}_{\omega_j}{}^T \mathbf{X}_{\omega_j})^{-1}), \tag{2.23}$$

where

$$\mathbf{X}_{\omega_j} = \begin{bmatrix} -a_{c_j} t_1 \sin(\hat{\omega}_j t_1) + a_{s_j} t_1 \sin(\hat{\omega}_j t_1) \\ \vdots \\ -a_{c_j} t_N \sin(\hat{\omega}_j t_N) + a_{s_j} t_N \sin(\hat{\omega}_j t_N) \end{bmatrix}. \tag{2.24}$$

If σ^2 is unknown, Eq. (2.23) becomes is a univariate Student's t distribution

$$p(\omega_j \,|\, \boldsymbol{\omega}_{-j}, \mathbf{a}_c, \mathbf{a}_s, \mathbf{D}, \sigma^2) \propto t\left(\hat{\omega}_j, s_{\omega_j}^2 (\mathbf{X}_{\omega_j}{}^T \mathbf{X}_{\omega_j})^{-1}, N-1\right). \tag{2.25}$$

with

$$s_{\omega_j}^2 = \frac{1}{N-1} (\mathbf{D} - \hat{\mathbf{D}})^T (\mathbf{D} - \hat{\mathbf{D}}), \tag{2.26}$$

where $\hat{\mathbf{D}} = \{\hat{d}(t_1), \hat{d}(t_2), ..., \hat{d}(t_N)\}$ whose components are defined by $\hat{d}(t_i) = \sum_{l=1}^{k} a_{c_l} \cos(\hat{\omega}_l t_i) + a_{s_l} \sin(\hat{\omega}_l t_i)$

A systematic form of GS algorithm [11, 12, 10] contains choosing initially arbitrary starting values $\{\mathbf{a}_c^{(0)}, \mathbf{a}_s^{(0)}, \boldsymbol{\omega}^{(0)}\}$ and drawing successively random samples from the full conditional distributions:

$$\begin{aligned} a_{c_j}^{(1)} &\sim p(a_{c_j} \,|\, \{a_{c_1}^{(1)}, ..., a_{c_{j-1}}^{(1)}, a_{c_{j+1}}^{(0)}, ..., a_{c_k}^{(0)}\}, \mathbf{a}_{s_j}^{(0)}, \omega^{(0)}, \mathbf{D}) \\ a_{s_j}^{(1)} &\sim p(a_{s_j} \,|\, \mathbf{a}_{c_j}^{(1)}, \{a_{s_1}^{(1)}, ..., a_{s_{j-1}}^{(1)}, a_{s_{j+1}}^{(0)}, ..., a_{s_k}^{(0)}\}, \omega^{(0)}, \mathbf{D}) \\ \omega_j^{(1)} &\sim p(\omega_j \,|\, \mathbf{a}_{c_j}^{(1)}, \mathbf{a}_{s_j}^{(1)}, \{\omega_1^{(1)}, ..., \omega_{j-1}^{(1)}, \omega_{j+1}^{(0)}, ..., \omega_k^{(0)}\}, \mathbf{D}), (j = 1, ..., k). \end{aligned} \tag{2.27}$$

At each iteration of the Gibbs sampler, we cycle through the set of conditional distributions and draw one sample from each. When a sample is drawn from one conditional distribution, the succeeding distributions are updated with the new value of that sample. At the K'th iteration we obtain the following drawings:

$$
\begin{aligned}
a_{c_j}^{(K+1)} &\sim p(a_{c_j} \left| \{a_{c_1}^{(K+1)}, ..., a_{c_{j-1}}^{(K+1)}, a_{c_{j+1}}^{(K)}, ..., a_{c_k}^{(K)}\}, \mathbf{a}_{s_j}^{(K)}, \omega_j^{(K)}, \mathbf{D}\right.) \\
a_{s_j}^{(K+1)} &\sim p(a_{s_j} \left| \mathbf{a}_{c_j}^{(K+1)}, \{a_{s_1,}^{(K+1)}, ..., a_{s_j-1}^{(K+1)}, a_{s_{j+1}}^{(K)}, ..., a_{s_k}^{(K)}\}, \omega_j^{(K)}, \mathbf{D}\right.) \\
\omega_j^{(K+1)} &\sim p(\omega_j \left| \mathbf{a}_{c_j}^{(K+1)}, \mathbf{a}_{s_j}^{(K+1)}, \{\omega_1^{(K+1)}, ..., \omega_{j-1}^{(K+1)}, \omega_{j+1}^{(K)}, ..., \omega_k^{(K)}\}, \mathbf{D}\right.).
\end{aligned}
\tag{2.28}
$$

For a large enough K, $a_c^{(K+1)}$, $a_s^{(K+1)}$ and $\omega^{(K+1)}$ can be considered as random variables drawn from their posterior PDF distributions. Therefore we are able to generate samples of these posterior PDFs for each parameter. Using these samples, all of the estimates about the their corresponding can then be found, such as the most probable values for them, the mean value, the marginal variances with respect to the most probable value etc. When σ^2 is unknown, we do the same thing as above except that the random numbers are drawn from the Student's t distribution.

2.5 Cramer–Rao Lower Bound

Given an estimation problem, one may ask: What is the variance of the best possible unbiased estimator? The answer is given by the Cramer–Rao lower bound (CRLB) [24, 17], which we will study in this section and it provides a theoretical lower limit for variance of estimator. If we consider the parameter vector θ and the signal to noise ratio (SNR), then the CRLB to the variance of unbiased estimator of the parameters θ for the signal model is determined in the form:

$$
\mathrm{Var}_{\boldsymbol{\theta}}(\hat{\theta}) \geq \mathrm{CRLB}(\theta) = \mathbf{J}^{-1}(\theta), \tag{2.29}
$$

where Fisher information matrix $\mathbf{J}(\theta)$ [24] is defined as an expectation of the second derivatives of the log likelihood function with respect to θ:

$$
\mathbf{J}(\theta) = E\left[-\frac{\partial^2 \ln P(\mathbf{D}|\theta, I)}{\partial \theta^2} \right]. \tag{2.30}
$$

for large N, $\mathbf{J}(\theta)$ is a diagonal matrix and its inversion is straightforward. The diagonal elements of its inversion yield the lower bound on the variance of the estimates asymptotically. When the noise is white Gaussian, we can use the an alternative form of CRLB which is easier than the general case in Eq. (2.30). In this case the Fisher information matrix becomes

$$
\mathbf{J}(\theta) = \frac{1}{\sigma^2} \sum_{j=1}^{N} \frac{\partial f_j(\theta)}{\partial \theta} \left(\frac{\partial f_j(\theta)}{\partial \theta} \right)^T. \tag{2.31}
$$

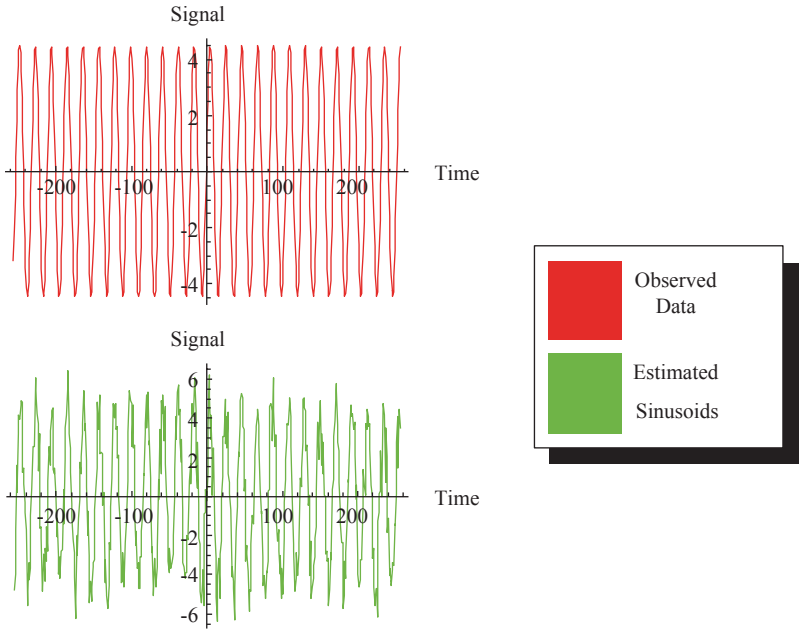


Fig. 2.1 Recovering signal from noisy data produced from a single harmonic frequency signal model

2.6 Computer Simulations

To demonstrate the proposed approach with examples which are used by previous researches [4, 11, 12, 38, 36, 37], we firstly created data samples according to a signal model with a single frequency:

$$d_i = 2\cos(0.3t_i) + 4\sin(0.3t_i) + e_i \quad (i = 1,...,512) \tag{2.32}$$

Here i runs in a symmetric time interval $-T$ and T (2T+1=N) and $e_i \sim N(0,1)$. We obtained noisy data samples ($N = 512$), shown in Fig. 2.1 and carried out Bayesian analysis. The proposed method requires initial values for the parameters to start the iteration. Instead of choosing them randomly from a uniform distribution [16], we first performed a Fast Fourier Transformation (FFT) of the data and then chose the locations of the peaks in the power spectrum density [4, 16], which is a squared magnitude of FFT, as an initial estimate for the frequencies. Once, initial frequencies were obtained, we carried on calculating the coefficients $\mathbf{a}_c$ and $\mathbf{a}_s$ as initial values for the amplitudes, respectively. The algorithm of GS, introduced in the paper was coded in *Mathematica* programming language and run on a workstation in two cases where the standard deviation of noise is known or not. In the case where the deviation of noise is unknown, the output of the computer simulation is illustrated in Table 2.1. The estimated parameter values are quoted

Table 2.1 Computer simulations for a single harmonic frequency model

Parameters	True values	Estimated values
ω_1	0.3	0.2999 ± 0.00009
a_{c_1}	2	2.041 ± 0.0623
a_{s_1}	4	3.992 ± 0.0628

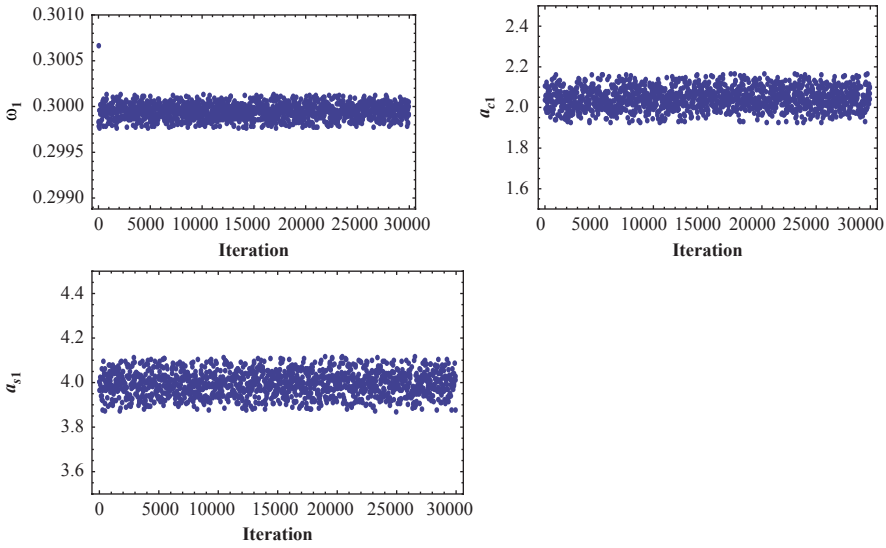


Fig. 2.2 MCMC parameter iterations

as (*value*) ± (*standard deviation*) and used to regenerate the given signal model, shown in Fig. 2.1.

It can be seen that a single frequency and its corresponding amplitudes are recovered very well.

In order to determine its convergence, there are several diagnostic tests [5, 34] we can do, both visual and statistical, to see if the chain appears to be converged. One intuitive and easily implemented diagnostic tool is a trace plot (or history plot) [34] which is a plot of the iteration number against the value of the draw of the parameter. If it has converged, the trace plot will move up and down around the mode of the distribution and the distribution of the parameters settles down to the target posterior PDF from which statistical inferences about the parameters can be made. A clear sign of non-convergence occurs when we observe some trending in the trace plot. In this case we can see whether our chain gets stuck in certain areas of the parameter space, which indicates bad mixing. Figure 2.2 shows the scatter plots of the model parameters, ω_1, a_{c_1} and a_{s_1}, respectively and indicates that the GS samples are densely placed around the estimated values of these parameters.

In our second example, we consider a signal model with two close harmonic frequencies:

Table 2.2 Computer simulations for two closed harmonic frequency model

Parameters	True values	Estimated values
ω_1	0.3	0.3001±0.0004
ω_1	0.31	0.3108±0.0004
a_{c_1}	0.5403	0.4821±0.0645
a_{c_2}	−0.4161	−0.3852±0.0642
a_{s_1}	−0.8415	−0.83±0.0644
a_{s_2}	−0.9093	−0.9005±0.0647

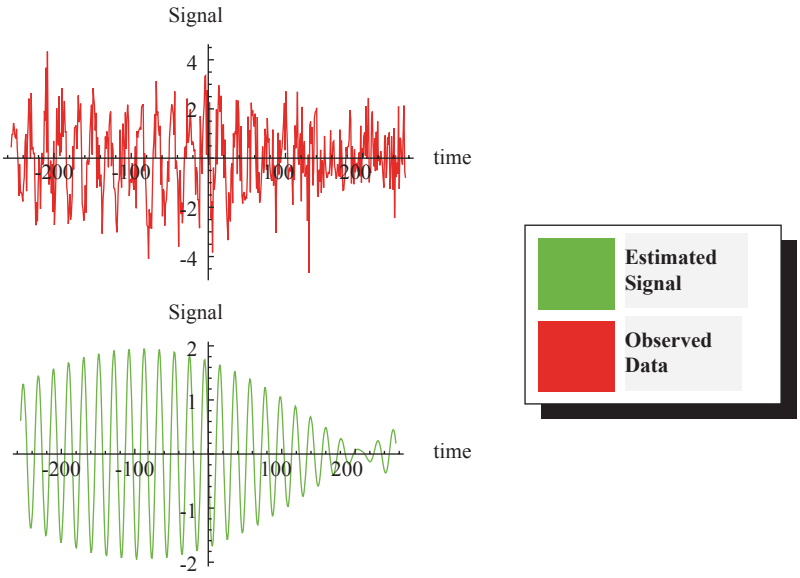


Fig. 2.3 Recovering signals from noisy data produced from two closed harmonic frequency signal model

$$d_i = 0.5403\cos(0.3t_i) - 0.8415\sin(0.3t_i) \\ -0.4161\cos(0.31t_i) - 0.9093\sin(0.31t_i), \tag{2.33}$$

In a similar way, we produced the same size data corrupted by the zero mean Gaussian noise with $\sigma = 1$, ran Mathematica code again in the case where the deviation of noise is unknown and obtained the results shown in Table 2.2. It indicates that all values of the parameters within the calculated accuracy are clearly recovered.

On the other hand, Fig. 2.3 shows the power of the method for recovering the signal from the noisy data using the estimated values of the parameters of sinusoids.

In general, we consider a signal model with five harmonic frequencies

Table 2.3 Computer simulations for five harmonic frequency signal model

Parameters	True values	Estimated values
ω_1	0.1	0.09979 ± 0.0003
ω_2	0.15	0.1498 ± 0.0002
ω_3	0.3	0.2999 ± 0.0008
ω_4	0.31	0.3095 ± 0.0002
ω_5	1	1.000 ± 0.0001
a_{c_1}	0.540302	0.6542 ± 0.0618
a_{c_2}	−0.832294	-0.8582 ± 0.0620
a_{c_3}	−4.94996	-4.756 ± 0.0624
a_{c_4}	−1.30729	-1.4 ± 0.06225
a_{c_5}	0.850087	0.7496 ± 0.0629
a_{s_1}	−0.841471	-0.9683 ± 0.0627
a_{s_2}	−1.81859	-1.82 ± 0.0631
a_{s_3}	−0.7056	-0.8228 ± 0.0624
a_{s_4}	1.5136	1.443 ± 0.0627
a_{s_5}	2.87677	2.8227 ± 0.0630

$$\begin{aligned} d_i = {} & \cos(0.1\ \mathrm{t}_i + 1) + 2\cos(0.15t_i + 2) \\ & + 5\cos(0.3t_i + 3) + 2\cos(0.31t_i + 4) + 3\cos(t_i + 5) + e_i \end{aligned} \tag{2.34}$$

The best estimates and the standard deviations with true values for all the parameters are tabulated in Table 2.3.

It is obvious that our results are closer to the true values and all the frequencies have been well resolved, even the third and fourth frequencies which are too closed are not to be separated by the Fourier power spectral density shown in Fig. 2.4.

These estimated values of parameters are used to regenerate the given signal model, shown in Fig. 2.4. In the case where the standard deviation of the noise is known, we obtained almost similar results.

The usual way the result from a spectral analysis is displayed is in the form of a power spectral density. Therefore, a comparison of Bayesian and Fourier spectral densities shown in Fig. 2.4 indicate separation of frequencies. DFT spectral density shows only four peaks among five frequencies but, Bayesian spectral density indicates five frequencies with high accuracies.

Moreover, we initially assumed that the values of the random noise in data were drawn from the Gaussian density. Figure 2.5 shows the exact and estimated PDF of the random noise in data. It is seen that the estimated (dotted) PDF is closer to the true (solid) PDF and the histogram of the errors, which is known as nonparametric estimator is also much closer to its true probability density. These results demonstrate how powerful Bayesian method is to separate noise from data.

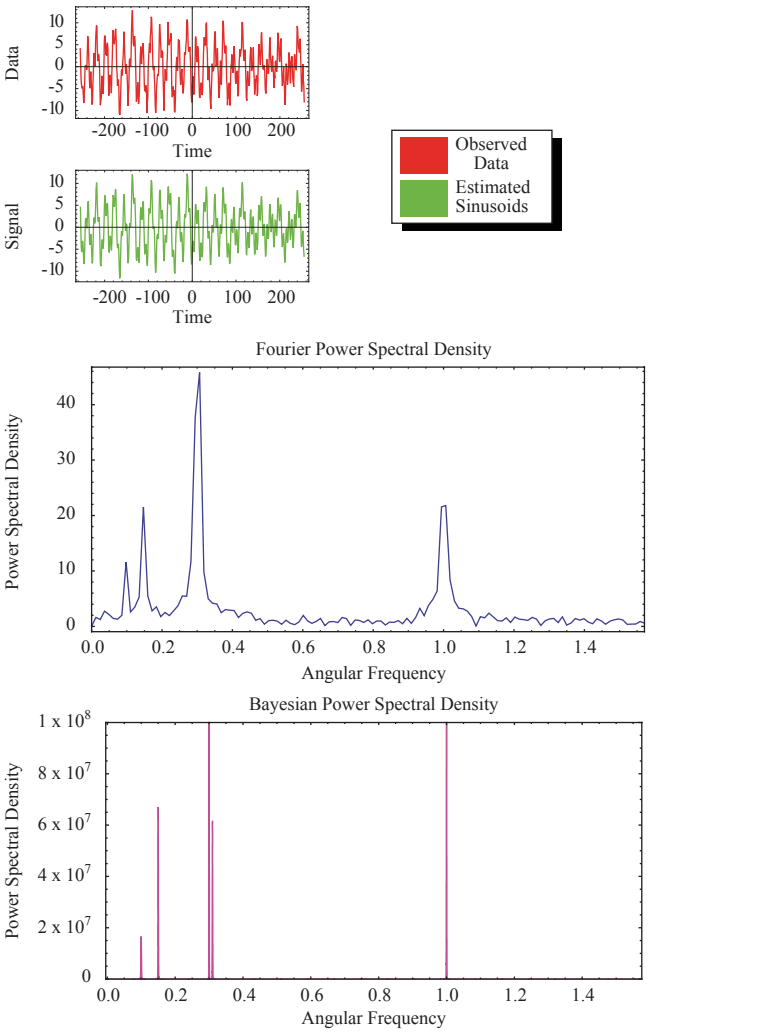


Fig. 2.4 Spectral analysis of multiple frequency model

The computational complexity of the GS algorithm is dependent on the length of data samples, a number of the parameters and few parameters that control convergence of the GS such as iteration number and size of samples needed to summarize the marginal posterior distribution of each parameter. Figure 2.6 shows only CPU time of different simulations for a variety of number of data samples and parameters and indicates that an increase in these numbers causes larger consumption of CPU time.

In order to evaluate the performance of GS, computer simulations were performed and compared with the classical estimators such as ML and DFT, as well as CRLBs of ω and α[1] expressed in decibel (dB):

[1] $\alpha = \sqrt{a_{c_1}^2 + a_{s_1}^2}$

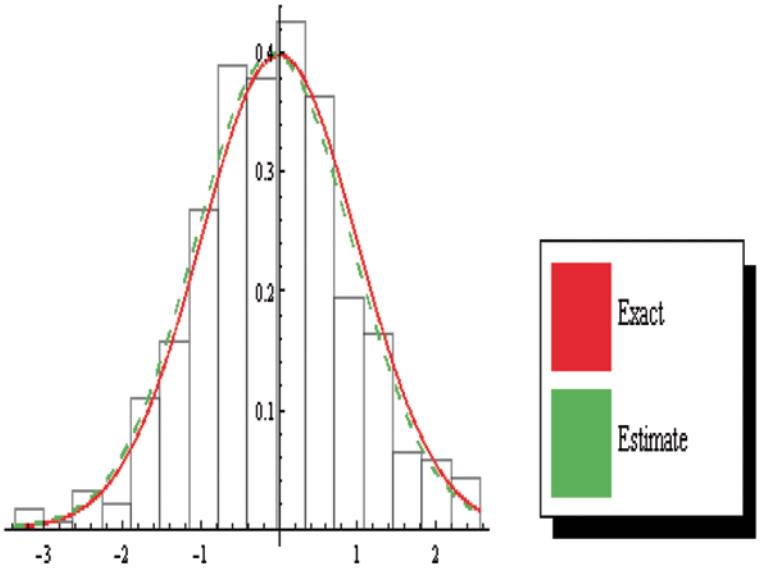
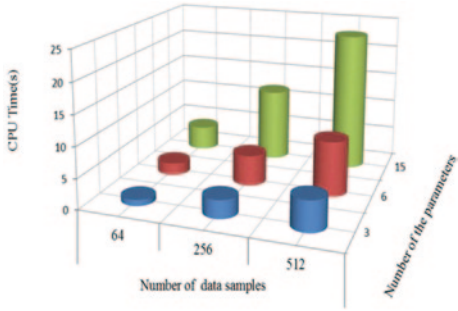


Fig. 2.5 Comparison of exact and estimate probability densities of noise in data

Fig. 2.6 Different simulation times with respect to number of parameters and data samples



$$CRLB(\omega) \simeq SNR + 10Log(N^3/12) \qquad (2.35)$$
$$CRLB(\alpha) \simeq 10Log(N) + 10Log(1/2\sigma^2),$$

which are a function of N and SNR. We fixed α to $\sqrt{2}$ and properly scaled $e(t_i)$ to obtain different SNRs, defined as $\text{SNR} = 10Log\dfrac{\alpha^2}{2\sigma^2}$. Unless stated otherwise, the angular frequency is chosen as $\omega = 0.3\pi$ and SNR = 20dB.

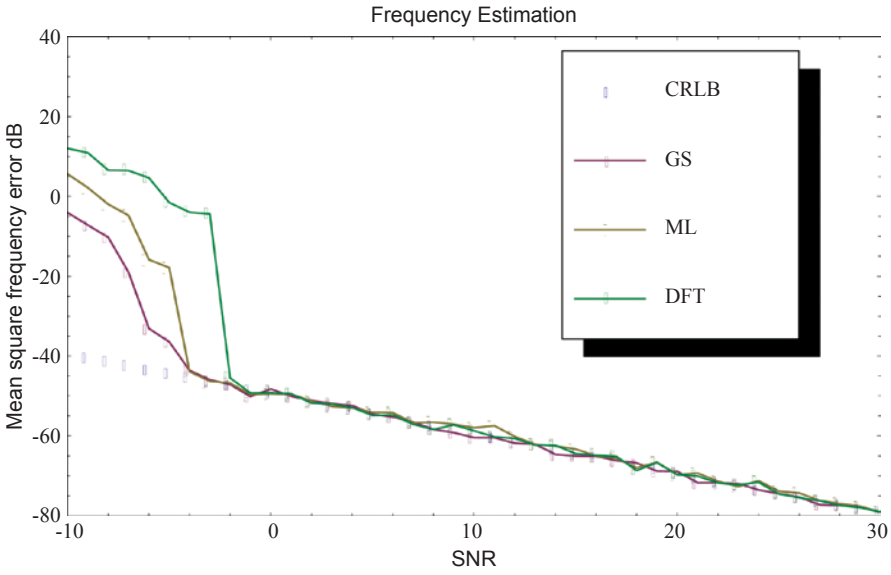


Fig. 2.7 Mean square frequency error versus SNR at $\omega = 0.3\pi$

We generated $N = 100$ data samples from a single real tone frequency signal model in a variety of noise levels. After 50 independent trials under the same noise level, MSEs of the estimated frequency and amplitude were obtained for each method. Their logarithmic values were plotted with respect to SNR ratios, which vary from −10 to 30 dB and shown in Figs. 2.7 and 2.8. They indicate the MSE performances for different estimators. The error curves in these figures were separated into two regions. The first one, on the left, shows that the estimator variances increase stronger than the CRLB and contain smaller threshold effects in Fig. 2.8 than that of Fig. 2.7. The second one, on the out of left indicates that the errors follow the CRLB and the curves close to it. In Fig. 2.7, it can be seen that GS, ML and DFT estimators have threshold about −5, −4 and −2 dB of the SNR, respectively and follow nicely with the CRLB after −1 dB. As expected, with increasing SNR, MSE values approaches to the CRLB but, with decreasing SNR, they get worse from it. This implies the higher the SNR, the lower CRLB. Moreover, all three estimators have same characters at high SNRs**.** The above argument treats only with the case in which a size of data samples $N = 100$ is used for the estimation. Therefore, one may ask how to vary accuracy of the estimation with N. To answer it, we set up an experiment in which the algorithms of three methods were run for 50 simulated data with different lengths. In this case Figs. 2.9 and 2.10 show the MSE performances of three estimators with different data length which varies from $N = 25$ to $N = 300$ at $\omega = 0.3\pi$ under SNR = 20 dB.

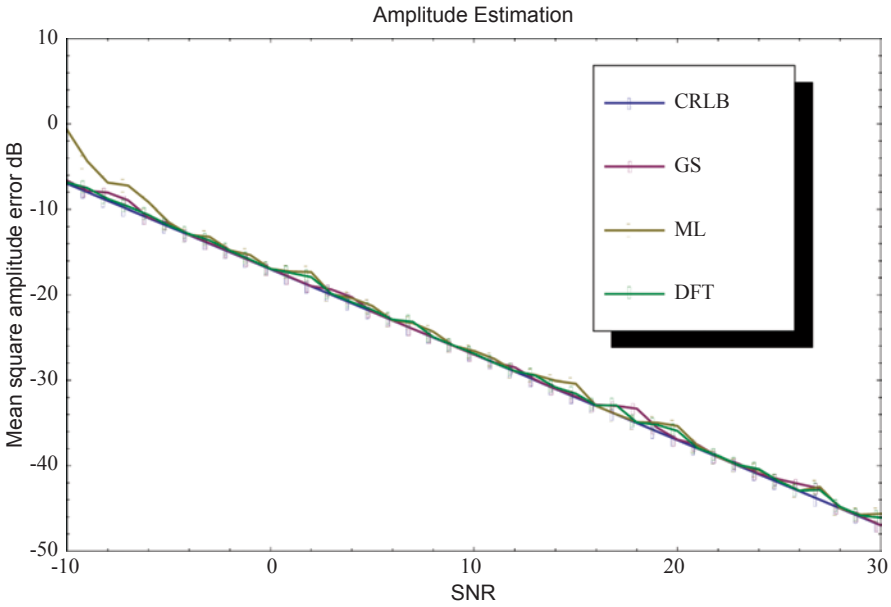
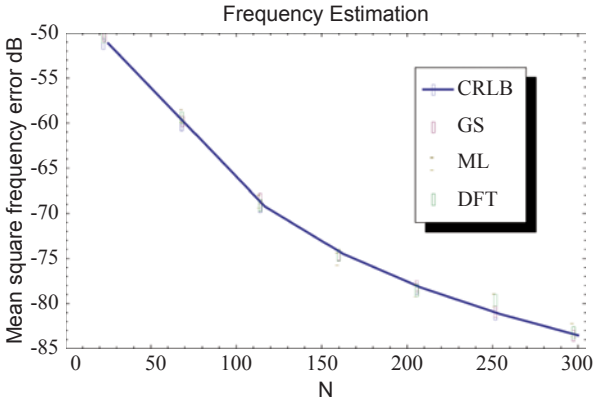


Fig. 2.8 Mean square amplitude error versus SNR at $\mathbf{a} = \sqrt{2}$

Fig. 2.9 Mean square frequency error versus N at $\omega = 0.3\pi$



They indicate that the larger data samples give the lower MSEs of frequencies than that of amplitudes. This implies that all three estimators are more effective for the frequency estimation, rather than amplitude estimation. On the other hand, if the estimator reaches the CRLB, it is called efficient. Therefore an efficiency parameter [31], defined as

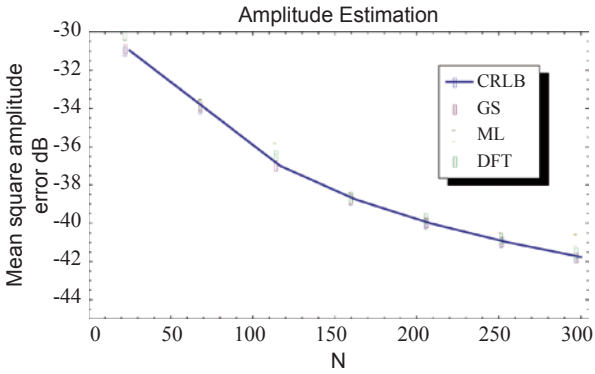


Fig. 2.10 Mean square amplitude error versus N at $\mathbf{a} = \sqrt{2}$

Table 2.4 Performance comparison of Bayesian methods for frequency estimation in single frequency signal model

SNR = 30 *dB*			*N* = 300	
Methods	MSE(dB)	Efficiency	MSE(dB)	Efficiency
GS	−79.175	100.041	−83.484	100.045
ML	−79.079	100.162	−82.850	100.81
DFT	−79.103	100.133	−83.401	100.144
CRLB	−79.208	100	−83.521	100

Table 2.5 Performance comparison of Bayesian methods for amplitude estimation in single frequency signal model

SNR = 30 *dB*			*N* = 300	
Methods	MSE(dB)	Efficiency	MSE(dB)	Efficiency
GS	−46.9800	100.021	−41.7415	100.046
ML	−45.6409	102.955	−40.8652	102.192
DFT	−46.0871	101.958	−40.8450	100.381
CRLB	−46.9897	100	−41.7609	100

$$\eta_i = \frac{CRLB_i}{MSE_i} \times 100, \tag{2.36}$$

indicates the closeness of estimators to the CRLB. Tables 2.4 and 2.5 contain the MSEs and the efficiency values η for the frequency and amplitude estimation obtained at the last states where SNR = 30dB and $N = 300$.

It can be seen that the efficiency value for the GS at SNR = 30dB and for $N = 300$ is much closer to the CRLB than that of the other methods in both frequency and amplitude estimation. Thus, it is said that GS is more effective than the others for higher SNR and larger data sample.

2.7 Conclusions

In this chapter we have presented a numerical procedure, namely Gibbs sampling, based on Bayesian inference for estimating parameters of multiple sinusoids embedded in noise. Overall results show that Bayesian approach can not only give us the best estimates for the parameters but, it can also tell us uncertainties associated with their estimated values. Experiments with synthetic signals show that GS performs frequency estimations with a high-resolution, according to the CRLB. On the other hand, it requires a maximization of full conditional marginal probability density of frequencies that can be difficult if SNR is low. Comparing with classical estimators such as ML and DFT, all three methods can give similar performance in higher SNRs and larger N.

The problem of detection of number of sinusoids which is a big part of spectral analysis is not included in this work but, Bayesian inference helps us to accomplish it. Therefore, it will deserve further investigations.

Acknowledgements This work has been supported by the Research Fund of Istanbul University with project numbers are UDP-33672 and YADOP-19681.

References

1. Andreiu C, Doucet A (1999) Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC, IEEE Transactions on Signal Processing, 47: 2667–2676
2. Bernardo JM, Smith AFM (2000) Bayesian theory, Willey Series in Probability and Statistics New York
3. Box GEP, Tiao C (1992) Bayesian inference in statistical analysis, New York
4. Bretthorst GL (1997) Bayesian spectrum analysis and parameter estimation, Lecture Notes in Statistics, Springer-Verlag Berlin Heidelberg New York
5. Brooks SP, Gelman A (1997) General methods for monitoring convergence of iterative simulations, Journal of Computational and Graphical Statistics, 7: 434–455
6. Cevri M, Ustundag D (2012) Bayesian recovery of sinusoids from noisy data with parallel tempering, IET Signal Process., 6 (7): 673–683
7. Cevri M, Ustundag D (2013) Performance analysis of Gibbs sampling for Bayesian extracting sinusoids, Proceedings of the 2013 International Conference on Systems, Control, Signal Processing and Informatics, Rhodes Island, Greece, 128–134
8. Cooley JW, Tukey, JW (1965) An algorithm for the machine calculation of complex Fourier series, Mathematics of Computation, 19: 297–301
9. Cox RT (1946) Probability, frequency, and reasonable expectation, American Journal of Physics, 14: 1–13
10. Diaconis P, Khare K, Coste LS (2008) Gibbs sampling, exponential families and orthogonal polynomials, Statistical Science, 23(2): 151–178
11. Dou L, Hodgson RJW (1995a) Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation I, Inverse Problem, 11:1069–1085
12. Dou L, Hodgson RJW (1995b) Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation II, Inverse Problem, 11:121–137
13. Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J. Amer. Statist. Assoc., 85:398–409

14. Gelman AB, Stern HS, Rubin DB (1995) Bayesian data analysis, Chapman & Hall/CRC
15. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. IEE Transactions on Pattern Analysis and machine Intelligence, 6:721–741
16. Gregory P (2005) Bayesian logical data analysis for the physical science, Cambridge University Press, United Kingdom
17. Händel P (2008) Parameter estimation employing a dual-channel sine-wave model under a Gaussian assumption, IEEE Transactions on Instrumentation and Measurement, 57(8): 1661–1669
18. Harney HL (2003) Bayesian inference: Parameter estimation and decisions, Springer-Verlag, Berlin Heidelberg
19. Hastings, W K (1970) Monte carlo sampling methods using Markov chains, and their applications, Biometrika, 57:97–109
20. Jackman S (2000) American journal of political science, 44(2):375–404
21. Jaynes ET (1987) Bayesian Spectrum and Chirp Analysis, In Proceedings of the Third Workshop on Maximum Entropy and Bayesian Methods, Ed. C. Ray Smith and D. Reidel, Boston, pp. 1–37
22. Jaynes ET (2003) Probability theory: The logic of science, Cambridge University Press, United Kingdom
23. Kay SM (1984) Accurate frequency estimation at low signal-to-noise ratio, IEEE Trans. Acoust., Speech, Signal Processing, ASSP-32, pp. 540–547
24. Kay SM (1993) Fundamentals of statistical signal processing: Estimation theory, Prentice-Hall, Englewood Cliffs, NJ, pp. 56–57
25. Kenefic RJ, Nuttall AH (1987) Maximum likelihood estimation of the parameters of tone using real discrete data, IEEE J. Oceanic Eng., 12 (1): 279–280
26. MacKay D (2003) Information theory, inference and learning algorithms, Cambridge University Press
27. Metropolis N, Rosenbluth A, Rosenbluth, M, Teller A, Teller E (1953) Equation of states calculations by fast computing machines, Journal of chemical physics, 21: 1087–1092
28. Michalopouloua ZH, Picarelli M (2005) Gibbs sampling for time-delay-and amplitude estimation in underwater acoustics, J. Acoust. Soc. Am., 117:799–808
29. Quinn BG (1994) Estimating frequency by interpolation using Fourier coefficients, IEEE Trans. Signal Process., 42 (5): 1264–1268
30. Rife DC, Boorstyn RR (1974) Single-tone parameter estimation from discrete-time observations, IEEE Transactions on Information Theory, 20:591–598
31. Ristic B, Arulampalam S, Gordon N (2004) Beyond the Kalman filter particle filters for tracking applications, Artech House, London
32. Stoica P, Moses RL (2005) Spectral analysis of signals, Prentice Hall
33. Swendsen RH, Wang JS (1986) Physical review of letters, 57:2607–2609
34. Tanner, MA (1996) Tools for Statistical Inference, 3rd ed. Springer-Verlag, New York
35. Tanner M, Wong W (1987) The calculation of posterior distributions by data augmentation (with discussion), J. Amer. Statist. Assoc., 82:528–550
36. Üstündağ D, Cevri M (2008) Estimating parameters of sinusoids from noisy data using Bayesian inference with simulated annealing, Wseas Transactions On Signal Processing, 7: 432–441
37. Üstündağ D, Cevri M (2011) Recovering sinusoids from noisy data using Bayesian inference with simulated annealing, Mathematical & Computational Applications, 16(2): 382–391
38. Ustundag D, Cevri M (2012) Simulated annealing—advances, applications and hybridizations, In Tech, Croatia, ISBN: 978-953-51-0710-1. pp. 67–90
39. Ustundag D, Cevri M (2013) Comparison of Bayesian methods for recovering sinusoids, Proceedings of the 2013 International Conference on Systems, Control, Signal Processing and Informatics, Rhodes Island, Greece, 120–127

Chapter 3
Controlling Chaotic Systems Via Time-Delayed Control

R. Farid, A. Ibrahim and B. Abou-Zalam

Abstract Based on Lyapunov stabilization theory, this paper proposes a proportional plus integral time-delayed controller to stabilize unstable equilibrium points (UPOs) embedded in chaotic attractors. The criterion is successfully applied to the classic Chua's circuit. Theoretical analysis and numerical simulation show the effectiveness of this controller.

Keywords Chaotic systems · Proportional plus integral time-delayed controller · Taylor approximation

3.1 Introduction

Dynamic chaos is a very interesting non-linear effect which has been intensively studied in science and engineering. The effect is very common, it has been detected in a large number of dynamic systems of various physical nature. However, this effect is usually irregular, complex and undesirable in practice, and it restricts the operating range of many electronic and mechanic devices. Recently, controlling this kind of complex dynamical systems has attracted a great deal of attention within the engineering society. Chaos control, in a broader sense, can be divided into two categories: one is to suppress the chaotic dynamical behavior [1–12] and the other is to generate or enhance chaos in nonlinear systems [13, 14]. Nowadays, different techniques and methods have been proposed to achieve chaos control. Among many methods, the time delayed feedback control DFC method [1]. This method utilizes the difference between the states and the delayed states as an input control provided that the delayed time is determined as the period of the unstable periodic orbits UPO to be stabilized. Furthermore [2, 3] also proposed a DFC based controller to stabilize the UPOs by virtue of the iterative learning control strategy. A time-delayed integrity controller is proposed in [4] to ensure the stabilization of UPOs in the case of sensor failures. Sliding mode control of uncertain unified chaotic systems is proposed in [5] based on a proportional plus integral sliding surface dislocated and

R. Farid (✉) · A. Ibrahim · B. Abou-Zalam
Department of industrial electronics and control engineering, Faculty of electronic engineering, Menofia University, Menuf, Egypt
e-mail: ramy5475@yahoo.com

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,* Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_3,

enhancing feedback control [6, 7] which multiply the independent variable of the system function with coefficient and take the result as feedback gain (the same coefficient for all states) based on Jacobi matrix, speed feedback control [6–8] multiply the derivative of independent variable with coefficient. And other feedback control techniques [9–12]. At the same time, chaos synchronization also is an important topic, and has obtained a lot of availability results [15–19].

The aim of this paper is to proposed new scheme of time-delayed controller based on proportional plus integral (PI) to stabilize unstable equilibrium points (UPOs) embedded in chaotic attractors based on Lyapunov stabilization theory.

The reset of the letter is organized as follows. In Sect. 3.2 the control problem is stated. In Sect. 3.3 PI time-delayed controller is proposed to stabilize UPOs using Lyapunov stabilization theory. The proposed controller is applied with numerical simulation to the classic Chua's circuit in Sect. 3.4. Finally, some conclusions are given in Sect. 3.5.

3.2 Problem Statement

Considering a chaotic system with state equation in the form

$$\dot{x} = Ax + g(x) \tag{3.1}$$

Where $x \in R^n$ is the state vector, $A \in R^{nxn}$ is constant matrix and $g(x)$ is a nonlinear vector on the state vector x.

Assuming that

$$g(x) - g(\tilde{x}) = M_{x,\tilde{x}}(x - \tilde{x}) \tag{3.2}$$

For a bounded matrix $M_{x,\tilde{x}}$ in which the elements are dependent on x and $\tilde{x}$. Most of the chaotic systems can be described by Eqs. (3.1 and 3.2).

General speaking, chaotic systems can be decomposed into a linear part and nonlinear function vector part.

Among many chaotic systems,

Lorenz system [14]

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -\rho \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 0 \\ -xz \\ xy \end{bmatrix} \tag{3.3}$$

RÖssler system [14]

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & a & 0 \\ 0 & 0 & -b \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ c + xz \end{bmatrix} \tag{3.4}$$

Chua system [14]

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} -\alpha & \alpha & 0 \\ 1 & -1 & 1 \\ 0 & -\beta & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} \alpha f(x) \\ 0 \\ 0 \end{bmatrix} \tag{3.5}$$

With

$$f(x) = bx + \frac{1}{2}(a-b)(|x+1| - |x-1|) \tag{3.6}$$

Have the same form as Eq. (3.1).

Our problem undertaken here is to construct a controller

$$u(t) = \left(u_1(t), u_2(t), ..., u_n(t)\right)^T \in R^{n\times 1}$$

To stabilize the UPOs within chaotic attractors. Therefore the controlled chaotic system can be described by

$$\dot{x}(t) = Ax(t) + g(x(t)) + u(t) \tag{3.7}$$

Suppose that the UPO to be stabilized is T-periodic, i.e.

$$\dot{x}(t-T) = Ax(t-T) + g(x(t-T)) \tag{3.8}$$

Defining the state error as

$$e(t) = x(t) - x(t-T)$$

The error dynamics is

$$\dot{e}(t) = Ae(t) + g(x(t)) - g(x(t-T)) + u(t)$$

Based on Eq. (3.2)

$$\dot{e}(t) = (A + M_{x(t),x(t-T)})e(t) + u(t) \tag{3.9}$$

With the help of the controller $u(t)$, the problem of stabilization of the T-periodic orbit becomes the problem of stabilization of Eq. (3.9) to either a periodic or equilibrium points.

3.3 Controller Design

A proportional plus integral time-delayed controller is proposed to stabilize UPOs embedded in chaotic attractors.

Controller $u(t)$ is chosen as

$$u(t) = -k_p(x(t) - x(t-T)) - k_i \int_{t_1}^{t_2} (x(t) - x(t-T))dt \tag{3.10}$$

Where k_p and k_i are diagonal matrices with diagonal gain elements $k_{p1}, k_{p2}, ..., k_{pn}$ and $k_{i1}, k_{i2}, ..., k_{in}$ respectively.

Theorem 1 if controller $u(t)$ is constructed as Eq. (3.10), then the error system Eq. (3.9) is globally exponentially stable for $T << 1$. if there exists a positive definite symmetric constant matrix P such that

$$(A + M_{x(t),x(t-T)} - k_p - Tk_i)^T P + P(A + M_{x(t),x(t-T)} - k_p - Tk_i) \leq \mu I < 0 \tag{3.11}$$

Where μ denotes a negative constant, and I is the identity matrix

Proof For $T << 1$, $x(t-T) = x(t) - \dot{x}(t)T + o(t^2)$, then by Taylor approximation, we have $x(t) - x(t-T) = \dot{x}(t)T$, so the controller $u(t)$ of Eq. (3.10) becomes

$$u(t) = -k_p(x(t) - x(t-T)) - k_i T \int_{t1}^{t2} \dot{x}(t)dt \tag{3.12}$$

For $t_2 = t$ and $t_1 = t - T$

$$u(t) = -k_p(x(t) - x(t-T)) - k_i T((x(t) - x(t-T)))$$
$$u(t) = -k_p e(t) - k_i T e(t) \tag{3.13}$$

By constructing (3.13) into (3.9)

$$\dot{e}(t) = (A + M_{x(t),x(t-T)} - k_p - Tk_i)e(t) \tag{3.14}$$

Now choose the Lyapunov function

$$V = e^T P e \tag{3.15}$$

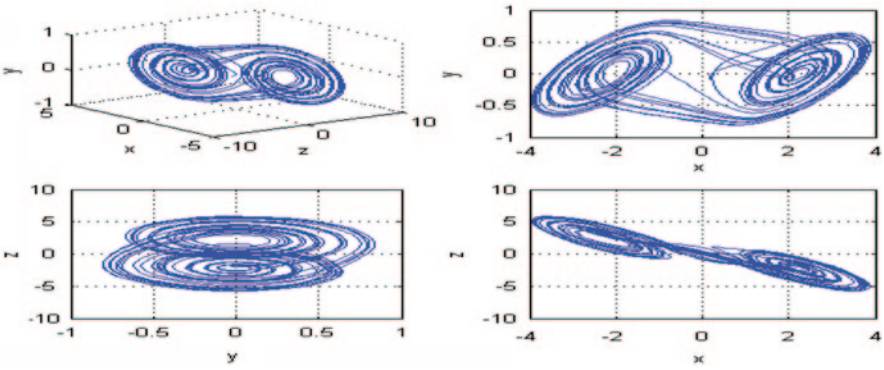


Fig 3.1 The attractors of Chua's circuit

Where P is a positive definite symmetric constant matrix. Then its derivative is

$$\begin{aligned}\dot{V} &= \dot{e}^T Pe + e^T P\dot{e} \\ &= [(A + M_{x(t),x(t-T)} - k_p - Tk_i)e]^T Pe + e^T P[(A + M_{x(t),x(t-T)} - k_p - Tk_i)e] \\ &= e^T[(A + M_{x(t),x(t-T)} - k_p - Tk_i)^T P + P(A + M_{x(t),x(t-T)} - k_p - Tk_i)]e \\ &\le \mu\|e\|^2 < 0\end{aligned} \tag{3.16}$$

Where $\|\cdot\|$ denotes the Euclidean norm.

3.4 Numerical Simulation

To demonstrate the use of chaos control criterion proposed herein, Chua's circuit is considered as an example of chaotic systems.

3.4.1 Chua's Circuit

Chua's circuit can be described by (5) and (6), where $\alpha > 0, \beta > 0, a < b < 0, f(.)$ is a piecewise linear function. Chua's circuit exhibits a chaotic behavior for $\alpha = 9.78, \beta = 14.97, a = -1.31$ and $b = -0.75$ as shown in Fig. 3.1, and in Eq. (3.6), we have

$$f(x(t)) - f(x(t-T)) = k_{x(t),x(t-T)}(x(t) - x(t-T)) \tag{3.17}$$

Where $k_{x(t),x(t-T)}$ is dependent on $x(t)$ and $x(t-T)$, and varies in the interval $[a,b]$ for $t \ge 0$ that is, $k_{x(t),x(t-T)}$ is bounded by the condition of $a \le k_{x(t),x(t-T)} \le b < 0$ graphical representation of $f(x)$ in [14].

System (1.5) has the same form of Eq. (3.1) with

$$A=\begin{bmatrix}-\alpha & \alpha & 0\\ 1 & -1 & 1\\ 0 & -\beta & 0\end{bmatrix},\ g(x(t))=\begin{bmatrix}-\alpha f(x(t))\\ 0\\ 0\end{bmatrix}$$

Consider

$$\begin{aligned} g(x(t))-g(x(t-T)) &= \begin{bmatrix}-\alpha(f(x(t)-f(x(t-T)) & 0 & 0\end{bmatrix}^T\\ &= \begin{bmatrix}-\alpha k_{x(t),x(t-T)}(x(t)-x(t-T)) & 0 & 0\end{bmatrix}^T\\ &= \begin{bmatrix}-\alpha k_{x(t),x(t-T)} & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}x(t)-x(t-T)\\ y(t)-y(t-T)\\ z(t)-z(t-T)\end{bmatrix}=M_{x(t),x(t-T)}e(t)\end{aligned} \tag{3.18}$$

Where

$$M_{x(t),x(t-T)}=\begin{bmatrix}-\alpha k_{x(t),x(t-T)} & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

From Eqs. (3.14 and 3.18), we get

$$\begin{aligned} &A+M_{x(t),x(t-T)}-k_p-Tk_i\\ &=\begin{bmatrix}-\alpha-k_{p1}-Tk_{i1}-\alpha k_{x(t),x(t-T)} & \alpha & 0\\ 1 & -1-k_{p2}-Tk_{i2} & 1\\ 0 & -\beta & -k_{p3}-Tk_{i3}\end{bmatrix}\end{aligned} \tag{3.19}$$

Choosing

$$P=\begin{bmatrix}p_1 & 0 & 0\\ 0 & p_2 & 0\\ 0 & 0 & p_3\end{bmatrix} \tag{3.20}$$

where p_1, p_2 and p_3 are positive constants, then

$$\begin{aligned} &(A+M_{x(t),x(t-T)}-k_p-Tk_i)^T P+P(A+M_{x(t),x(t-T)}-k_p-Tk_i)-\mu I\\ &=\begin{bmatrix}-2p_1\left(\alpha+k_{p1}+Tk_{i1}+\alpha k_{x(t),x(t-T)}+\dfrac{\mu}{2p_1}\right) & p_1\alpha+p_2 & 0\\ p_1\alpha+p_2 & -2p_2\left(1+k_{p2}+Tk_{i2}+\dfrac{\mu}{2p_2}\right) & p_2-p_3\beta\\ 0 & p_2-p_3\beta & -2p_3\left(k_{p3}+Tk_{i3}+\dfrac{\mu}{2p_3}\right)\end{bmatrix}\end{aligned} \tag{3.21}$$

In Eq. (3.21), we have

$$\Delta_1 = -2p_1\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1}\right) \tag{3.22}$$

$$\Delta_2 = \det\begin{bmatrix} -2p_1\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1}\right) & p_1\alpha + p_2 \\ p_1\alpha + p_2 & -2p_2\left(1 + k_{p2} + Tk_{i2} + \frac{\mu}{2p_2}\right) \end{bmatrix}$$

$$= 4p_1p_2\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1}\right)\left(1 + k_{p2} + Tk_{i2} + \frac{\mu}{2p_2}\right) - \left(p_1\alpha + p_2\right)^2 \tag{3.23}$$

$$\Delta_3 = \det\begin{bmatrix} -2p_1\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1}\right) & p_1\alpha + p_2 & 0 \\ p_1\alpha + p_2 & -2p_2\left(1 + k_{p2} + Tk_{i2} + \frac{\mu}{2p_2}\right) & p_2 - p_3\beta \\ 0 & p_2 - p_3\beta & -2p_3\left(k_{p3} + Tk_{i3} + \frac{\mu}{2p_3}\right) \end{bmatrix}$$

$$= 2p_1(p_2 - p_3\beta)^2\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1}\right)$$
$$-2p_3\left(k_{p3} + Tk_{i3} + \frac{\mu}{2p_3}\right)\left(4p_1p_2\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1}\right)\right.$$
$$\left.\left(1 + k_{p2} + Tk_{i2} + \frac{\mu}{2p_2}\right) - \left(p_1\alpha + p_2\right)^2\right) \tag{3.24}$$

Assuming that $\Delta_1 < 0, \Delta_2 > 0, \Delta_3 < 0$ and $p_2 = p_3\beta$ for simplicity we obtain

$$
\begin{aligned}
k_{p1} + Tk_{i1} &> \left(-\alpha - \alpha k_{x(t),x(t-T)} - \frac{\mu}{2p_1} \right) \\
k_{p2} + Tk_{i2} &> \frac{\left(p_1\alpha + p_2 \right)^2}{4p_1p_2\left(\alpha + k_{p1} + Tk_{i1} + \alpha k_{x(t),x(t-T)} + \frac{\mu}{2p_1} \right)} - 1 - \frac{\mu}{2p_2} \\
k_{p3} + Tk_{i3} &> -\frac{\mu}{2p_3}
\end{aligned}
\tag{3.25}
$$

Since $\alpha > 0, \beta > 0, a \le k_{x(t),x(t-T)} \le b < 0$, we know that, if we choose suitable $k_{p1}, k_{p2}, k_{p3}, k_{i1}, k_{i2}, k_{i3}$ such that

$$
\begin{aligned}
k_{p1} + Tk_{i1} &> \left(-\alpha(a+1) - \frac{\mu}{2p_1} \right) \\
k_{p2} + Tk_{i2} &> \frac{\left(p_1\alpha + p_2 \right)^2}{4p_1p_2\left(\alpha + k_{p1} + Tk_{i1} + a\alpha + \frac{\mu}{2p_1} \right)} - 1 - \frac{\mu}{2p_2} \\
k_{p3} + Tk_{i3} &> -\frac{\mu}{2p_3}
\end{aligned}
\tag{3.26}
$$

then (3.11) will be satisfied

By selecting $\mu = -0.2$, $P = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 14.97 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $T = 0.1$ we can get $k_{p1} = 2$, $k_{p2} = 6$, $k_{p3} = 0$, $k_{i1} = 20$, $k_{i2} = 40$, $k_{i3} = 0$ to satisfy (3.26). Figure 3.2 show the effectiveness of the proposed controller which is activated from $t = 20$ s.

3.5 Conclusions

In this paper, proportional plus integral time-delayed feedback scheme for chaos control based on Lyapunov stabilization theory is proposed. In particular, we can find many unstable periodic orbits and stabilized them through PI time-delayed feedback control. Theoretical analysis and numerical simulation for classic Chua's circuit show the effectiveness of this technique.

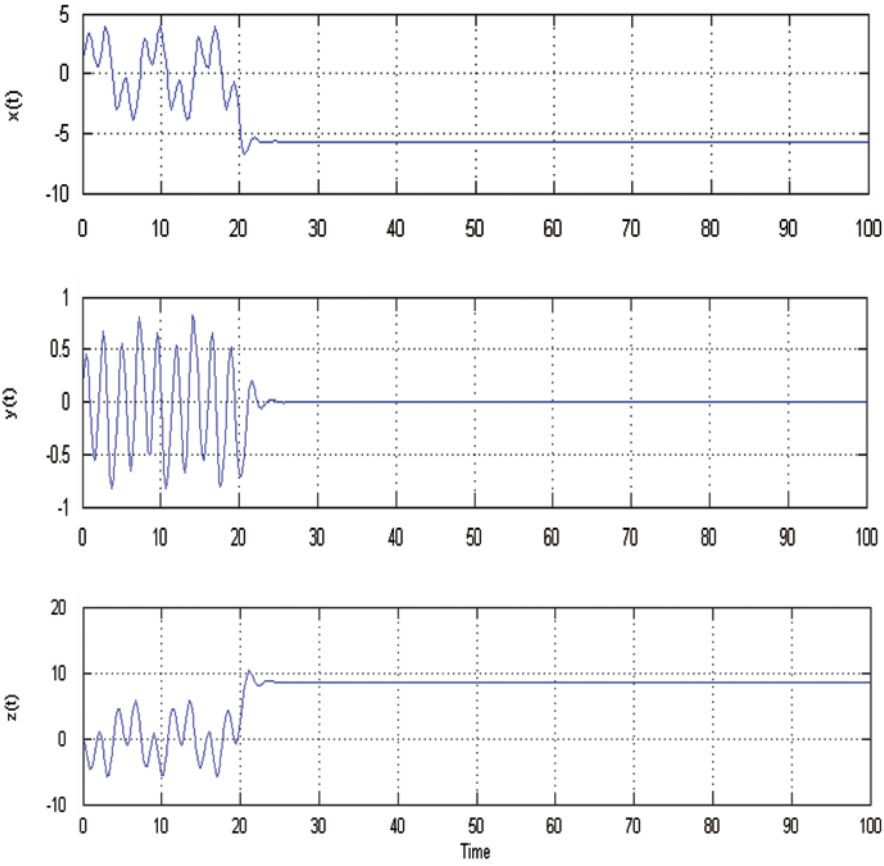


Fig. 3.2 The controlled states of Chua's circuit based on PI time-delayed controller with $k_{p1} = 2$, $k_{p2} = 6$, $k_{p3} = 0$ and $k_{i1} = 20$, $k_{i2} = 40$, $k_{i3} = 0$ which are activated from $t = 20$ s

References

1. K. Pyragas, Physics Letters A (1992), Continuous control of chaos by self-controlling feedback, 170, 421–428.
2. Maoyin Chen, D. Zhou, Y. Shang, A simple time-delayed method to control chaotic systems (2004), Chaos Solitons Fractals 22, 1117–1125.
3. Maoyin Chen, Y. Shang, D. Zhou, Repetitive learning control of continuous chaotic systems (2004b), Chaos Solitons Fractals, 22, 161–169.
4. Maoyin Chen, D. Zhou, Y. Shang, Integrity control of chaotic systems (2006), Physics Letters A, 350, 214–220.
5. Gunyaz Ablay, Nonlinear Analysis: Hybrid systems (2009), sliding mode control of uncertain unified chaotic systems, 3, 513–535.

6. Congxu Zhu, Nonlinear Analysis (2009), feedback control methods for stabilizing unstable equilibrium points in a new chaotic system, 71, 2441–2446.
7. Congxu Zhu, Z. Chen, Physics Letters A (2008), Feedback control strategies for the Liu chaotic system, 372, 4033–4036.
8. Chaohai Tao, C. Yang, Y. Luo, H. Xiong, F. Hu (2005). Speed feedback control of chaotic system. Chaos Solitons Fractals, 23, 259–263.
9. X. Liao, P. Yu, Chaos control for the family of Rossler systems using feedback controllers Chaos Solitons Fractals 29 (2006) 91–107.
10. X. Wang, L. Tian, S. Jiang, L. Yu, Feedback Control and Synchronization of Chaos for the Coupled Dynamos Dynamical System, Journal of Information and Computing Science, 1 (2006) 2, pp 93–100 ISSN 1746-7659, England, UK.
11. L. Fronzoni, M. Giocondo, CONTROLLING CHAOS WITH PARAMETRIC PERTURBATIONS, International Journal of Bifurcation and Chaos, Vol.8, No.8 (1998) 1693–1698.
12. Lu Jun-an, H. Baoxing, Wu Xiaoqun, Control of a unified chaotic system with delayed continuous periodic switch, Chaos Solitons Fractals 22 (2004) 229–236
13. Xiao F. Wang and G. Chen, Generating topologically conjugate chaotic systems via feedback control (2003), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: FUNDAMENTAL THEORY AND APPLICATIONS, VOL.50, NO. 6, 812–817.
14. Kit-Sang Tang, Kim F. Man, G. Zhong, and G. Chen, Making a continuous-time minimum-phase system chaotic by using time-delay feedback (2001), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: FUNDAMENTAL THEORY AND APPLICATIONS, VOL.48, NO.5,641–645.
15. G. JIANG and W. K. S. TANG, A global synchronization criterion for coupled chaotic systems via unidirectional linear error feedback approach, International Journal of Bifurcation and Chaos, Vol.12, No.10 (2002) 2239–2253.
16. G. JIANG, W. X. Zheng and G. Chen, Global chaos synchronization with channel time-delay, Chaos Solitons Fractals 20 (2004) 267–275.
17. J. M. Peña, Characterizations and stable tests for the Routh conditions and for total positivity Linear Algebra and its Applications 393 (2004) 319–332.
18. R. FARID, A. IBRAHIM, B. ABO-ZALAM, Synchronization of chaotic systems based on observer design under noisy environment, Proceedings of the 10th WSEAS International Conference on Automation & Information
19. R. FARID, A. IBRAHIM, B. ABO-ZALAM, Chaos Synchronization based on PI Fuzzy Observer, Proceedings of the 10th WSEAS International Conference on Fuzzy systems

Chapter 4
Analytical Results for a Small Multiple-Layer Parking System

S. R. Fleurke and A. C. D. van Enter

Abstract In this article a multilayer parking system of size $n=3$ is studied. We prove that the asymptotic limit of the particle density in the center approaches a maximum of 1/2 in higher layers. This means a significant increase of capacity compared to the first layer where this value is 1/3. This is remarkable because the process is solely driven by randomness. We conjecture that this result applies to all finite parking systems with $n \geq 2$.

Keywords Car parking problem · Multi-layer car parking · Particle deposition · Random sequential adsorption

4.1 Introduction

Suppose we have a lattice L(x, r) consisting of sites (x, r) with positions $x \in \{-2, -1, 0, 1, 2\}$ and heights $r \in N$. At each position particles arrive according to independent Poisson processes $N_t(x)$. We impose boundary conditions $N_t(-2)=N_t(2)=0$. The particles pile up across the layers but they are not allowed to "interfere" with particles earlier deposited in neighboring sites at the same layer. In other words, the horizontal distance between two particles has to be at least 2. Furthermore, the model has no screening i.e. the particles are always deposited in the lowest possible layer (see Fig. 4.1).

Our model can be formulated more precisely in the following way.

1. The state-space is $F := (L, N^+)^{\{0,1\}}$
2. The process $\kappa_t(x, r)=1$ if there is a particle at (x, r) at time t and 0 otherwise.
3. When a particle arrives at site x at time t, it will be deposited at $h_t(x) := \min\{r: \kappa_t(y, r)=0, \forall_{y \in Nx}\}$, where neighborhood set N_x consists of site x and the sites with distance 1 from it.

S. R. Fleurke (✉)
Radiocommunications Agency Netherlands, Postbus 450, 9700 AL Groningen, The Netherlands
e-mail: sjoert.fleurke@agentschaptelecom.nl

A. C. D. van Enter
Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands
e-mail: a.c.d.van.enter@rug.nl

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_4,

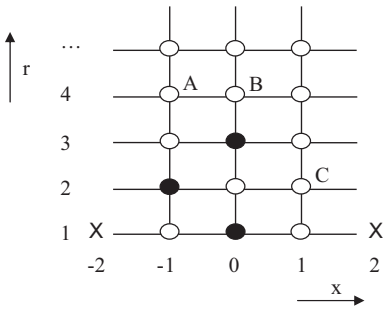


Fig. 4.1 Parking lattice consisting of three positions where parking is allowed. Three particles have already arrived consecutively at positions 0, −1, and 0. The next particle will be deposited either in A, B, or C depending on the position (−1, 0, or 1 respectively) where it arrives. The '×' symbols at −2 and 2 indicate that at those x-positions no particles arrive

The density $\rho_t(x, r)$ of a site at $(x, r) \in L$ is defined as the expectation of the occupancy of that site at time t, or $\rho_t(x, r) := E\,\kappa_t(x, r)$. The end-density of a site is $\rho_\infty(x, r)$.

Our models can be viewed either as particle deposition, car parking [1, 2], or as models for random sequential adsorption [3]. In this article we will use the terminology of particle deposition. We focus on the densities of the sites in the center, i.e. those with coordinates $(0, r)$, $r \in N^+$. The majority of the existing literature in which discrete parking is analytically treated is about monolayer models [2, 4, 5], while most literature about multi-layer models is based on simulations [6, 7]. However, in [8] it was shown that in an infinite parking system the second layer has a higher capacity than the first layer and in [9] time-dependent density formulas for the first few layers of small finite parking systems are calculated.

In this paper we continue the work on calculating the particle densities in a small multi-layer parking model. We hope our result will lead to further insights also in systems with bigger sizes.

4.2 Particle Densities in the Case of Deposition at Three Vertices on an Interval

In this section we will analytically calculate the end-densities in the case of a system with three vertices.

Theorem 4.1: Consider a multilayer parking system with three vertices. The average density at vertex 0 at height $h+1 \geq 1$ and at time t obeys the following formula

$$\rho_t^{(h+1)}(0) = \sum_{k=0}^{h}\left[\binom{h}{k}\binom{h+k}{k}\left(\frac{1}{3}\right)^{h+k+1} + 2\binom{h}{k}\sum_{j=0}^{k-1}\binom{h+1}{j}\left(\frac{1}{3}\right)^{h+j+1}\right]$$
$$-\sum_{k=0}^{h}\left[\binom{h}{k}\binom{h+k}{k}\left(\frac{1}{3}\right)^{h+k+1}\sum_{i=0}^{h+k}\frac{(3t)^i}{i!} + 2\binom{h}{k}\sum_{j=0}^{k-1}\binom{h+1}{j}\left(\frac{1}{3}\right)^{h+j+1}\sum_{i=0}^{h+j}\frac{(3t)^i}{i!}\right]e^{-3t} \tag{4.1}$$

4.2.1 Proof of Theorem 4.1

The proof of this result is based on the fact that a new particle that arrives at x at time t will always be deposited in layer $h_t(x)+1$. Therefore the derivative of the density at a height $y+1$ at time t is equal to the probability that $H_t(x)=y$.

For the height stochastic variable $H_t(0)$ we can state that

Lemma 4.1:

$$H_t(0) = N_t(0) + \max(N_t(-1), N_t(1)) \tag{4.2}$$

where $N_t(x)$ is the number of Poisson arrivals at site x at time t.

Proof:

Recall from the Introduction that the height $H_t(0)$ at position 0 is defined as the total number of layers containing one or two particles. So, we may write

$$\begin{aligned} H_t(0) &= \sum_{r=1}^{\infty} k_t(-1,r) + k_t(0,r) + k_t(1,r) - k_t(-1,r)k_t(1,r) \\ &= N_t(0) + N_t(-1) + N_t(1) - \sum_{r=1}^{\infty} k_t(-1,r)k_t(1,r) \end{aligned} \tag{4.3}$$

The value of the last term may be written as

$$\sum_{r=1}^{\infty} \kappa_t(-1,r)\kappa_t(1,r) = \begin{cases} N_t(-1) & \text{if } N_t(-1) \le N_t(1) \\ N_t(1) & \text{if } N_t(-1) > N_t(1) \end{cases} \tag{4.4}$$

Or more simply

$$\sum_{r=1}^{\infty} k_t(-1,r)k_t(1,r) = \min\left(N_t(-1), N_t(1)\right) \tag{4.5}$$

Combining this result with Eq. 4.3 completes the proof of the lemma.

The next step is to calculate the probability $\Pr(H_t(0)=h)$. Therefore we first need to derive the density of the term $\max(N_t(-1), N_t(1))$.

Lemma 4.2:

$$\begin{aligned} &\Pr(\max(N_t(-1), N_t(1)) = n) = \left(e^{-t}\frac{t^n}{n!}\right)^2 \\ &+2e^{-t}\frac{t^n}{n!}e^{-t}\sum_{j=0}^{n-1}\frac{t^j}{j!} \end{aligned} \tag{4.6}$$

Proof:

$$\begin{aligned}
&\Pr(\max(N_t(-1), N_t(1)) = n) = \Pr([N_t(-1) = n] \cap [N_t(1) < n]) \\
&+ \Pr([N_t(1) = n] \cap [N_t(-1) < n]) + \Pr(N_t(-1) = N_t(1) = n) \\
&= 2\Pr([N_t(-1) = n] \cap [N_t(1) < n]) + \Pr(N_t(-1) = N_t(1) = n) \\
&= 2\Pr([N_t(-1) = n])\Pr(N_t(1) < n + \Pr(N_t(1) = n)^2 \\
&= 2e^{-t}\frac{t^n}{n!}e^{-t}\sum_{j=0}^{n-1}\frac{t^j}{j!} + \left(e^{-t}\frac{t^n}{n!}\right)^2 \\
&= 2e^{-t}\frac{t^n}{n!}e^{-t}\sum_{j=0}^{n}\frac{t^j}{j!} - e^{-t}\frac{t^n}{n!}
\end{aligned} \tag{4.7}$$

The combination of lemma 4.1 and lemma 4.2 provides us a useful expression for the height. Since the probability that $H_t(x)=y$ equals the derivative of the density of the site at height $y+1$ at time t we can continue as follows.

Proof:

$$\begin{aligned}
&\dot{\rho}_t^{(h+1)}(0) = \Pr(H_t(0) = h) \\
&= \Pr(N_t(0) + \max(N_t(-1), N_t(1)) = h) \\
&= \sum\nolimits_{k=0}^{h} \Pr(N_t(0) = h-k)\Pr(\max(N_t(-1), N_t(1)) = k) \\
&= \sum_{k=0}^{h}\frac{t^{h-k}}{(h-k)!}\left(2e^{-t}\frac{t^k}{k!}\sum\nolimits_{j=0}^{k-1}e^{-t}\frac{t^j}{j!} + \left(e^{-t}\frac{t^k}{k!}\right)^2\right) \\
&= 2t^h e^{-3t}\sum\nolimits_{k=0}^{h}\left[\frac{1}{k!(h-k)!}\sum\nolimits_{j=0}^{k-1}\frac{t^j}{j!}\right] + t^h e^{-3t}\sum\nolimits_{k=0}^{h}\frac{1}{k!(h-k)!}\frac{t^k}{k!}
\end{aligned} \tag{4.8}$$

Integrating this expression results in the time-dependent densities ρ_t^{h+1} for layer $h+1$. So, we have

$$\begin{aligned}
\rho_t^{(h+1)}(0) = 2\sum_{k=0}^{h}&\left[\frac{1}{k!(h-k)!}\sum_{j=0}^{k-1}\frac{\int_0^t x^{h+j}e^{-3x}dx}{j!}\right] \\
&+ \sum_{k=0}^{h}\frac{\int_0^t x^{h+k}e^{-3x}dx}{(h-k)!k!^2}
\end{aligned} \tag{4.9}$$

Now we use the identity

$$\int e^{-ax}x^s dx = -\frac{s!}{a^{s+1}}e^{-ax}\sum_{i=0}^{s}\frac{(ax)^i}{i!} \tag{4.10}$$

and get

$$\begin{aligned}\rho_t^{(h+1)}(0) &= 2\sum_{k=0}^{h}\left[\frac{1}{k!(h-k)!}\sum_{j=0}^{k-1}\frac{(h+j)!}{j!}\frac{1-e^{-3t}\sum_{i=0}^{h+j}\frac{(3t)^i}{i!}}{3^{h+j+1}}\right]\\ &\quad+\sum_{k=0}^{h}\frac{(h+k)!}{k!^2(h-k)!}\sum_{j=0}^{k-1}\frac{\left(1-e^{-3t}\sum_{i=0}^{h+k}\frac{(3t)^i}{i!}\right)}{3^{h+k+1}}\\ &= 2\sum_{k=0}^{h}\binom{h}{k}\sum_{j=0}^{k-1}\binom{h+j}{j}\frac{\left(1-e^{-3t}\sum_{i=0}^{h+j}\frac{(3t)^i}{i!}\right)}{3^{h+j+1}}\\ &\quad+\sum_{k=0}^{h}\binom{h}{k}\binom{h+k}{k}\frac{\left(1-e^{-3t}\sum_{i=0}^{h+k}\frac{(3t)^i}{i!}\right)}{3^{h+k+1}}\\ &= \sum_{k=0}^{h}\left[\binom{h}{k}\binom{h+k}{k}\frac{1}{3^{h+k+1}}+2\binom{h}{k}\sum_{j=0}^{k-1}\binom{h+j}{j}\frac{1}{3^{h+j+1}}\right]\\ &\quad-\sum_{k=0}^{h}\left[\binom{h}{k}\binom{h+k}{k}\frac{\sum_{i=0}^{h+k}\frac{(3t)^i}{i!}}{3^{h+k+1}}+2\binom{h}{k}\sum_{j=0}^{k-1}\binom{h+j}{j}\frac{\sum_{i=0}^{h+j}\frac{(3t)^i}{i!}}{3^{h+j+1}}\right]e^{-3t}\end{aligned} \tag{4.11}$$

This may be rewritten as (with r = h + 1).

$$\begin{aligned}\rho_t^{(r)}(0) &= \sum_{k=0}^{r-1}\left[\frac{\binom{r-1}{k}\binom{r+k-1}{k}}{3^{r+k}}+2\binom{r-1}{k}\sum_{j=0}^{k-1}\binom{r+j-1}{j}\frac{1}{3^{r+j}}\right]\\ &\quad-\sum_{k=0}^{r-1}\left[\binom{r-1}{k}\binom{r+k-1}{k}\frac{\sum_{i=0}^{r+k-1}\frac{(3t)^i}{i!}}{3^{r+k}}\right.\\ &\quad\left.+2\binom{r-1}{k}\sum_{j=0}^{k-1}\binom{r+j+1}{j}\frac{\sum_{i=0}^{r+j-1}\frac{(3t)^i}{i!}}{3^{r+j}}\right]e^{-3t}\end{aligned} \tag{4.12}$$

For the first few layers Theorem 4.1 provides:

$$\begin{aligned}
\rho_t^{(1)} &= \frac{1}{3} - \frac{1}{3}e^{-3t} \\
\rho_t^{(2)} &= \frac{11}{27} - \left(\frac{11}{27} + \frac{11}{9}t + \frac{1}{3}t^2\right)e^{-3t} \\
\rho_t^{(3)} &= \frac{35}{81} - \left(\frac{35}{81} + \frac{35}{27}t + \frac{35}{18}t^2 + \frac{7}{9}t^3 + \frac{1}{12}t^4\right)e^{-3t} \\
\rho_t^{(4)} &= \frac{971}{2187} - \left(\frac{971}{2187} + \frac{971}{729}t + \frac{971}{486}t^2 \frac{971}{486}t^3 + \frac{283}{324}t^4 + \frac{17}{108}t^5\right)e^{-3t}
\end{aligned} \tag{4.13}$$

A plot of these functions is shown in Fig. 4.2. Confer [9] where the first three layers were calculated using a different approach.

4.3 Calculation of the End-Densities

Close inspection of Eq. 4.13 reveals that as time goes to infinity the densities of the first four layers tend towards 1/3, 11/27, 35/81, and 971/2187 respectively. Calculating end-densities for higher layers can be done directly from Theorem 4.1.

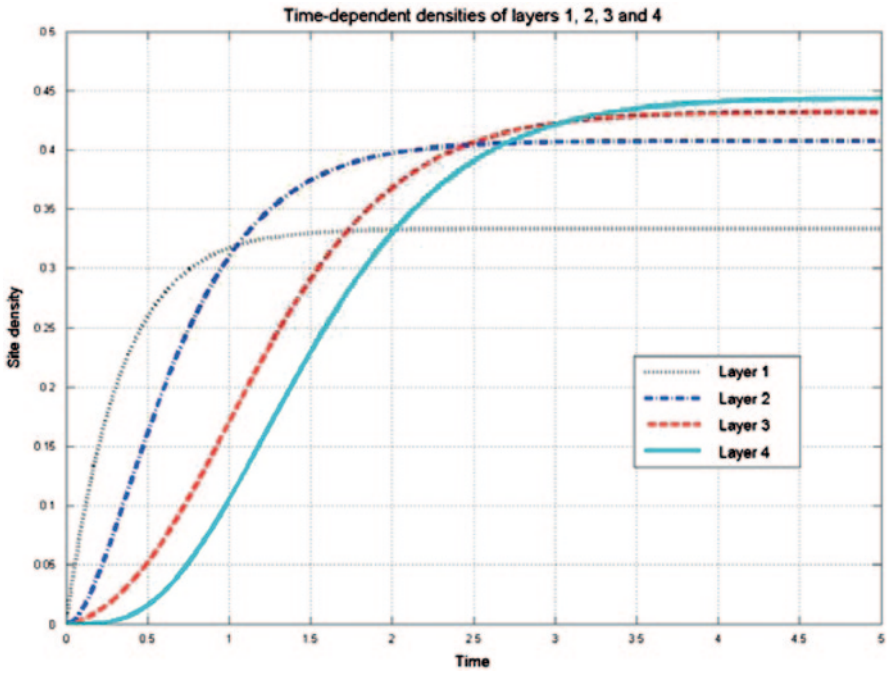


Fig. 4.2 Particle densities at the sites (0, r) as a function of time in the cases r is 1, 2, 3, and 4 according to Eq. 4.13

Table 4.1 End-densities calculated using Theorem 4.1

Layer	End-density	Approximately
1	1/3	0.3333
2	11/27	0.4074
3	35/81	0.4321
4	971/2187	0.4440
5	8881/19683	0.4512
6	80811/177147	0.4562
7	733209/1594323	0.4599
8	6640491/14348907	0.4628
9	60067809/129140163	0.4651
10	542880971/1162261467	0.4671

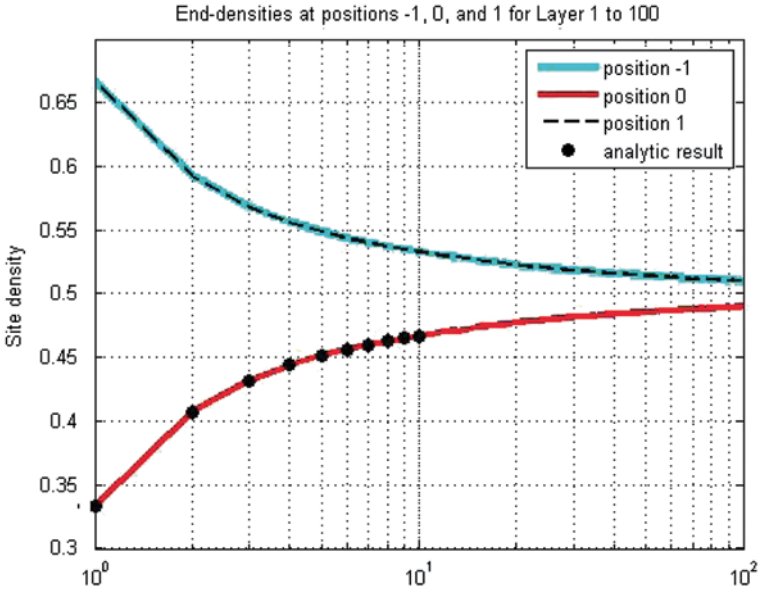


Fig. 4.3 End-densities as a function of the layer created by 10^7 simulations. The analytic results from Table 4.1 are plotted as well to demonstrate its concurrence

See Table 4.1 for the exact values of the end-densities of the first 10 layers and its decimal approximations.

A plot of these constants for the first 100 layers is shown in Fig. 4.3. It can be seen clearly that the graph of these end-densities appears to approach the value of 1/2. In this section we will prove that this is indeed the case.

Define $\rho^{(r)} := \lim_{t\to\infty} \rho_t(0, r)$. Then we have the following result.

Theorem 4.2: The density at high layers converges in time to the value

$$\lim_{r\to\infty} \rho_t^{(r)} = \frac{1}{2} \tag{4.14}$$

To prove this we can take the result of Theorem 4.1 and focus on the constant term.

$$\rho_t^{(h+1)} = \sum_{k=0}^{h}\left[\binom{h}{k}\binom{h+k}{k}\left(\frac{1}{3}\right)^{h+k+1} + 2\binom{h}{k}\sum_{j=0}^{k-1}\binom{h+j}{j}\left(\frac{1}{3}\right)^{h+j+1}\right] \tag{4.15}$$

We may rewrite this more conveniently as

$$\begin{aligned}\rho_t^{(r)} = &\frac{1}{2}\sum_{k=0}^{r-1}\Pr\left(X_{r-1,\frac{1}{2}} = k\right)\Pr\left(Y_{r,\frac{1}{3}} = k\right)\\ &+\sum_{k=0}^{r-1}\Pr\left(X_{r-1,\frac{1}{2}} = k\right)\sum_{j=0}^{r-k}\Pr\left(Y_{r,\frac{1}{3}} = j\right)\end{aligned} \tag{4.16}$$

Where we used the notation $r=h+1 \in N^+$, $X_{n,\,p} \sim B(n,\,p)$ or $\Pr(X_n, p=k) = \binom{n}{k}(1-p)^{n-k}\,p^k$, and also $Y_{r,p} \sim NB(r,\,p)$ or $\Pr(Y_{r,p}=k) = \binom{r+k-1}{k}(1-p)^r\,p^k$. We will treat the first and second term of Eq. 4.16 separately in the following lemmas.

The first term of Eq. 4.16 converges to zero when $r \to \infty$, or

$$\lim_{r\to\infty}\frac{1}{2}\sum_{k=0}^{r-1}\Pr\left(X_{r-1,\frac{1}{2}} = k\right)\Pr\left(Y_{r,\frac{1}{3}} = k\right) = 0 \tag{4.17}$$

Proof:

$$\begin{aligned}&\sum_{k=0}^{r-1}\Pr\left(X_{r-1,\frac{1}{2}} = k\right)\Pr\left(Y_{r,\frac{1}{3}} = k\right)\\ &= \sum_{k=0}^{r-1}\Pr\left(X_{r-1,\frac{1}{2}} = k \cap Y_{r,\frac{1}{3}} = k\right)\\ &= \sum_{k=0}^{r-1}\Pr\left(Y_{r,\frac{1}{3}} = X_{r-1,\frac{1}{2}} \,\middle|\, X_{r-1,\frac{1}{2}} = k\right)\Pr\left(X_{r-1,\frac{1}{2}} = k\right)\\ &= \Pr\left(Y_{r,\frac{1}{3}} = X_{r-1,\frac{1}{2}}\right)\end{aligned} \tag{4.18}$$

This represents the probability that the number of successes (with Pr(Success) = 1/3) after r failures equals the number of successes in a Binomial experiment of $r-1$ trials and Pr(Success) = ½. When we let $r \to \infty$ both $X_{r-1,\,1/2}$ and $Y_{r,\,1/3}$ will converge to continuous Gaussian distributions, so that this probability vanishes.

Lemma 4.3: The second term of Eq. 4.16 converges to ½, or

$$\lim_{r\to\infty}\sum_{k=0}^{r-1}\Pr\left(X_{r-1,\frac{1}{2}} = k\right)\sum_{j=0}^{r-k}\Pr\left(Y_{r,\frac{1}{3}} = j\right) = \frac{1}{2} \tag{4.19}$$

Proof:

$$\sum_{k=0}^{r-1} \Pr\left(X_{r-1,\frac{1}{2}} = k \right) \sum_{j=0}^{r-k} \Pr\left(Y_{r,\frac{1}{3}} = j \right)$$
$$= \sum_{k=0}^{r-1} \Pr\left(X_{r-1,\frac{1}{2}} = k \right) \Pr\left(Y_{r,\frac{1}{3}} \le r-k \right) \tag{4.20}$$

Now we will use the symmetry of the negative binomial distribution for large r. Note that $\Pr(Y_{r,p} < r-k) = \Pr(Y_{r,p} > k)$ in this case where $p = 1/3$.

$$\lim_{r\to\infty} \frac{1}{2}\left(\Sigma_{k=0}^{r-1} \Pr(X_{r-1,\frac{1}{2}} = k) \Pr\left(Y_{r,\frac{1}{3}} < r-k \right) + \Sigma_{k=0}^{r-1} \Pr\left(X_{r-1,\frac{1}{2}} = k \right) \Pr\left(Y_{r,\frac{1}{3}} < k+1 \right) \right)$$
$$= \lim_{r\to\infty} \frac{1}{2}\left(\Sigma_{k=0}^{r-1} \Pr(X_{r-1,\frac{1}{2}} = k) \Pr\left(Y_{r,\frac{1}{3}} > k \right) + \Sigma_{k=0}^{r-1} \Pr\left(X_{r-1,\frac{1}{2}} = k \right) \Pr\left(Y_{r,\frac{1}{3}} < k+1 \right) \right)$$
$$\approx \frac{1}{2} \Sigma_{k=0}^{r-1} \Pr\left(X_{r-1,\frac{1}{2}} = k \right) = \frac{1}{2} \tag{4.21}$$

4.3.1 Alternative Proof

We note however that this result also follows from the following consideration. After a while the differences in height between position −1 and 1 increase to the order of the square root of the total number of dropped particles. This follows by application of the Central Limit Theorem to $K_t := |N_t(-1) - N_t(1)|$.

The probability that a new particle drops at a side vertex happens with probability 2/3. So, the probability that this particle raises the height equals 1/2 times 2/3, which is 1/3. This equals the probability that a particle drops on the center vertex, by which the height always increases. For 1/3 of the dropped particles the height does not increase. Thus half of the newly filled layers contain an occupied center vertex, and half will contain two occupied side vertices, which implies density one half.

4.3.2 Larger Parking Systems

The calculation of (end-)densities in larger systems is much more complicated. It is always possible to calculate the densities on the first layer [4] or the first few layers [9] but going beyond the first few layers in systems with bigger sizes probably requires more advanced methods.

However, it is interesting to ask oneself whether the behavior of the small system demonstrated in this article does also appear in larger systems. Do systems of bigger

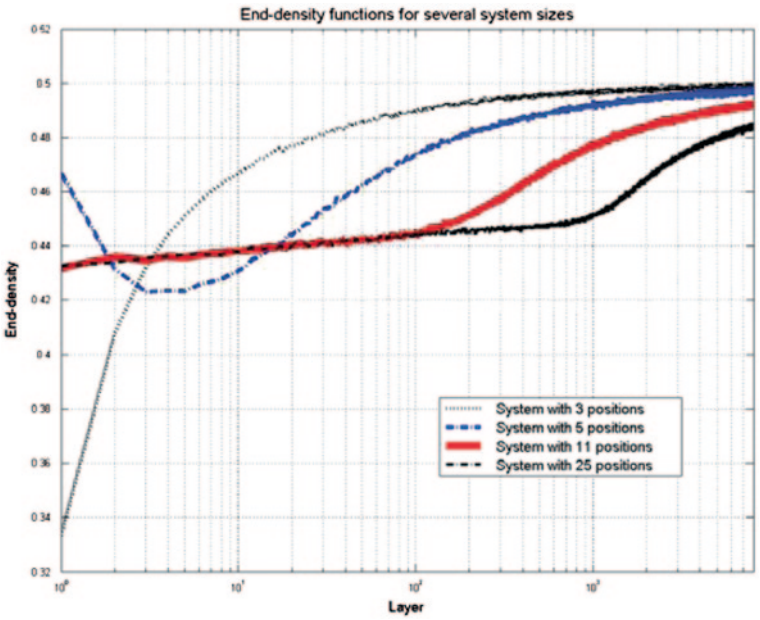


Fig. 4.4 The fact that the end-densities of the sites in the center converge to 1/2 in the case of three positions is not unique. Simulation results of bigger systems suggest that this behavior is not uncommon for finite-sized systems. However, it appears that the bigger the system, the more layers it takes to approach the limit of 1/2

sizes also generally have higher end-densities in higher layers than in lower layers, and if so, do those end-densities ultimately approach the maximum value of ½ as well?

We conjecture that this is the case for all finite-sized systems. Although we are not able to give hard evidence for this we can provide some simulation results (Fig. 4.4) supporting our view and justifying further research.

4.4 Conclusions

In this paper we introduced a parking system consisting of three positions. The formula for the time dependent densities of the center position for all layers was analytically derived.

Although similar work has been done on the model with screening (see [10, 11]) to our knowledge this is the first time that densities in a multi-layer particle deposition model without screening were calculated analytically.

We paid special attention to the densities of the center sites when $t \to \infty$, the so called end-densities. We proved that they increase as a function of the layer number and eventually approach the density 1/2.

We showed that in the case of a small system with three positions it can be easily understood why the end-density converges to this value. But this is not the case with larger systems although our simulation results do suggest similar end-density behavior. Although not yet fully understood, it thus seems that these randomly driven finite parking systems tend to use the parking space of the center positions more efficiently over time.

Acknowledgement This research was partially supported by the Radiocommunications Agency of the Netherlands.

References

1. Renyi A (1958) On a one-dimensional problem concerning random space-filling. Publ. Math. Inst. Hung. Acad. Sci. 3:109–127
2. Hemmer PC (1989) The random parking problem. J. Stat. Phys. 57:865–869
3. Evans JW (1993) Random and cooperative sequential adsorption. Rev. Mod. Phys. 64(4):1281–1327
4. Cohen R, Reiss H (1963) Kinetics of reactant isolation I. one-dimensional problems. J. Chem. Phys. 38(3):680–691
5. Dehling HG, Fleurke SR, Külske C (2008) Parking on a random tree. J. Stat. Phys. 133(1):151–157
6. Nielaba, P, Privman, V (1992) Multilayer adsorption with increasing layer coverage. Phys. Rev. A 45:6099–6102
7. Dehling HG, Fleurke SR (2007) The sequential frequency assignment process. Proc. of the 12th WSEAS Internat. Conf. on Appl. Math., Cairo, Egypt, pp 280–285
8. Fleurke SR, Külske C (2009) A second-row parking paradox. J. Stat. Phys. 136(2):285–295
9. Fleurke SR (2011) Multilayer particle deposition models (Groningen Thesis), published by VDM Verlag Dr. Muller, Saarbrücken, p 33
10. Mountford TS, Sudbury A (2012) Deposition processes with hardcore behavior. J. Stat. Phys. 146:687–700
11. Fleurke SR, Külske C (2010) Multilayer parking with screening on a random tree. J. Stat. Phys. 139(3):417–431

Chapter 5
Three Dimensional Pulsatile Non-Newtonian Flow in a Stenotic Vessel

I. Husain, C. Langdon and J. Schwark

Abstract This study investigates the pulsatile simulations of non_Newtonian flows in a stenotic vessel. Four non-Newtonian blood models, namely the Power Law, Casson, Carreau and the Generalized Power Law, as well as the Newtonian model of blood viscosity, are used to investigate the flow effects induced by these different blood constitutive equations. The aim of this study are three fold: firstly, to investigate the variation in wall shear stress in an artery with a stenosis at different flow rates and degrees of severity; secondly, to compare the various blood models and hence quantify the differences between the models and judge their significance and lastly, to determine whether the use of the Newtonian blood model is appropriate over a wide range of shear rates.

Keywords Fluid flows · Blood · Stenosis · Non-Newtonian · Pulsatile · Simulations

5.1 Introduction

This paper presents the second of a two-part study on the numerical simulations of blood flow in a representative model of an arterial stenosis in the common carotid artery for various degree of severity using five blood rheological models.

The partial obstruction of arteries due to a stenosis is one of the most frequent anormalies in blood circulation. It is well known that, once such an obstruction is formed, the blood flow is significantly altered and fluid dynamic factors such as velocity, pressure or shear stress play an important role as the stenosis continues to develop [1]. So far, the specific role of these factors is not yet well understood. The ability to accurately describe the flow through a stenosed vessel would provide the possibility of diagnosing these diseases in its earlier stages. Furthermore,

I. Husain (✉) · C. Langdon · J. Schwark
Department of Mathematics, Luther College—University of Regina,
Regina, SK S4S 0A2, Canada
e-mail: Iqbal.Husain@uregina.ca

C. Langdon
e-mail: Chris.Langdon@uregina.ca

J. Schwark
e-mail: Justin.Schwark@uregina.ca

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_5,

the presence of the anomaly itself may produce flow disturbances such as vortex formation, which has been reported as a contributing factor to atherogenesis and thrombogenesis [2].

The aim of this study are three fold: firstly, to investigate the variation in wall shear stress in an artery with a stenosis at different flow rates and degrees of severity; secondly, to compare the various blood models and hence quantify the differences between the models and judge their significance and lastly, to determine whether the use of the Newtonian blood model is appropriate over a wide range of shear rates.

5.2 Analysis and Modelling

5.2.1 Governing Equations

The blood flow is assumed to be laminar and incompressible and therefore the Navier-Stokes equations for 3D incompressible flow are given by

$$\nabla \cdot V = 0 \tag{5.1}$$

$$\rho\left(\frac{\partial V}{\partial t} + V \cdot \nabla V\right) = -\nabla \cdot \tau - \nabla p \tag{5.2}$$

where V is the 3D velocity vector, p is the pressure, ρ is th density and τ the shear stress term.

Four different non-Newtonian blood flow models as well as the simple Newtonian model are considered in this study. The effects of these models on the flow field and the wall shear stress in the vicinity of a stenosis are examined. These models are given below [3].

5.2.1.1 Blood Models

a. **Newtonian model**

$$\mu = 0.00345 \qquad Pa \cdot s \tag{5.3}$$

b. **Power Law Model**

$$\mu = \mu_0 (\dot{\gamma})^{n-1} \tag{5.4}$$

c. **Casson Model**

$$\mu = \frac{\left[\sqrt{\tau_y} + \sqrt{\eta|\dot{\gamma}|}\right]^2}{|\dot{\gamma}|} \tag{5.5}$$

d. **Carreau Model**

$$\mu = \mu_\infty + (\mu_0 - \mu_\infty)\left[1 + (\lambda\dot{\gamma})^2\right]^{(n-1)/2} \tag{5.6}$$

e. **Generalized Power Law Model**

$$\mu = \lambda|\dot{\gamma}|^{n-1} \tag{5.7}$$

where

$$\lambda(\dot{\gamma}) = \mu_\infty + \Delta\mu \exp\left[-\left(1 + \frac{|\dot{\gamma}|}{a}\right)\exp\left(\frac{-b}{|\dot{\gamma}|}\right)\right],$$

$$n(\dot{\gamma}) = n_\infty - \Delta n \exp\left[-\left(1 + \frac{|\dot{\gamma}|}{c}\right)\exp\left(\frac{-d}{|\dot{\gamma}|}\right)\right]$$

5.2.1.2 Geometry

The flow geometry comprises a tube of diameter D and can be divided into three regions, the inlet, the deformed and the outlet region. In the case of the stenosis, the lengths of these regions are $4D$, $2D$ and $20D$, respectively. The radius of the undeformed inlet and outlet is $R_0 = \frac{D}{2}$.

In the case of the stenosis, the radius of the constricted region is given by

$$R = R_0\left[1 - \left(\frac{R_0 - R_{\min}}{R_0}\right)\left(\frac{1 - \cos(\pi x / D)}{2}\right)^2\right] \quad 0 \le x \le 2D \tag{5.8}$$

where $R_{\min}$ is the minimum radius at the centre of the stenosis. In this study, three different degrees of stenosis were used, 20, 50 and 80%.

5.2.1.3 Assumptions and Boundary Conditions

It is assumed that the arterial walls are rigid and no-slip condition is imposed at the walls. At the outlet, stress-free conditions are applied and the pressure is set to zero. Finally, the velocity profile at the inlet is regarded to be that of fully developed flow in a straight tube and can be derived analytically for both the Newtonian and the Power Law fluids [4]. The forms are

$$u = \overline{u}\left[1-\left(\frac{r}{R_0}\right)^2\right] \qquad 0 \le r \le R_0 \tag{5.9}$$

where u is the velocity component in the $\overline{x}$ direction for the Newtonian flow and

$$u = \overline{u}\left(\frac{3n+1}{n+1}\right)\left[1-\left(\frac{r}{R_0}\right)^{\frac{n+1}{n}}\right] \qquad 0 \le r \le R_0 \tag{5.10}$$

for the non-Newtonian flow. In transient flow, the pulsatile flow at the inlet is given by a time varying forcing function given in Fig. 5.1 below. This forcing function was scaled to yield a maximum inflow velocity of $\overline{u}$ with a heart rate of approximately 60 beats per minute.

5.2.1.4 Solution Methodology

The governing equations are highly nonlinear and must be solved numerically using techniques of computational fluid dynamics. In this study, these equations are solved using the finite element method as implemented by COMSOL (COMSOL Inc., Los Angeles, CA). The flow geometries for the stenosis was first created using Matlab. Then a finite element mesh was placed on this geometry. Briefly, an inlet plane of the artery is meshed in 2D using triangles and this mesh is extruded along the centerline of the artery to create a 3D mesh consisting of hexadrel elements. The mesh used for all computations consisted of 9,708 elements and 15,048 nodes for the stenosis.

The governing equations were solved completely using the boundary conditions for fully developed flow (5.9) and (5.10) at the inlet along with the pulsatile forcing function for the transient case.

5.3 Results and Discussion

Transient simulations were performed using all five models given above. Three different degrees of stenosis were used namely 20, 50 and 80 %.

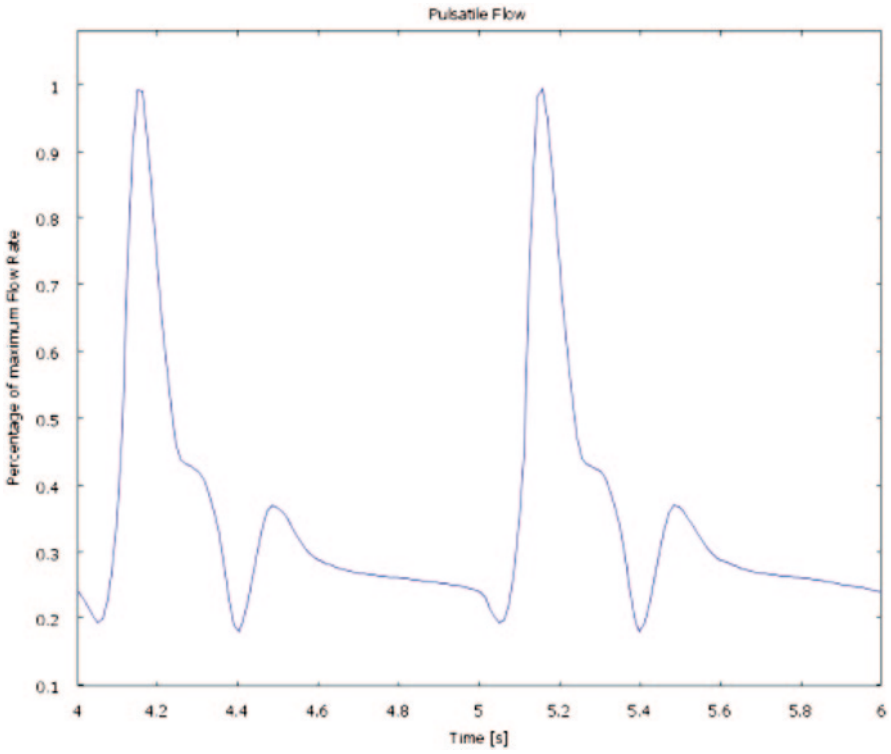


Fig. 5.1 Physiological flow waveform in the carotid artery used to drive the inlet velocity boundary conditio as a function of time

Figure 5.2 below shows that all of the non-Newtonian models considered here except the Power Law model produce a higher pressure difference than the Newtonian model. Specifically, the highest pressure drop is induced by the Generalized Power Law model and the lowest by the Power Law model. Similar pattern in pressure differences are obtained at higher flow rates.

The distribution of the wall shear stress (WSS) is one of the most important hemodynamic parameter due to its direct relevance in artherosclerosis formation. Figure 5.3 shows the distributions of maximum shear stress for various degrees of severity of the stenosis for all models. It is evident that WSS increases with increasing severity. All models show close agreement with the Newtonian model except for the Power Law model. At 50% stenosis, the WSS predicted by this model is significantly lower than the rest. Figure 5.4 shows the distribution of WSS along the geometry at various times. Maximum shear stresses are reached just before the throat of the stenosis. The magnitude of this value increases with higher flow rates. This peak is followed by a negative value indicating the presence of backflow. Further downstream, the WSS steadily regains its undisturbed value.

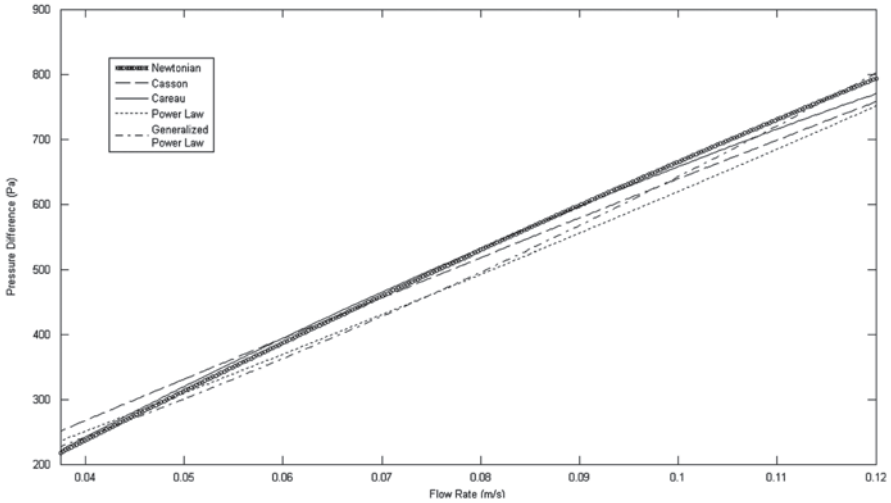


Fig. 5.2 Pressure difference versus flow rate

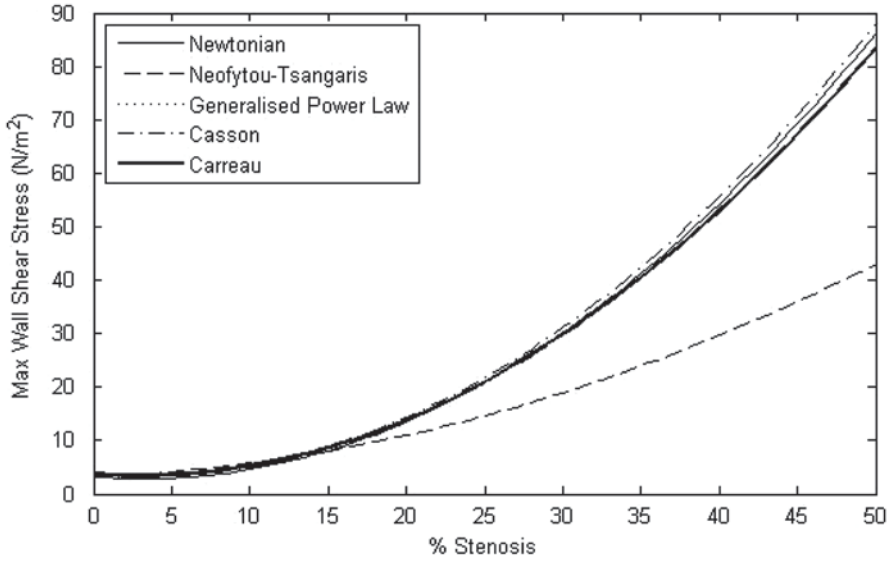


Fig. 5.3 Wall Shear stress versus percent stenosis for various models, with 0.11196 m/s inflow rate

The Power Law model gives a much lower $\tau_W^{\max}$ value because it exhibits a lower viscosity at the throat of the stenosis where the shear stress is high. As the flow rate increases, these WSS differences from various models become more prominent indicating significant differences in model behaviour.

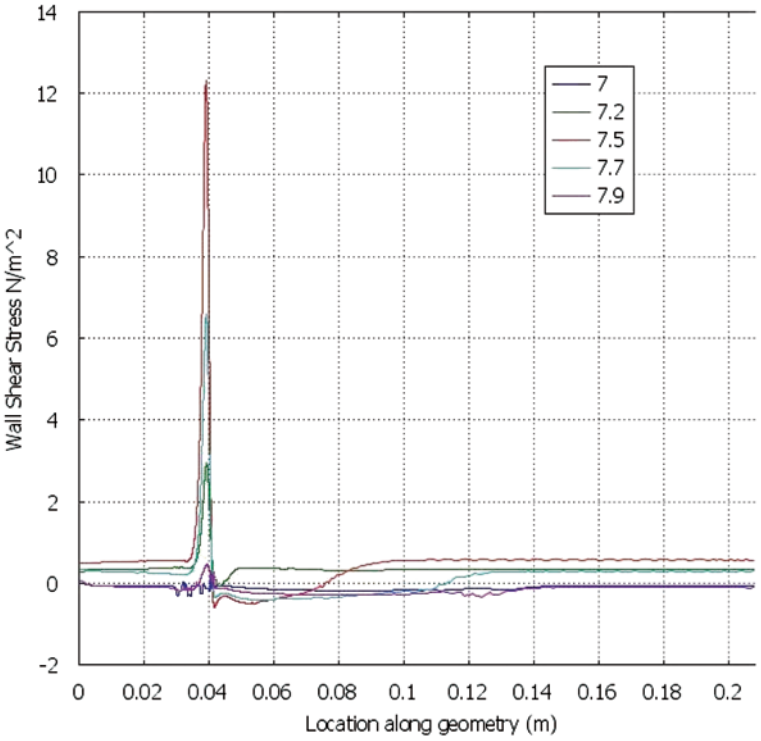


Fig. 5.4 Wall shear stress for 50% stenosis pulsatile Generalized Power Law model at various time intervals, with 0.11196 m/s max inflow rate

Transient simulations were performed using the Generalized Power Law Model for the stenosis and each simulation was from $t = 0$ to 10.0 s, yielding a heart rate of approximately 60 beats per minute.

Figure 5.5 shows the distribution of maximum WSS with shear rate in a stenosis. Again, WSS increases with increasing shear rate with the Power Law model deviating significantly from the rest.

The maximum wall shear stress occurs in the middle of the cycle corresponding to the maximum inflow velocity. The distribution of shear rates in a 50% stenosed artery is shown in Fig. 5.6. The regions of high shear are confined to the throat of the stenosis and immediately downstream of the stenosis.

The maximum and minimum WSS values are in close agreement for the Generalized Power Law, Carreau and the Newtonian models. The Power Law model gives a much lower value because it exhibits a lower viscosity at the throat of the stenosis where the shear stress is high. As the flow rate increases, these WSS differences from the first three models become less prominent indicating insignificant differences in model behavior at high shear rates.

Similar result are obtained when the diameter of the common carotid artery is assumed to be as large as 0.8 cm. The maximum wall shear stress and shear rates

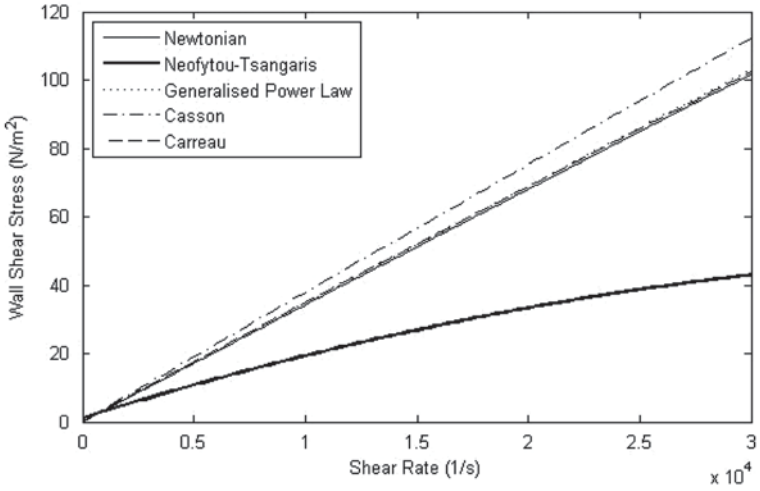


Fig. 5.5 WSS versus shear rate in a stenosis

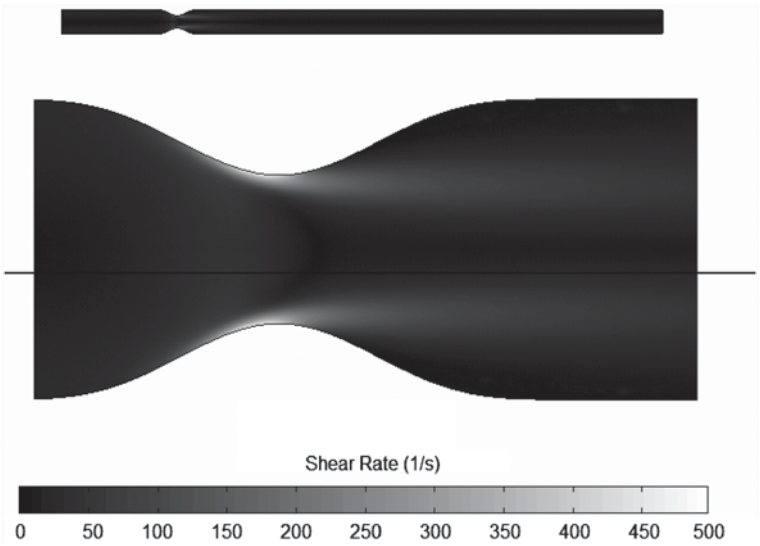


Fig. 5.6 Shear rate distribution in a 50% stenosed vessel

values are lower when compared to the 0.64 diameter artery but the differences in model behaviour are analogous.

It is evident from these results that the Power Law model tends to break down at higher shear rates in that it reduces the viscosity of the blood to levels below the Newtonian level which theoretically should not be possible. This is noticeable in the 80% stenosed model. The pressure difference predicted by this model is less than

the Newtonian model, indicating a lower than Newtonian viscosity. This method also show very low wall shear stress levels, dropping below Newtonian levels at fairly low shear rates, for example at the medium flow rate at 50% stenosed, the WSS levels are less than Newtonian levels. This model is relatively easy to use but predict decreasing viscosity at higher strain, contrary to the generally accepted observation that blood behaves as a Newtonian fluid for strains above $100s^{-1}$.

At low shear rates the Casson model shows near Newtonian behavior with the behavior becoming less Newtonian as the shear rate increases for a period, then reverting to near Newtonian behavior as the rate continues to increase. This model takes the haematocrit factor H (the volume fraction of red blood cells in whole blood) into account, with the parameters given (obtained from data fitting) suggesting a value of H of 37%. However, it is reported that this yields a limiting viscosity at high shear slightly above the usual Newtonian value. The results obtained here suggest the same with WSS values above the Newtonian values at low and very high shear rates. This model appears to be fairly accurate at high shear rates, but not at low shear rates.

The Carreau model produces values that are in close agreement with that of the Newtonian model at shear rates well above $100s^{-1}$. Our results indicate this to be the case. Both the WSS and the pressure difference tends to the Newtonian values at shear rates in excess of $1{,}000s^{-1}$. This model by design reverts to Newtonian numbers as shear rates approach infinity. The basis for this model is the constant Newtonian viscosity, modified to non-Newtonian such that the modification tends to zero as the limit of the shear rate goes to infinity.

Finally, the Generalized Power Law model gave results that are in closest agreement with the Newtonian values at mid-range and high shear rates. At low shear rates, this model gives values that are close to that of the Power Law and the Carreau models. While the Power Law model breaks down at high shear, our results show a close agreement between the Generalized Power Law and the Carreau models even at high shear rates as shown in Fig. 5.5. The Generalized Power Law model is widely accepted as a general model for non-Newtonian blood viscosity. It includes the Power Law model at low shear rate and the Newtonian model at mid-range and high shear rates. There is also good agreement between the Generalized Power Law and the Carreau model for low shear rates.

5.4 Conclusions

A study of the effects of modeling blood flow through a stenosis using five different blood rheological models is presented. The flow field and wall shear stress distributions produced by each model are investigated for various flow rates and degrees of abnormality. The results show that there are significant differences between simulating blood as a Newtonian or non- Newtonian fluid. It is found that the Newtonian model is a good approximation in regions of mid-range to high shear but the Generalized Power Law model provides a better approximation of wall shear stress at low shear.

These conclusions are presented under the assumption that the arterial walls are rigid and zero pressure is assumed at the outlet. A more realistic simulation would include elastic walls and incorporate the effects of upstream and downstream parts of the circulatory system into the boundary conditions. This is a long term objective of this study.

References

1. Berger SA, Jou, L-D (2000) Flows in stenotic vessels. Annual Reviews of Fluid Mechanics, 32:347–382.
2. Ku DN (1997) Blood flow in arteries. Annual Review of Fluid Mechanics, 29:399–434.
3. Johnsto BM et al. (2004) Non-Newtonian blood flow in human right coronary arteries: steady state simulations. J. Biomechanics, 37:709–720.
4. Neofytou P, Tsangaris S (2005) Flow effects of blood constitutive equations in 3D models of vascular anomalies. Int. J. Numer. Meth. Fluids, 51:489–510.

Chapter 6
A Polynomial Matrix Approach to the Descriptor Systems

W. Kase

Abstract In this chapter, we will propose an analysis method of the descriptor systems using the regularizing polynomial matrix. The regularizing matrix compensates the singularity of the descriptor systems, like an interactor matrix. We will show that the degree of the regularizing polynomial matrix presents a structure aspect of a given descriptor system.

Keywords Linear multivariable systems · Descriptor systems · Polynomial matrix · Regularizing matrix

6.1 Introduction

The descriptor systems [1] are convenient and natural modeling process for the practical plants. The state space method [2] and the geometric approach [3] are used to study the structure properties and to design the controllers. Comparing these methods, there are not so many literatures using the polynomial matrix approach [4]. Since the impulsive modes in the descriptor systems cause the improper transfer function, it is natural to to discuss the treatment of the improper transfer function using the polynomial matrices.

In this chapter, we will propose an analysis method of the descriptor systems using the regularizing polynomial matrix. The regularizing polynomial matrix compensates the singularity of the descriptor systems, like an interactor matrix for rational function matrices [5]. In fact, the regularizing matrix is almost equivalent to an interactor. Although some derivation methods of the interactor were proposed, almost of all were complex. Mutoh and Ortege proposed the algebraic equation, which the coefficient matrices of the interactor should be satisfied [6]. But the solution method in [6] was not adequate for computer calculations. The authors proposed a solution of the equation in [6] using Moore-Penrose pseudo-inverse [7]. Since a function to calculate the pseudo-inverse is available in some standard softwares for control engineering, the method is adequate for computer calculations.

W. Kase (✉)
Department of Electrical and Electronic Systems Engineering,
Osaka Institute of Technology, 5-16-1 Omiya, Asahi-ku, Osaka 535-8585, Japan
e-mail: kase@ee.oit.ac.jp

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_6,

We will show that the degree of the regularizing polynomial matrix presents a structural aspect of a given descriptor system. That is, there exists the regularizing matrix of degree one if a given system has no impulsive mode. There exists the regularizing matrix of degree two if a given system has some impulsive modes. We will also discuss a condition for the impulsive controllability of the descriptor systems using the analysis. We will also discuss the feedback controller design which removes the impulsive modes of the descriptor systems.

6.2 Regularizing Polynomial Matrix

Consider the following $q \times m$ $(q \le m)$ polynomial matrix $D(s)$:

$$\begin{aligned} D(s) &= D_0 + sD_1 + \cdots + s^{\mu} D_{\mu} \\ &= \boldsymbol{D} S_{I_m}^{\mu}(s) \end{aligned} \tag{6.1}$$

where

$$\begin{aligned} \boldsymbol{D} &= \begin{bmatrix} D_0 & D_1 & \cdots & D_{\mu} \end{bmatrix} \\ S_{I_m}^{\mu}(s) &= \begin{bmatrix} 1 & s & \cdots & s^{\mu} \end{bmatrix}^T \end{aligned} \tag{6.2}$$

$D(s)$ is called *regular* if D_{μ} has full rank q. The problem considered in this Sect is to find a $q \times q$ nonsingular polynomial matrix $L(s)$ which makes μ-th degree's coefficient matrix of $L(s)D(s)$ be full rank and the coefficient matrices which degrees are greater than μ be zeros. $L(s)$ is called a regularizing polynomial matrix of $D(s)$. The existence of such matrix is clear by considering the interactor for $D(s)/s^{\mu+1}$. In the following, we will consider the direct derivation of $L(s)$ not using the interactor.

Assume that $L(s)$ has the following structure

$$\begin{aligned} L(s) &= L_0 + sL_1 + \cdots s^w L_w \\ &= \boldsymbol{L} S_{I_q}^{w}(s) \\ \boldsymbol{L} &= \begin{bmatrix} L_1 & L_2 & \cdots & L_w \end{bmatrix} \end{aligned} \tag{6.3}$$

where the integer w will be defined later. Then, $L(s)D(s)$ can be written by

$$\begin{aligned} & L(s)D(s) \\ &= \boldsymbol{L} S_{I_q}^{w}(s) \begin{bmatrix} D_0 & D_1 & \cdots & D_{\mu} \end{bmatrix} S_{I_m}^{\mu}(s) \\ &= \boldsymbol{L} \begin{bmatrix} D_0 & D_1 & \cdots & D_w & \cdots & D_{\mu} & 0 & \cdots & 0 \\ 0 & D_0 & \cdots & D_{w-1} & \cdots & D_{\mu-1} & D_{\mu} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & D_0 & \cdots & D_{\mu-w} & D_{\mu-w+1} & \cdots & D_{\mu} \end{bmatrix} S_{I_m}^{\mu+w}(s) \end{aligned} \tag{6.4}$$

where $D_{\mu-w}=0$ if $\mu<w$. Assume that the μ-th degree's coefficient matrix of $L(s)D(s)$ is $K\in\Re^{q\times m}$. If $L(s)$ is the regularizing matrix, then the following equality must hold from the above relation:

$$\boldsymbol{L}\boldsymbol{T}_w=\boldsymbol{J} \tag{6.5}$$

where

$$\begin{aligned}\boldsymbol{T}_{\boldsymbol{w}} &= \begin{bmatrix} D_\mu & 0 & \cdots & 0 \\ D_{\mu-1} & D_\mu & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ D_{\mu-w} & D_{\mu-w+1} & \cdots & D_\mu \end{bmatrix} \\ \boldsymbol{J} &= \begin{bmatrix} K & 0 & \cdots & 0 \end{bmatrix}. \end{aligned} \tag{6.6}$$

Considering the structure of $\boldsymbol{J}$, set

$$\boldsymbol{L}=\boldsymbol{J}\boldsymbol{T}_w^{+}=K\boldsymbol{T}_w^{+}(1:m,:), \tag{6.7}$$

where $\boldsymbol{T}_w^{+}(1:m,:)$ denote the submatrix constituted of the first m-th rows of $\boldsymbol{T}_w^{+}$. Substituting the above equation to Eq. 6.5,

$$K\boldsymbol{T}_w^{+}(1:m,:)\boldsymbol{T}_w=\boldsymbol{J}. \tag{6.8}$$

Define Λ by

$$\Lambda=\boldsymbol{T}_w^{+}(1:m,:)\begin{bmatrix} D_\mu \\ D_{\mu-1} \\ \vdots \\ D_{\mu-w} \end{bmatrix}, \tag{6.9}$$

the first m-th columns of Eq. 1.8 can be written by

$$K\Lambda=K. \tag{6.10}$$

That is, if Eq. 6.5 is solvable, its special solution is given by Eq. 6.7 and K must satisfy Eq. 6.10. Let $U\begin{bmatrix}\Gamma & 0\\ 0 & 0\end{bmatrix}V^T$ denote the singular value decomposition (SVD) of $\boldsymbol{T}_w$ using some nonsingular matrix Γ and unitary matrices U and V. Then, $\boldsymbol{T}_w^{+}$ is given by

$$\boldsymbol{T}_w^{+}=V\begin{bmatrix}\Gamma^{-1} & 0\\ 0 & 0\end{bmatrix}U^T$$

and

$$\boldsymbol{T}_w^{+}\boldsymbol{T}_w = V\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}V^T.$$

Therefore, Λ can be written by

$$\Lambda = V(1:m,:)\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}V^T(:,1:m) \geq 0. \tag{6.11}$$

Eq. 6.10 means that K is the left eigenvectors of Λ which correspond to the eigenvalues at $\lambda = 1$. Since Λ is a real symmetric matrix, the geometric multiplicity of the eigenvalue *one* in Λ equals to the algebraic multiplicity. Thus we can find a set of linearly independent eigenvectors for the eigenvalue *one*. Therefore,

1. w is the least integer when Λ has p multiple eigenvalue at $\lambda = 1$.
2. K is constituted of corresponding left eigenvectors.

Example 1 Consider the following polynomial matrix:

$$D(s) = \begin{bmatrix} s+1 & s+2 & s+3 \\ s+4 & s+5 & s+6 \end{bmatrix}.$$

For the above case, $q = 2$, $m = 3$ and $\mu = 1$. D_0 and D_1 are given by

$$D_0 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Setting $w = 2$, $\boldsymbol{T}_2$ is given by

$$\boldsymbol{T}_2 = \begin{bmatrix} D_1 & 0 & 0 \\ D_0 & D_1 & 0 \\ 0 & D_0 & D_1 \end{bmatrix}$$

and then Λ is given by

$$\Lambda = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$$

which has the eigenvalue at $\lambda = 1$ with multiplicity $2 = p$. The left eigenvectors of Λ corresponding to $\lambda = 1$ are given by [1 0 –1] and [0 1 2] and thus K is given by

$$K = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix}.$$

Therefore, $L(s)$ can be calculated by

$$L(s) = \begin{bmatrix} K & 0 & 0 \end{bmatrix} \boldsymbol{T}_2^+ S_1^2(s)$$
$$= \begin{bmatrix} .5385 & .5385 \\ -.3846 & -.3846 \end{bmatrix} + s \begin{bmatrix} -1.3077 & .3077 \\ 1.0769 & -.0769 \end{bmatrix} + \frac{s^2}{3} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

6.3 Applications to Descriptor Systems

The descriptor system is given by the following equations:

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) \end{aligned} \tag{6.12}$$

where $x(t) \in \Re^n$ is the descriptor vector, $u(t) \in \Re^m$ is a control input vector, $y(t) \in \Re^q$ is an output vector, and E, $A \in \Re^{n \times n}$, $B \in \Re^{n \times m}$ and $C \in \Re^{q \times n}$ are constant matrices. It is assumed that rank $E = r < n$ and (E, A) is *regular*, i.e., $\det(sE - A) \neq 0$ for almost of all s.

It is known that there are three modes for the descriptor system 6.12. In the followings, we will analyze the impulsive mode using the regularizing matrix.

Let $\varphi(s)$ denote the characteristic polynomial of (E, A), i.e.,

$$\varphi(s) = \det(sE - A), \quad \deg \varphi(s) := d. \tag{6.13}$$

The zeros of the above polynomial are called *dynamics mode* of the system 6.12. Since E is singular, the system 6.12 has infinite mode. If $r = d$, then the infinite mode is called *static*. If $d < r$, then the system 6.12 has *impulsive mode*.

Lemma 1: If the regularizing polynomial matrix of $sE - A$ can be described as a first order polynomial matrix, i.e.,

$$L(s) = L_0 + sL_1, \quad L_1 \neq 0, \tag{6.14}$$

then the system 6.12 has no impulsive modes. Conversely, if the system 6.12 has no impulsive modes, then there exists a first order regularizing polynomial matrix.

(Proof). Consider the SVD of E as follows:

$$E = U \begin{bmatrix} E_1 & 0 \\ 0 & 0 \end{bmatrix} V^T, \quad U, V \in \Re^{n \times n}, \tag{6.15}$$

wher E_1 is nonsingular. According to the above decomposition, A and $L(s)$ are decomposed by

$$
\begin{aligned}
A &= U\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} V^T, \\
L(s)U &= [L_{01} \;\; L_{02}] + s[L_{11} \;\; L_{12}], \\
A_{11} &\in \Re^{r\times r}, \qquad A_{12} \in \Re^{r\times(n-r)}, \\
A_{21} &\in \Re^{(n-r)\times r}, \quad A_{22} \in \Re^{(n-r)\times(n-r)}, \\
L_{i1} &\in \Re^{n\times r}, \qquad L_{i2} \in \Re^{n\times(n-r)}, \quad i = 0,1.
\end{aligned} \tag{6.16}
$$

It is known that the system 6.12 has no impulsive modes if and only if A_{22} is nonsingular. Thus, it will be shown that nonsingularity of A_{22} if $L(s)$ is given by Eq. 6.14. Now,

$$
\begin{aligned}
&L(s)(sE - A) \\
&= \left([L_{01} \;\; L_{02}] + s[L_{11} \;\; L_{12}]\right)\begin{bmatrix} sE_1 - A_{11} & -A_{12} \\ -A_{21} & -A_{22} \end{bmatrix} V^T \\
&= s^2[L_{11}E_1 \;\; 0] - s\left[[L_{11} \;\; L_{12}]\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}\right] \\
&-\left([L_{01} \;\; L_{02}]\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + L_{01}E_1[L_{11} \;\; L_{12}]\right) V^T
\end{aligned} \tag{6.17}
$$

Since $L(s)$ is a regularizing matrix, the second degree coefficient matrix must be zero. Thus, $L_{11}E_1 = 0$. Since E_1 is nonsingular,

$$
E_1 = 0. \tag{6.18}
$$

Then, Eq. 6.17 can be written by

$$
\begin{aligned}
L(s)(sE - A) \;\; = \;\; & -[L_{01} \;\; L_{02}]\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} V^T \\
& -s[L_{12}A_{21} - L_{01}E_1 \;\; L_{12}A_{22}]V^T.
\end{aligned}
$$

Again, since $L(s)$ is a regularizing matrix, the first degree coefficient matrix must be nonsingular. Thus, $L_{12}A_{22}$ must have column full rank. Therefore, A_{22} must be nonsingular.

Conversely, if A_{22} is nonsingular, define $L(s)$ by

$$
L(s) = \begin{bmatrix} I_r & 0 \\ 0 & sI_{n-r} \end{bmatrix} U^T. \tag{6.19}
$$

Lemma 2: If the system 6.12 has some impulsive modes, then there exists a regularizing polynomial matrix which degree is greater than or equals to two. Conversely, if the regularizing polynomial matrix of $sE - A$ cannot be described as a

first degree polynomial matrix, but there exists a regularizing polynomial matrix which degree is greater than or equals to two, i.e.,

$$L(s) = L_0 + sL_1 + s^2 L_2,\ L_2 \neq 0,$$

then the system 6.12 has some impulsive modes.

(Proof). Consider the Weierstrass form of (E, A) as follows:

$$S^{-1}ET = \begin{bmatrix} I_d & 0 \\ 0 & N \end{bmatrix}, \quad S^{-1}AT = \begin{bmatrix} \Lambda & 0 \\ 0 & I_{n-d} \end{bmatrix}. \tag{6.20}$$

where S and T are nonsingular, and N is given by

$$\begin{aligned} N &= diag\{N_1,\ N_2,\ \ldots,\ N_\alpha\} \in \Re^{(n-d)\times(n-d)}, \\ N_i &= \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \in \Re^{k_i \times k_i}, \quad \sum_{i=1}^{\alpha} k_i = n-d. \end{aligned} \tag{6.21}$$

If the system has some impulsive modes, then $k_i \geq 2$ for some i. In this case $N_i \neq 0$ and thus $N \neq 0$. Then, there exists a unimodular matrix $U_2(s)$ which degree is $\max(k_i - 1)$ such that

$$U_2(s)(sN - I_{n-d}) = I_{n-d},\ U_2(s) = (sN - I_{n-d})^{-1} \tag{6.22}$$

Then, a regularizing polynomial matrix $L(s)$ is given by

$$L(s) = \begin{bmatrix} I_d & 0 \\ 0 & sU_2(s) \end{bmatrix} S^{-1}. \tag{6.23}$$

Since $U_2(s)$ is at least first order polynomial matrix, the order of $L(s)$ is greater than or equals to two.

Conversely, assume that the regularizing polynomial matrix of $sE - A$ cannot be described as a first degree polynomial matrix, but there exists a regularizing polynomial matrix which degree is greater than or equals to two. From the definition of the regularizing polynomial matrix, there exists an $n \times n$ matrix $\bar{A}$ such that

$$L(s)(sE - A) = sI - \bar{A}.$$

Then, $(sI - \bar{A})^{-1}$ can be written by

$$(sI - \bar{A})^{-1} = \frac{s^{n-1}I + \text{lower degree terms}}{\det(sI - A)}.$$

Since $L(s)$ is assumed to be the polynomial matrix which degree is greater than one,

$$\begin{aligned}(sE-A)^{-1} &= \{L(s)(sE-A)\}^{-1}L(s)\\ &= (sI-\overline{A})^{-1}L(s)\end{aligned}$$

is improper and thus (E, A) has some impulsive modes.

The system 6.12 is said to be impulsive mode controllable if there exists a feedback gain matrix F such that $sE-A+BF$ has no impulsive mode. From the above Lemmas, we can obtain a necessary and sufficient condition for impulsive controllability.

Theorem 1: The system 6.12 is impulsive controllable if and only if there exists a feedback gain matrix $F \in \Re^{m\times n}$ such that

$$\operatorname{rank}\begin{bmatrix} E & 0\\ A-BF & E\\ I_n & 0\end{bmatrix} = \operatorname{rank}\begin{bmatrix} E & 0\\ A-BF & E\end{bmatrix}. \tag{6.24}$$

(Proof). From Lemma 1, the closed-loop system

$$\begin{aligned}&Ex(t) = (A-BF)x(t)+Bu(t),\\ &y(t) = Cx(t)\end{aligned}$$

has no impulsive modes if there exists a first degree regularizing polynomial matrix. In this case, $w=1$, and then T_2 and J are given by

$$\boldsymbol{T}_2 = \begin{bmatrix} E & 0\\ -A+BF & E\end{bmatrix}, \quad \boldsymbol{J} = [I_n \quad 0].$$

Eq. 6.7 is solvable if and only if

$$\operatorname{rank}\begin{bmatrix}\boldsymbol{T}_2\\ \boldsymbol{J}\end{bmatrix} = \operatorname{rank}\boldsymbol{T}_2.$$

Thus, Eq. 6.24 can be obtained from the above equation.

Conversely, if Eq. 6.24 holds, then there exists a first degree regularizing matrix. Then, from Lemma 1, the closed-loop system has no impulsive modes, i.e., the open-loop system is impulsive controllable.

Lemma 3: Define the SVD of E by Eq. 6.15. Corresponding decomposition of A is defined by Eq. 6.16 and decomposition of B and F are defined by

$$\begin{aligned}&U^TBV = \begin{bmatrix}B_1\\ B_2\end{bmatrix} && B_1 \in \Re^{r\times m},\\ &&& B_2 \in \Re^{(n-r)\times m},\\ &U^TFV = [F_1 \quad F_2], && F_1 \in \Re^{m\times r}, \quad F_2 \in \Re^{m\times(n-r)}.\end{aligned} \tag{6.25}$$

Then, the system 6.12 is impulsive controllable if and only if there exists a gain matrix F_2 which makes $A_{22} - B_2 F_2$ be nonsingular.

(Proof). Since E_1 is nonsingular,

$$\operatorname{rank}\begin{bmatrix} E & 0 \\ A - BF & E \\ I_n & 0 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} E_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ A_{11} - B_1 F_1 & A_{12} - B_1 F_2 & E_1 & 0 \\ A_{21} - B_2 F_1 & A_{22} - B_2 F_2 & 0 & 0 \\ I_r & 0 & 0 & 0 \\ 0 & I_{n-r} & 0 & 0 \end{bmatrix}$$

$$= \operatorname{rank}\begin{bmatrix} E_1 & 0 & 0 \\ 0 & 0 & E_1 \\ 0 & A_{22} - B_2 F_2 & 0 \\ 0 & I_{n-r} & 0 \end{bmatrix}$$

Thus, Eq. 6.25 holds if and only if there exists a gain matrix K_2 which makes $A_{22} - B_2 F_2$ be nonsingular.

From the view point of the transfer function matrix, $(sE - A)^{-1} B$ is proper if and only if $sE - A$ is row proper. Thus, the problem is to find the feedback gain matrix which makes $sE - A + BF$ be row proper. By an elementary row operation matrix W, sE can be decomposed by

$$sWE = \begin{bmatrix} sE_1 \\ E_2 \end{bmatrix}, \quad \begin{matrix} E_1 \in \Re^{\mu \times n} \\ E_2 \in \Re^{(n-\mu)\times n}, \end{matrix}$$
$$\operatorname{rank} E_1 = \mu. \tag{6.26}$$

According to the decomposition, A and B are also decomposed by

$$WA = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \quad \begin{matrix} A_1 \in \Re^{\mu \times n}, \\ A_2 \in \Re^{(n-\mu)\times n}, \end{matrix}$$
$$WB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \begin{matrix} B_1 \in \Re^{\mu \times m}, \\ B_2 \in \Re^{(n-\mu)\times m}. \end{matrix} \tag{6.27}$$

Theorem 2: Let

$$\overline{A} := \begin{bmatrix} E_1 \\ A_2 \end{bmatrix}, \quad \overline{B} := \begin{bmatrix} 0_{\mu \times m} \\ B_2 \end{bmatrix}. \tag{6.28}$$

The descriptor system is impulsive controllable if and only if there exists a feedback gain matrix $\overline{F}$ such that $\overline{A} - \overline{B}\overline{F}$ does not have any uncontrollable eigenvalues at the origin.

Example 2: Consider the following E, A and B:

$$E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Then, a regularizing polynomial matrix $L(s)$ of $sE - A$ is given by

$$L(s) = \begin{bmatrix} -s & -s^2 \\ 0.5 & -0.5s \end{bmatrix}$$

and thus there exist some impulsive modes for a given system by Lemma 2. In fact, $sE - A$ is a unimodular polynomial matrix and thus $d = 0$. Since rank $E = 1 > d$, the system has an impulsive mode.

Set $F = [f_1 \; f_2]$. Then,

$$\begin{aligned} \operatorname{rank} \begin{bmatrix} E & 0 \\ A - BF & E \end{bmatrix} &= \operatorname{rank} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ f_1 & f_2 + 1 & 0 & 0 \end{bmatrix} \\ &= \operatorname{rank} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ f_1 & f_2 + 1 & 0 \end{bmatrix}. \end{aligned}$$

If we choose $f_1 \neq 0$ and f_2 arbitrary, Eq. 6.25 holds. Therefore, the system is impulsive controllable by Theorem 1.

On the other hand,

$$\overline{A} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \overline{B} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad \overline{A}\overline{B} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

Since the pair $(\overline{A}, \overline{B})$ is controllable, we can find a feedback gain matrix $\overline{F}$ which makes $\overline{A} - \overline{B}\overline{F}$ be nonsingular.

6.4 Conclusins

In this chapter, a regularizing polynomial matrix was proposed. Using the matrix, an approach to the descriptor systems by polynomial matrix was proposed. A feedback controller design which removes the impulsive modes was shown.

References

1. Luenberger DG (1977) Dynamic equation in descriptor form, IEEE Trans. Automat. Contr., 22:312–321.
2. Dai L (1989) Singular control systems, Springer-Verlag, Berlin
3. Lewis A (1992) A tutorial on the geometric analysis of linear time-invariant implicit systems, Automatica, 28:119–137
4. Verghese GC, Levy BC, Kailath T (1981) A generalized state-space for singular systems, IEEE Trans. Automat. Contr., 26:811–831
5. Wolovich WA, Falb PL (1976) Invariants and canonical form under dynamic compensations, SIAM J. Contr. and Optim., 14:996–1008
6. Mutoh Y, Ortega R (1993) Interactor structure estimation for adaptive control of discrete-time multivariable nondecouplable systems, Automatica, 29:635–647
7. Kase W, Mutoh Y (2009) A simple derivation of interactor matrix and its applications, Int. J. Systems Sciences 40:1197–1205

Chapter 7
Analysis of the Electric Field Distribution in a Wire-Cylinder Electrode Configuration

K. N. Kiousis, A. X. Moronis and W. G. Früh

Abstract The electric field distribution in an air gap between a wire-cylinder electrode configuration, has been studied by implementing Finite Element Analysis. The electrodes were assumed to be surrounded by air at normal conditions, while high dc voltage has been applied across them, with positive polarity at the wire. Numerical analysis on the maximum electric field intensity along the wire-cylinder gap axis, as well as on the potential distribution in the air surrounding the electrodes has been carried out, considering different geometrical characteristics of the electrodes. The applied mesh parameters were optimized, in terms of accuracy and processing power. The maximum field intensity was mainly associated with the wire radius r and the electrode gap length d. The cylindrical electrode radius R had a limited impact on the maximum electric field intensity but, on the other hand, it had a strong effect in the distribution of the electric field lines. Finally, a formula for the estimation of the maximum electric field intensity is proposed.

Keywords Finite element analysis · HV electrodes · Modeling · Numerical analysis

7.1 Introduction

The study of the electric field strength distribution is of great importance for the design and dimensioning of high voltage equipment [1–3]. The electric field strength is the key parameter that defines the behaviour of insulating materials under high electric field stress. There are numerous applications of high voltage technology in electrical power systems, in industry and research. The experimental measurement

K. N. Kiousis (✉) · W. G. Früh
Institute of Mechanical, Process and Energy Engineering, Heriot-Watt University, Edinburgh EH14 4AS, UK
e-mail: konstantinosq@gmail.com

W. G. Früh
e-mail: w.g.fruh@hw.ac.uk

A. X. Moronis
Energy Technology Department, Technological Educational Institute of Athens, Aegaleo 12210, Greece
e-mail: amoronis@teiath.gr

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,* Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_7,

of the field strength in air gaps is in fact difficult and not quite accurate, due to the presence of sensing elements which may affect the distribution of the electric field. On the other hand, computer methods can provide instant and accurate results and are capable of solving problems in more complex conditions. Some of the most commonly used software applications implement the Finite Element Analysis (*FEA*).

FEA modelling in electrostatics is based on the application of a set of differential equations that describe the problem, considering certain boundary conditions, in order to come to a unique solution. *FEA* modelling breaks the problem down into a large number of regions, each with a simple geometry (e.g. triangles), defined by a mesh with a very large number of nodes. Then the problem is transformed from a small but difficult to solve problem into a big but relatively easy to solve problem, involving a very large number of unknown quantities.

Despite the large number of computational studies of the electric field distribution in uniform and non-uniform electric fields at different electrode configurations (e.g. parallel planes, tip-plane, concentric cylinders, wire-wire etc.) found in literature [4–10], there is no study available for the wire-cylinder electrode arrangement. On the other hand, experimental investigations, by means of corona discharge current [11], [12] and current distribution [13], have already been conducted for wire-cylinder electrode pairs.

The goal of this study is the fine modeling and analysis of the electric field strength in a wire-cylinder electrode arrangement in atmospheric air at normal conditions, considering the geometrical characteristics of the electrodes. In this study the wire radius r ranged from *1* to *500μm*, the cylinder radius R from *1* to *20 mm* and the gap d between the electrodes ranged from *1* to *10 cm*. These dimensions are quite common in experimental studies of corona discharge currents and the corresponding electro-hydrodynamic (*ehd*) effects, which can be found in bibliography [11–13].

This analysis was based on *FEA* techniques. On this purpose, open source *FEA* modeling software *F.E.M.M. ver. 4.2* has been implemented. The mesh parameters have been fully investigated in order to optimize the applied mesh around the specific areas of interest, such as the inter-electrode region and especially at points along the line defining the shortest distance d between the electrodes, thus ensuring the accuracy of the results.

7.2 Governing Equations

In our case we have a typical electrostatics problem which is governed by the well-known Gauss's and Poisson's equations, assuming homogenous field and steady state conditions [1–3]:

$$E = -\nabla V \tag{7.1}$$

$$\nabla^2 V = -\rho / \varepsilon_0 \tag{7.2}$$

where E is the electric field intensity, V is the applied voltage, ρ is the space charge density and ε_0 is the dielectric permittivity of air. The electric field should satisfy the charge conservation law:

$$\nabla \cdot j = 0 \tag{7.3}$$

where j is the current density. The latter is defined as:

$$j = \rho \cdot u = \rho \cdot \mu \cdot E \tag{7.4}$$

where u is the ion drift velocity and μ is the ion mobility. (7.1), (7.2), (7.3) and (7.4) can be combined to obtain:

$$\nabla\left\{\left(\nabla^2 V\right)\left(\nabla V\right)\right\} = 0 \tag{7.5}$$

In theory, the physical problem is reduced to the mathematical problem of solving (7.5) with the appropriate boundary conditions. The *FEA* model provides numerical results for the voltage distribution at each node of the applied mesh. The electric field strength may then be easily determined by (7.1) around the user-defined domain, where the mesh is constructed.

In such a computational analysis, the *solver precision*, the *boundary conditions*, the *bounding box size* defining the domain and the *mesh distribution* are of great importance for the accuracy of the results [14–18].

7.3 Electrode Geometry

The wire-cylinder electrode pair under consideration is shown in Fig. 7.1. A thin cylindrical wire of radius r is placed parallel to a cylinder with significantly larger radius R, at distance d. The wire radius r, the cylinder radius R and the air gap d between the electrodes are critical parameters which define the geometry and, in this way, determine the electric field strength. In this study, r was ranging between *1* and *500μm*, R between *1* and *20 mm*, while the gap d between the electrodes was ranging between *1* and *10 cm*. In addition, it has been assumed that positive potential has been applied to the wire while the cylinder is grounded.

Theoretically, these electrodes may have infinite length, but due to the longitudinal axis symmetry, the electric field or potential distribution may only change in the radial direction, perpendicular to the wire or the cylinder surface, along the gap. Therefore, this three-dimensional problem may be minimized into a two-dimensional problem that requires a much smaller number of nodes and less computing power.

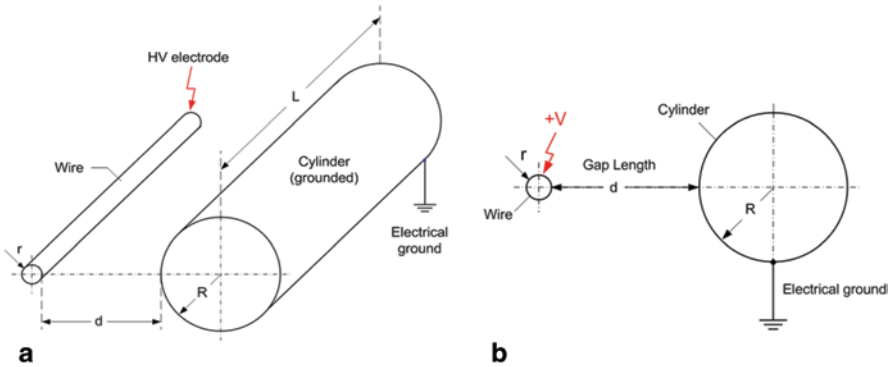


Fig. 7.1 **a** Perspective plan of the electrodes arrangement in space and **b** simplified planar model, due to longitudinal axis symmetry

7.4 Modelling Parameters

F.E.M.M. solves (7.5) for the potential V, over the user defined domain with the user defined sources and boundary conditions. It discretizes the problem domain using triangular elements, which form a mesh consisting of a large number of nodes. The solution over each element is approximated by a linear interpolation of the values of potential at the three vertices of the triangle [19]. In our case, a two-dimensional planar electrostatic problem was defined with a solver precision 10^{-8}.

Due to the symmetry of the electrode geometry along the gap axis, half-plane modeling has been applied. The problem's domain was defined by the bounding box shown in Fig. 7.2. This box sets the limits of the surrounding dielectric medium, which in our case was atmospheric air. The bounding box size was defined by the fixed distances $A = k \cdot D$ between its sides and the electrodes, where $D = 2r + d + 2R$ was the total length of the electrodes assembly (air gap included) and k was a scaling constant.

The determination of the bounding box size is generally critical, since a small box may affect the electric field distribution and lead to errors, or, on the other hand, a large box may unnecessarily lead to a very large number of nodes, demanding more processing power. Preliminary analysis with different k values, has shown that a suitable choice would be $k = 3$ [20]. This value had been kept constant throughout all simulations.

Dirichlet conditions [21] were explicitly defined on the problem's boundaries. The wire and cylinder outer surfaces were considered to be equipotentials with fixed voltages *1 kV* and *0 V*, respectively. Subsequently, all electric field strength results were defined per *kV* of the applied voltage. Since the applied voltage may vary in practice, valid results may be easily obtained in any case, by just multiplying the electric field strength at *1 kV*, by the number of applied kilovolts. This can be

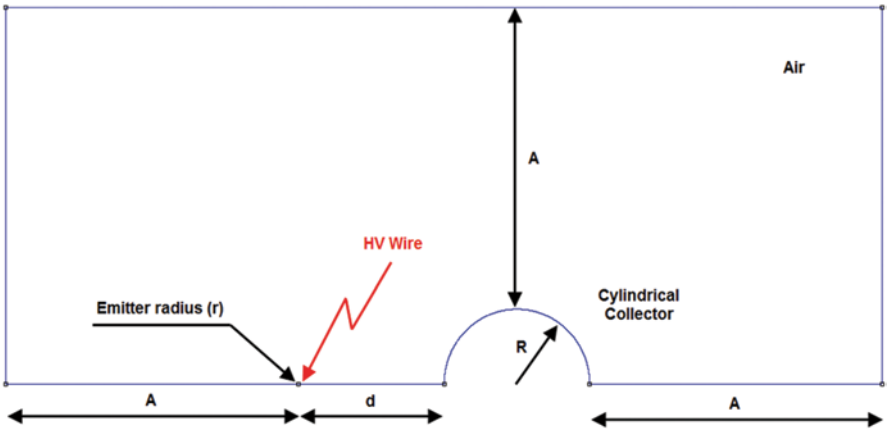


Fig. 7.2 The model of the two electrodes and the bounding box of the surrounding air, where the mesh is applied

easily explained by (7.1). For example, considering any fixed pair of wire-cylinder electrodes with *1 kV* voltage difference, then, if $V_1(x, y)$ is the potential and $E_1(x, y)$ is the corresponding electric field strength at any point (x, y), then, at αkV, the potential would be $V_a(x, y) = a \cdot V_1(x, y)$ and the corresponding electric field strength $E_a(x, y)$ could be determined by (7.1) as follows:

$$E_a(x,y) = -\nabla V_a(x,y) = -\nabla\left(a \cdot V_1(x,y)\right) = a \cdot \left(-\nabla V_1(x,y)\right) = a \cdot E_1(x,y) \qquad (7.6)$$

It becomes clear that the electric field strength at αkV equals α times the electric field at *1 kV*.

There is a set of key parameters to the *F.E.M.M.* model so as to ensure proper mesh formation. The mesh discretization at distances very close to the electrode surfaces depends mainly on two parameters, the *maximum arc segment degrees* and the *minimum angle*. These determine the size of the triangular elements near the outer surface of the electrodes, where the electric field and voltage gradients get their maximum values, thus demanding very fine analysis by a dense mesh. On the other hand, the mesh distribution along the gap is a function of the *local element size along line* parameter. The density of the mesh elements in other areas such as the inter-electrode space away from the electrode surfaces are functions of another key parameter, the *mesh size*.

Analytical study of the influence of each one of the mesh parameters has been carried out, by running a large number of simulations, using different *maximum arc segment angles*, *minimum angles* of the triangular mesh, *local elements size along line* and *mesh sizes*, in order to accomplish convergence of the results. In this way an optimal mesh has been configured, in terms of accuracy and processing power consumption, with the key parameter values given in Table 7.1.

Table 7.1 Comparison between the default values of the *F.E.M.M.* key parameters and the selected optimized values

F.E.M.M. key parameter	Default values	Selected values (optimized)
Minimum angle (degrees)	30	31
Maximum arc segment (degrees)	5	0.5
Local element size along line (μm)	Auto	10
Mesh size (μm)	Auto	Auto
Nodes	3632	22523
Elements	6906	41026

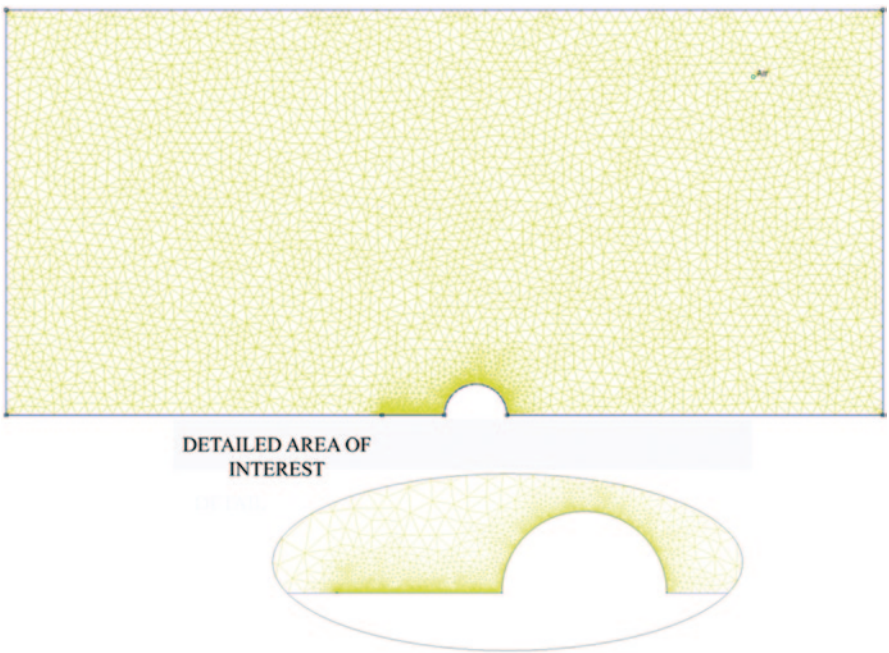


Fig. 7.3 Optimized mesh layout and detailed area of interest. Wire-cylinder configuration with geometrical parameters: $r = 25\mu m$, $R = 15\ mm$ and $d = 3\ cm$. (Here $A = 3D$)

An example of the optimized mesh is shown in Fig. 7.3 for a wire-cylinder electrode setup with $r = 25\mu m$, $R = 15\ mm$ and $d = 3\ cm$. The mesh is denser in the areas of interest, i.e. near the high voltage and the grounded electrode, as well as along the gap axis.

For verification purposes, the optimized mesh has been used in order to estimate the maximum field intensity E_{max}, at well-known geometries, similar to the wire-cylinder pair, such as two identical cylindrical conductors in parallel (where $R/r = 1$), for which analytical formulas can be found in bibliography. In this way, the accuracy of the optimized mesh could be easily tested.

Table 7.2 Comparison between the electric field intensity (theoretical and *F.E.M.M.* results) for two identical parallel conductors at *1 kV* potential difference

r(μm)	d(cm)	Theoretical E_{max} (V/m)	Optimized Mesh results E_{max} (V/m)	Relative error (%)
1	1	$54{,}296 \times 10^6$	$54{,}183 \times 10^6$	0.21
25	3	$2{,}825 \times 10^6$	$2{,}803 \times 10^6$	0.78
100	5	807250	799771	0.93
250	7	357014	353483	0.99
500	10	190261	187465	1.47

In the case of two parallel cylindrical conductors E_{max} is given by the analytical formula [14]:

$$E_{\max} = \frac{V}{d} \cdot \frac{\sqrt{\left(d/2r\right)^2 + \left(d/r\right)}}{\ln\left(1 + \left(d/2r\right) + \sqrt{\left(d/2r\right)^2 + \left(d/r\right)}\right)} \tag{7.7}$$

where V is the applied voltage, d is the distance between the two electrodes and r is the electrode radius.

F.E.M.M. simulations, that have been conducted with the optimized mesh for two parallel cylindrical wires with r and d values within the limits of our study ($r = 1\text{-}500\mu m$ and $d = 1\text{-}10\ cm$), have provided results which are in good agreement with theoretical expectations, in all cases. Such results are given in Table 7.2, where both theoretical and simulation values for E_{max} are shown, along with the corresponding relative error.

7.5 Simulation Results for the Wire-Cylinder Electrodes

As expected, the simulation results have shown that the maximum electric field strength E_{max} is located at the outer surface of the wire electrode, at the least distant point from the cylinder (see Figs. 7.4a, 7.5a and 7.6), while the minimum field strength E_{min} has been identified at distance x, depending on the R/r ratio (see Fig. 7.6).

On the other hand, the potential distribution across the gap d is shown in Fig. 7.4b and 7.5b, where it becomes clear that equipotentials are in fact cylindrical surfaces with displaced centers along the axis of the electrode gap.

The variation of the electric field intensity across the gap, for different R/r ratios is given in Fig. 7.6, where the normalized field $E(x)/E_{max}$ is shown, at distance x from the wire's surface, expressed as a percentage of the total electrode gap length d.

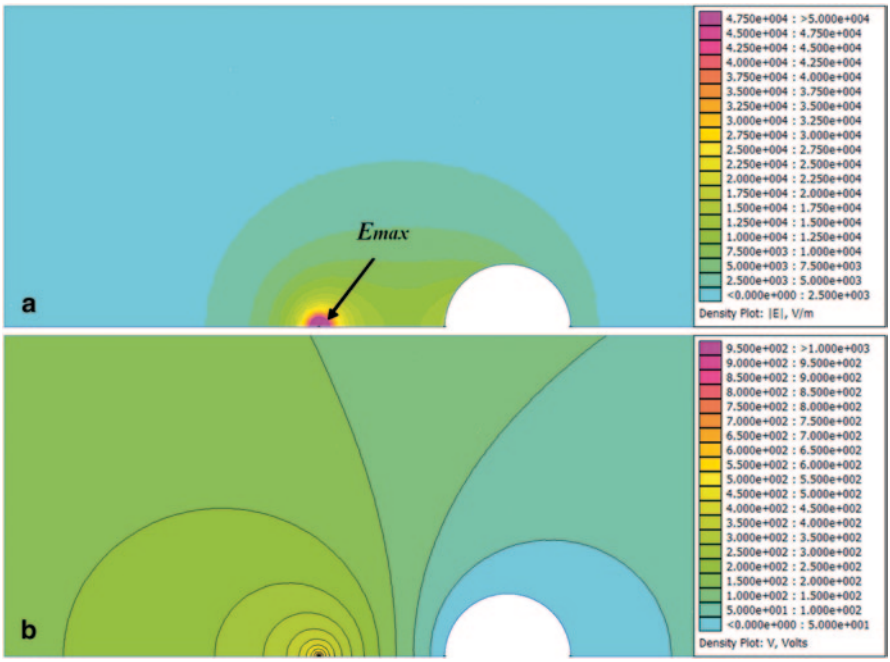


Fig. 7.4 **a** Electric field strength and **b** potential distribution. Wire-cylinder electrodes with $r = 25\mu m$, $R = 15$ *mm* and $d = 3$ *cm*, at *1 kV* potential difference

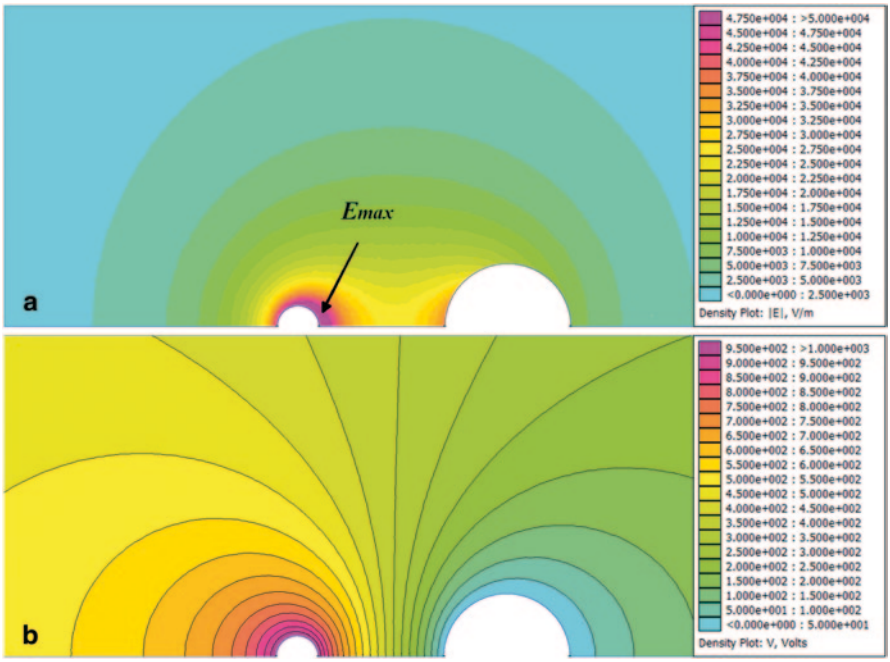


Fig. 7.5 **a** Electric field strength and **b** potential distribution. Wire-cylinder electrodes with $r = 5$ *mm*, $R = 15$ *mm* and $d = 3$ *cm*, at *1 kV* potential difference

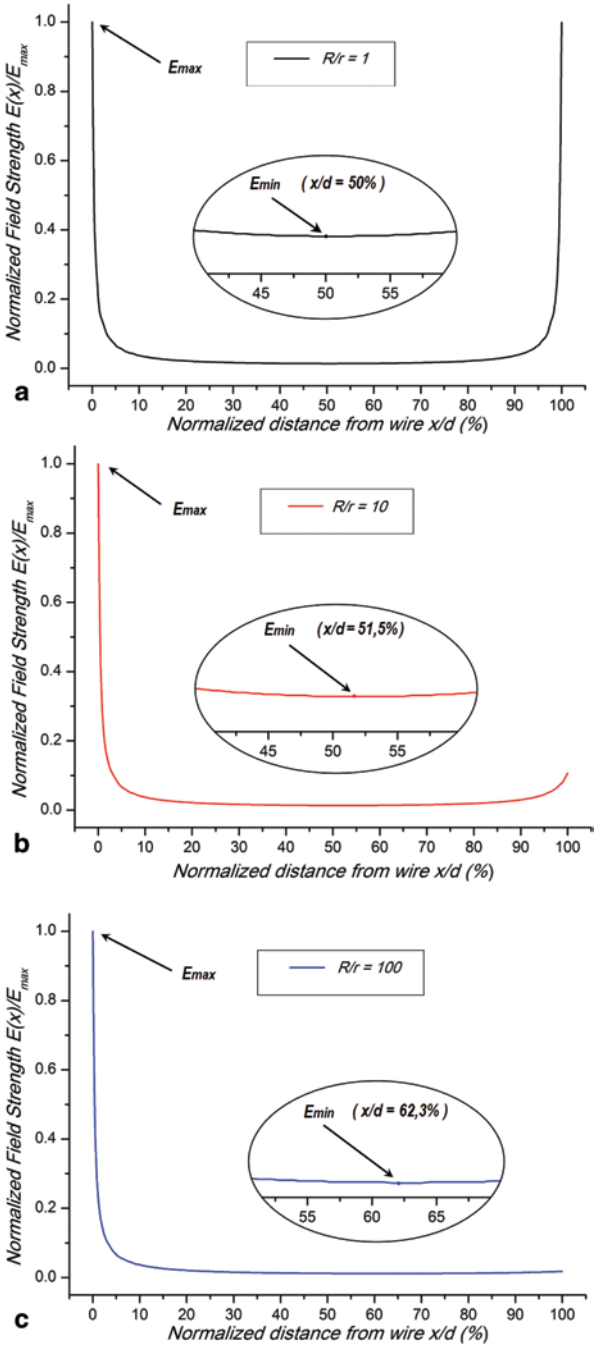


Fig. 7.6 Normalized electric field intensity along the gap axis, and detail where E_{min} is shown. In this case $r = 100\mu m$, $d = 3\ cm$ and **a** $R/r = 1$, **b** $R/r = 10$, and **c** $R/r = 100$. Similar results can be obtained for different d and r values as well

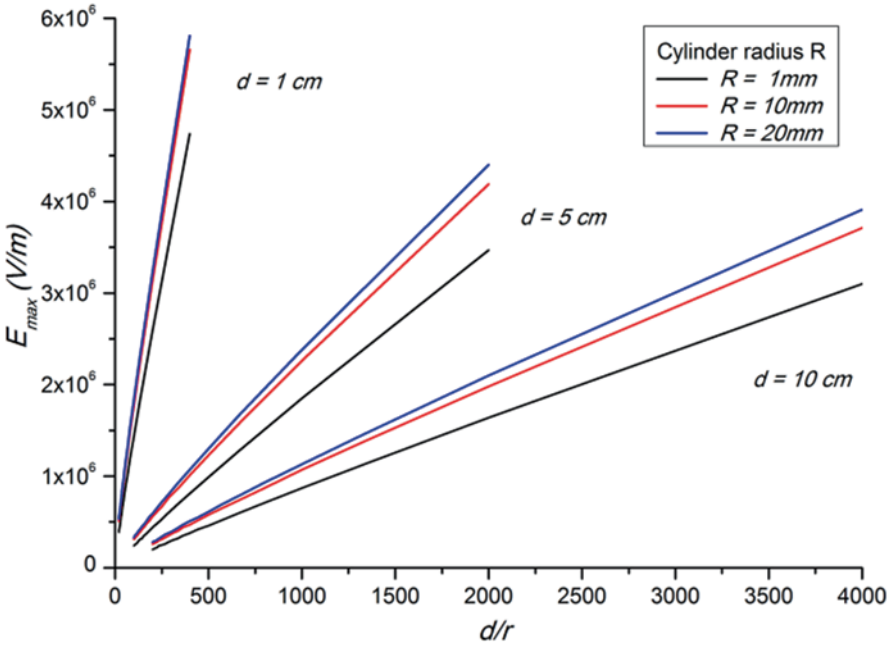


Fig. 7.7 Variation of E_{max} with the *d/r* ratio. Electrode gap at *1 cm*, *5 cm* and *10 cm* and potential difference *1 kV* in all cases

It can be seen that the field intensity gets its maximum value at the wire's surface (where $x = 0$ or $x/d = 0\%$) then diminishes along the gap until its minimum value and, finally, increases at a certain level depending on the *R/r* ratio, at the cylinder's surface (where $x = d$ or $x/d = 100\%$). From another point of view, *R/r* ratio may be considered as a measure of the electric field inhomogeneity, since larger *R/r* ratios result in a more inhomogeneous field distribution along the gap axis (see Fig. 7.6).

The dependence of the maximum electric field intensity E_{max} on the wire radius *r*, the cylinder radius *R* and the electrode gap *d* has also been examined. Figure 7.7 shows typical curves of E_{max} versus *d/r* ratio, for different gaps d and cylinder radii *R*, while Fig. 7.8 shows the variation of E_{max} versus *d/R* ratio for different wire radii *r* and gaps *d*. From these results, it becomes clear that E_{max} is strongly affected by the *d/r* ratio, in a linear way, and secondly, by the *d/R*, ratio.

Gap distance *d* remains a critical parameter in all cases. Generally, maximum electric field intensities can be reached by using thin wires, small electrode gaps and large cylinder radii, which is a reasonable finding, since the electric field distribution is thus becoming strongly inhomogeneous.

On the other hand, Fig. 7.9 and Fig. 7.10 show how the minimum field strength E_{min} along the gap axis, is affected by *d/r*, *d/R* and the electrode gap *d*. Here it seems that the gap distance *d* is the dominant parameter, while the wire radius *r* comes next.

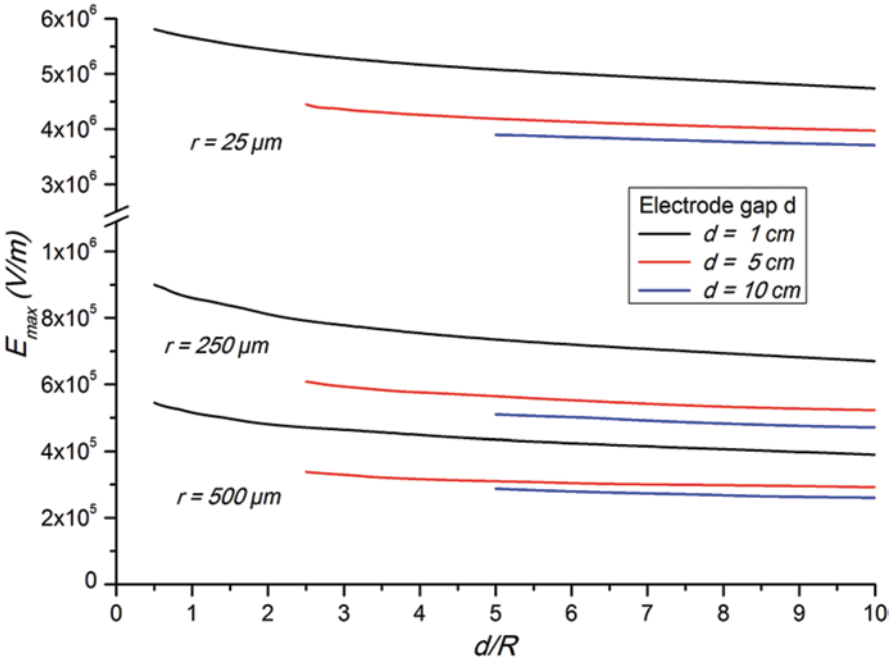


Fig. 7.8 E_{max} variation with the *d/R* ratio. Wire radius at *25µm*, *250µm* and *500µm* and potential difference at *1 kV* in all cases

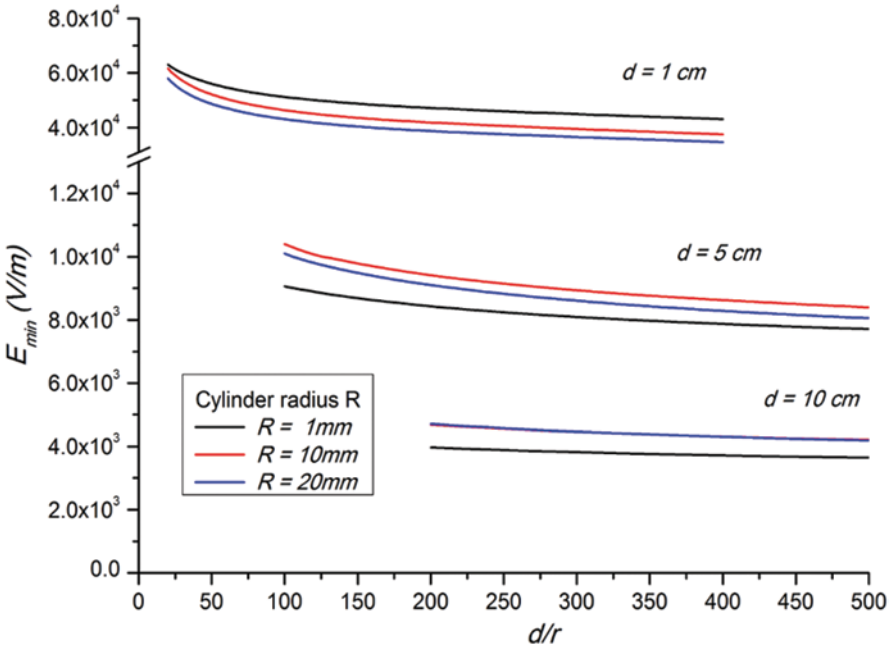


Fig. 7.9 Variation of E_{min} with the *d/r* ratio. Electrode gap at *1 cm*, *5 cm* and *10 cm* and potential difference at *1 kV* in all cases

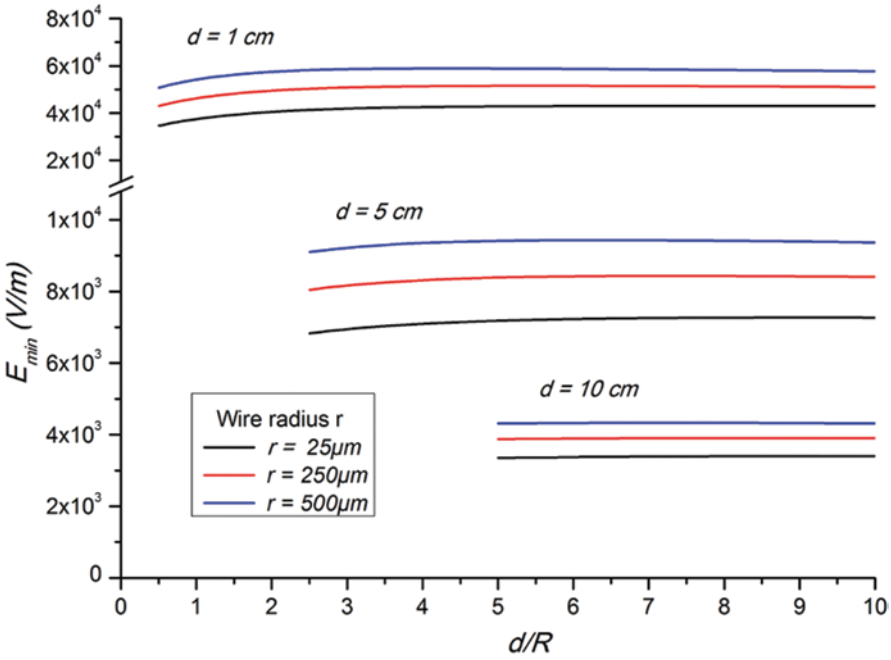


Fig. 7.10 E_{min} variation with the *d/R* ratio. Wire radius at *25μm*, *250μm* and *500μm* and potential difference at *1 kV* in all cases

7.6 Proposed Formula for E_{max} in the Case of a Wire-Cylinder Electrode Arrangement

According to simulation results, E_{max} increases linearly with *d/r* ratio. On the other hand, it is also dependent on the *d/R* ratio. The gap distance *d* is found to be critical in determining the electric field in all cases. This comes in agreement with theoretical expectations, since the ratio *V/d* is frequently used in bibliography as a standard measure of the mean value of the electric field strength in any gap [22–26]. Besides, most of the formulas for E_{max} in well-known geometries are usually expressed as the product of *V/d* by a geometrical constant, as in (7.7) [1–3].

According to the above, an effort has been made to introduce a formula for the maximum electric field strength E_{max} in the wire-cylinder arrangement. Detailed analysis of all simulation results has shown that the maximum electric field intensity $E_{maxW\text{-}C}$ within the limits of this study can be approximated by the following formula:

$$E_{\max W-C} = \frac{V}{d} \cdot \frac{\gamma_1}{\ln\left[\gamma_1\left(\gamma_2 + 2\right)\right]} \tag{7.8}$$

where V is the applied voltage, $\gamma_1 = d/r$ and $\gamma_2 = d/R$ (dimensionless factors).

On the other hand, similarly defining $\gamma' = d/2r$ in (7.7), we have:

$$E_{\max} = \frac{V}{d} \cdot \frac{\sqrt{\gamma'(\gamma'+2)}}{\ln\left[(\gamma'+1)+\sqrt{\gamma'(\gamma'+2)}\right]} \tag{7.9}$$

It should be noted that in the case of two identical cylindrical conductors in parallel $r = R$ ($\gamma_1 = \gamma_2 = \gamma$) and for $d >> r$ we have $\gamma >> 1$, $\gamma + 2 \approx \gamma$.

Then (7.8) becomes:

$$E_{\max} = \frac{V}{d} \cdot \frac{\gamma}{2 \cdot \ln(\gamma)} \tag{7.10}$$

In fact (7.10) equals (7.9) for $d >> r$, since $\gamma' >> 1$, $\gamma' + 1 \approx \gamma'$ and $\gamma' + 2 \approx \gamma'$ (also considering that $\gamma' = \gamma/2$).

7.7 Discussion

The results of (7.8) for all possible combinations of the critical geometrical parameters r, R and d, within the limits of this study ($r \leq 500\mu m$, $R \geq 1$ *mm* and $d \geq 1$ *cm*), are in good agreement with the corresponding maximum field intensity E_{max} values estimated by the *FEA* simulation.

Typical graphs of the change in relative error for E_{max} with the geometrical parameters r, R and d are given in Fig. 7.11. These graphs show that the error diminishes as d and R increase with respect to the wire radius r.

According to Fig. 7.11, the relative error remains small, below *4%* (worst case) and decreases with increasing gap and cylinder radius.

Practically speaking, the proposed formula for E_{max} can be effectively used for electrode pairs constructed by thin wires parallel to cylinders of considerably larger radii at distances of a few centimetres or more. Such electrode arrangements have been used in previous work [11–13] and are suitable for corona or ionic wind applications, due to the high inhomogeneity of the produced electric field. The determination of the maximum electric field in these cases is always one of the most critical design parameters.

Moreover, the electric field utilization factor $n = E_{av}/E_{max}$, (where $E_{av} = V/d$), which is frequently used to indicate the electric field inhomogeneity [27], can be easily defined from (7.8) as:

$$n_{W-C} = \frac{\ln\left[\gamma_1(\gamma_2+2)\right]}{\gamma_1} \tag{7.11}$$

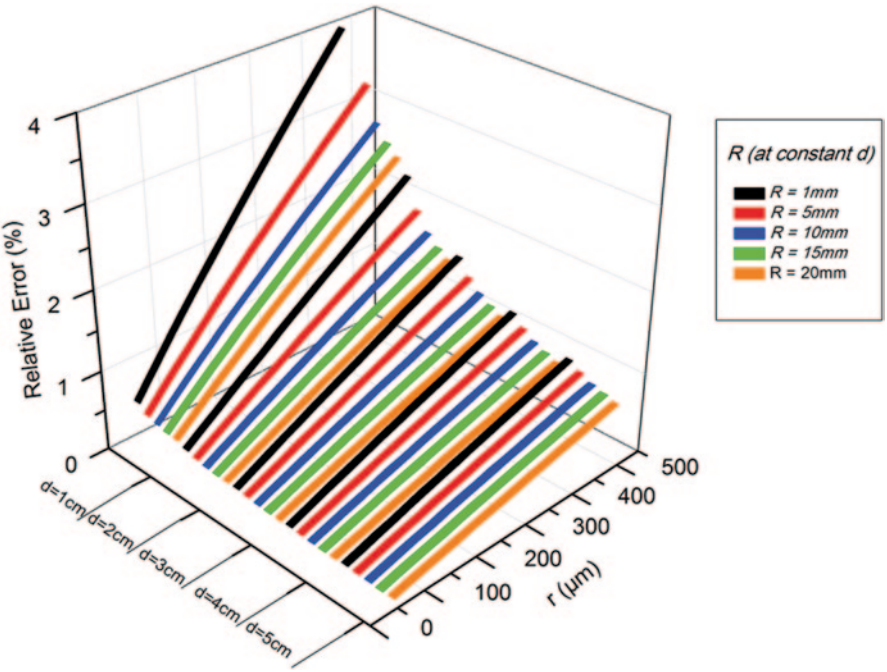


Fig. 7.11 Representation of the relative error between simulated data and the empirical formula results for E_{max} according to (7.8)

7.8 Conclusions

The electric field distribution in a typical wire-cylinder electrode configuration in air, under high voltage dc application has been studied with the aid of dedicated simulation software implementing the Finite Element Analysis. The applied mesh parameters have been optimized and validated, in order to ensure the accuracy of the results. The maximum electric field strength E_{max}, as well as the minimum electric field strength E_{min}, has been examined, considering geometrical characteristics of the electrodes such as the wire radius r, the electrode spacing d and the cylindrical electrode radius R.

Simulations have shown that E_{max} is mainly associated with the wire electrode radius r. Generally, smaller wire radii result in higher field intensities around the wire, especially at the wire's surface, where E_{max} is observed. In addition, E_{max} was found to be inversely proportional to the electrode gap d. Moreover, larger cylinder radii R lead to higher E_{max} values, for constant r and d. On the other hand, E_{min} is strongly related to the electrode gap d, while the wire radius r and the cylindrical electrode radius R have a limited impact on the minimum electric field intensity.

Finally, an empirical formula for the estimation of the maximum electric field intensity has been proposed. The correlation of the simulated data with the empirical formula results was found to be satisfactory, with an absolute error lower than *4 %* in all cases.

References

1. Khalifa M, (1990) High-Voltage Engineering Theory and Practice. Marcel Dekker Inc., New York
2. Naidu CL, Kamaraju V, (1996) High Voltage Engineering. Mc Graw Hill, New York
3. Wadhwa CL, (2007) High Voltage Engineering. New Age International Ltd
4. Hidaka K, Kouno T, (1982) A method for measuring electric field in space charge by means of pockels device. J Electrostatics, 11:195–211
5. McAllister IW, (2002) Electric fields and electrical insulation. In: IEEE Transactions, Dielectrics and Electrical Insulation, 9:672–696
6. Maglaras A, Maglaras L, (2004) Modeling and analysis of electric field distribution in air gaps, stressed by breakdown voltage. In: Math. Methods and Computational Techniques in Electrical Engineering, WSEAS, Athens, pp 1–8
7. Maglaras A, (2004) Numerical Analysis of Electric Field in Air Gaps, Related to the Barrier Effect. In: 1st International Conference from Scientific Computing to Computational Engineering, Athens, pp. 857–865
8. Mackerle J, (2000) Finite element and boundary element modeling of surface engineering systems: A bibliography (1996–1998). J Finite Elements in Analysis and Design, 34:113–124
9. Rezouga M, Tilmatine A, Ouiddir R, Medles K, (2009) Experimental Modelling of the Breakdown Voltage of Air Using Design of Experiments. J Advances in Electrical and Computer Engineering, 9:41–45
10. Rau M, Iftemie A, Baltag O, Costandache D, (2011) The Study of the Electromagnetic Shielding Properties of a Textile Material with Amorphous Microwire. J Advances in Electrical and Computer Engineering, 11:17–22
11. Kiousis KN, Moronis AX, (2011) Experimental Investigation of EHD Flow in Wire to Cylinder Electrode Configuration. In: PES, IASTED, Crete, pp. 21–26
12. Kantouna K, Fotis GP, Kiousis KN, Ekonomou L Chatzarakis GE, (2012) Analysis of a Cylinder-Wire-Cylinder Electrode Configuration during Corona Discharge. In: Circuits, Systems, Communications, Computers and Applications (CSCCA), WSEAS, Iasi, pp. 204–208
13. Morrison RD, Hopstock DM, (1979) The distribution of current in wire-to-cylinder corona. J Electrostatics, 6:349–360
14. Kuffel E, Zaengl WS, Kuffel J, (2000) High Voltage Engineering Fundamentals. Newnes, Oxford
15. Sylvester PP, Ferrari RL, (1996) Finite Elements for Electrical Engineers 3rd Edition. Cambridge University Press, New York
16. Jin J, (2002) Finite Element Method in Electromagnetics 2nd Edition. Wiley-IEEE Press, New York
17. Rao SS, (1999) The Finite Element Method in Engineering. Butterworth-Heinemann, Boston
18. Mackerle J, (1993) Mesh generation and refinement for FEM and BEM – A bibliography (1990–1993). J Finite Elements in Analysis Design, 15:177–188
19. Meeker D, (2010) Finite Element Method Magnetics Ver 4.2 User's Manual
20. Kiousis KN, Moronis AX, (2013) Modeling and Analysis of the Electric Field and Potential Distribution in a Wire-Cylinder Air Gap. In: Computer Engineering and Applications (CEA '13), WSEAS, Milan, pp. 35–40
21. Hlavacek I, Krizek M, (1987) On a super convergent finite element scheme for elliptic systems. I. Dirichlet boundary condition. J App Math, 32:131–154
22. Matsumoto T, (1968) DC corona loss of coaxial cylinders. J Electr. Eng. Japan 88, 12:11–19
23. Giubbilini P, (1988) The current-voltage characteristics of point-to-ring corona. J Applied Physics, 64: 3730–3732
24. Waters RT, Rickard TS, Stark WB, (1970) The Structure of the Impulse Corona in a Rod/Plane Gap I The Positive Corona. In: Proc. R. Soc., pp. 1–25

25. Ferreira GF, Oliveira ON, Giacometti JA, (1996) Point-to-plane corona: Current-voltage characteristics for positive and negative polarity with evidence of an electronic component. J Applied Physics, 59:3045–3049
26. Carreno F, Bernabeu E, (1994) On wire-to-plane positive corona discharge. J Physics D: Appl. Physics, 27:2136
27. Arora R, Mosch W, (2011) High Voltage and Electrical Insulation Engineering. John Wiley and Sons Inc., New Jersey

Chapter 8
Oblique Newtonian Fluid Flow with Heat Transfer Towards a Stretching Sheet

F. Labropulu and A. Ghaffar

Abstract Oblique stagnation point flow and heat transfer towards a stretching sheet of a viscous fluid is investigated. The governing equations are transformed to a system of ordinary differential equations and then solved numerically for various values of the parameters. It is observed that the dual solution exists for velocity and temperature for certain values of velocity ratio parameter.

Keywords Oblique · Stagnation point · Heat transfer · Stretching sheet

8.1 Introduction

The viscous fluid flow over a stretching sheet is important because of its practical application in engineering processes and in different industries such as glass-fibred production, paper production, wire drawing and extraction of polymer sheet and so on [1]. A closed form solution for steady, two dimensional stretching sheet was found by Crane [2] where the velocity varies linearly with the distance from a fixed point. Following Crane's work many researchers such as Gupta and Gupta [3], Brady and Acrivos [4], Wang [5–7] and Usha and Sridharan [8] considered the case of a stretching sheet in their work. Moreover, stagnation point flow over stretching sheet was investigated by Mahapatra and Gupta [9], Ishak et al. [10], Layek et al. [11] and Nadeem et al. [12]. There exist also a very interesting series of papers by Liao [13, 14], Xu and Liao [15], and Tan et al. [16] on dual solutions of boundary layer flows over a stretching surface. Lok et al. [17] investigated steady flow of a viscous fluid impinging at some angle of incidence on stretching sheet and found that the free stream obliqueness is the shift of the stagnation point towards the incoming flow and it depends on the inclination angle, while Stuart [18], Tamada [19], Dorrepaal [20, 21] and Labropulu et al. [22] also contributed to oblique stagnation point flow. A very good analysis of the oblique stagnation point flow can be found in the book by Pozrikidis [23]. Also, Blyth and Pozrikidis [24] and Tooke and Blyth [25] have presented an interesting analysis of oblique stagnation point at a plane wall. Their analysis shows that oblique flow consists of orthogonal stagnation

F. Labropulu (✉) · A. Ghaffar
Luther College—Mathematics, University of Regina, Regina, SK S4S 0A2, Canada
e-mail: fotini.labropulu@uregina.ca

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_8,

point flow to which is added a shear flow whose vorticity is fixed at infinity. Study of the oblique stagnation point flow of a viscous fluid towards a stretching gives a different perspective to the behaviour of stagnation point flow.

The objective of the present study is to analyze the oblique stagnation point flow of a viscous fluid towards a stretching sheet with heat transfer.

8.2 Basic Equations

Consider the steady oblique stagnation point flow of a viscous fluid towards a stretching sheet. Following Tooke and Blyth [25], we assume that $\bar{u} \sim \bar{u}_e(\bar{x}, \bar{y})$ and $\bar{v} \sim \bar{v}_e(\bar{y})$ have the form $\bar{u}_e = a\bar{x} + b(\bar{y} - \bar{\beta})$, $\bar{v}_e = -a(\bar{y} - \bar{\alpha})$ where $a > 0, b > 0$, $\bar{\alpha}$ and $\bar{\beta}$ are dimensional constants. It is also assumed that the surface is stretched in its own plane with velocity $u_w(\bar{x}) = c\bar{x}$, where c is a constant and that the plate has a constant temperature $T_w(\bar{x})$, while the uniform temperature of the ambient fluid is T_∞, where $T_w(\bar{x}) > T_\infty$. Under these assumptions the steady, two-dimensional, forced convection flow of a viscous fluid can be written as

$$\frac{\partial \bar{u}}{\partial \bar{x}} + \frac{\partial \bar{v}}{\partial \bar{y}} = 0 \tag{8.1}$$

$$\bar{u}\frac{\partial \bar{u}}{\partial \bar{x}} + \bar{v}\frac{\partial \bar{u}}{\partial \bar{y}} = -\frac{1}{\rho}\frac{\partial \bar{p}}{\partial \bar{x}} + \upsilon \bar{\nabla}^2 \bar{u} \tag{8.2}$$

$$\bar{u}\frac{\partial \bar{v}}{\partial \bar{x}} + \bar{v}\frac{\partial \bar{v}}{\partial \bar{y}} = -\frac{1}{\rho}\frac{\partial \bar{p}}{\partial \bar{y}} + \upsilon \bar{\nabla}^2 \bar{v} \tag{8.3}$$

$$\bar{u}\frac{\partial \bar{T}}{\partial \bar{x}} + \bar{v}\frac{\partial \bar{T}}{\partial \bar{y}} = \alpha^* \bar{\nabla}^2 \bar{T} \tag{8.4}$$

where $\bar{u}$ and $\bar{v}$ are the velocity components along the $\bar{x}-$ and $\bar{y}-$ axes, respectively, $\bar{T}$ is the fluid temperature, $\bar{p}$ is the pressure, ρ is the density, υ is the kinematic viscosity, α^* is the thermal diffusivity of the fluid $\bar{\nabla}^2$ is the Laplacian in Cartesian coordinates $(\bar{x}, \bar{y})$. Equations (8.1) to (8.4) will be solved subject to the following boundary conditions.

$$\begin{aligned}
&\bar{v} = 0, \quad \bar{u} = u_w(\bar{x}) = c\bar{x} \quad \text{at} \quad \bar{y} = 0\\
&\bar{T} = T_w(\bar{x}) = T_\infty + c\bar{x} \quad \text{at} \quad \bar{y} = 0\\
&\bar{u}_e = a\bar{x} + b(\bar{y} - \bar{\beta}) \text{ as } \quad \bar{y} \to \infty\\
&\bar{v}_e = -a(\bar{y} - \bar{\alpha}) \text{ as } \quad \bar{y} \to \infty\\
&T = T_\infty, \quad \bar{p} = p_e \quad \text{as} \quad \bar{y} \to \infty
\end{aligned} \tag{8.5}$$

Introducing the non-dimensional variables

$$x = (a/\nu)^{1/2}\,\bar{x}, \quad y = (a/\nu)^{1/2}\,\bar{y}$$
$$u = \bar{u}/(a\nu)^{1/2}, \quad v = \bar{v}/(a\nu)^{1/2}$$
$$p = \bar{p}/(\rho a\nu),\ T = \frac{\bar{T} - T_\infty}{T_w - T_\infty} \tag{8.6}$$

and substituting them into Eqs. (8.1) to (8.4) yield

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{8.7}$$

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{\partial p}{\partial y} + \nabla^2 u \tag{8.8}$$

$$u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{\partial p}{\partial y} + \nabla^2 v \tag{8.9}$$

$$u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} = \frac{1}{\mathrm{Pr}}\nabla^2 T \tag{8.10}$$

The boundary conditions (8.5) become

$$v = 0, \quad u = u_w(x) = \varepsilon x, \quad T = x \quad \text{at} \quad y = 0$$
$$u_e = x + \gamma(y - \beta), \quad v_e = -(y - \alpha) \quad \text{as} \quad y \to \infty$$
$$T = 0, \quad p = p_e \quad \text{as} \quad y \to \infty \tag{8.11}$$

where $\varepsilon = c/a$ is the constant velocity ratio parameter corresponding to the oblique stagnation point flow towards a stretching surface, $\gamma = b/a$ represents the shear in the free stream and Pr is the Prandtl number.

Using Eqs. (8.8) and (8.9), and boundary conditions (8.11), the non-dimensional pressure $p = p_e$ of the inviscid or far flow can be expressed as

$$\mathrm{p_e} = -\frac{1}{2}(x^2 + y^2) + \gamma(\beta - \alpha)x + \alpha y + \mathrm{Constant} \tag{8.12}$$

The physical quantities of interest are the skin friction and the local heat flux from the flat plate which can be written in dimensional form as

$$\tau_w = \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)_{y=0} \tag{8.13}$$

$$q_w = -\left(\frac{\partial T}{\partial y}\right)_{y=0} \tag{8.14}$$

We seek solutions of Eqs. (8.7) to (8.10) of the form

$$\begin{aligned} u &= x\,F'(y) + \gamma\,G'(y), \quad v = -F(y) \\ T &= x\,\theta_1(y) + \gamma\,\theta_2(y) \end{aligned} \tag{8.15}$$

where the functions $F(y)$ and $G(y)$ are referring as the normal and tangential components of the flow, respectively, and prime denotes differentiation with respect to y. Substituting Eq. (8.15) into Eqs. (8.8) to (8.10) and eliminating the pressure p using $p_{xy} = p_{yx}$, we obtain the following differential equations after one integration

$$F''' + FF'' - F'^2 + 1 = 0 \tag{8.16}$$

$$G''' + FG'' - F'G' = \beta - \alpha \tag{8.17}$$

$$\theta_2'' + \Pr(F\theta_1' - F'\theta_1) = 0 \tag{8.18}$$

$$\theta_2'' + \Pr(F\theta_2' - G'\theta_1) = 0 \tag{8.19}$$

The boundary conditions become

$$F(0) = 0, \quad F'(0) = \varepsilon, \quad F'(\infty) = 1 \tag{8.20}$$

$$G(0) = 0, \quad G'(0) = 0, \quad G''(\infty) = 1 \tag{8.21}$$

$$\theta_1(0) = 1, \quad \theta_1(\infty) = 0 \tag{8.22}$$

$$\theta_2(0) = 0, \quad \theta_2(\infty) = 0 \tag{8.23}$$

Employing (8.15), the dimensionless skin friction and the heat transfer can now be written as

$$\tau_w = xF''(0) + \gamma G''(0) \tag{8.24}$$

$$q_w = -[x\,\theta_1'(0) + \gamma\,\theta_2'(0)] \tag{8.25}$$

where the values of $F''(0)$ and $G''(0)$ can be calculated from Eqs. (8.16) and (8.17) with the boundary conditions (8.20) and (8.21) and the values of $\theta_1'(0)$ and $\theta_2'(0)$

Table 8.1 Comparison of the values of $F''(0)$ for various values of ε

ε	Present results	Wang [27]
0	1.2325	1.232588
0.1	1.1465	1.14656
0.2	1.0511	1.05113
0.5	0.7133	0.71330
1	0.0000	0.0000
2	−1.8873	−1.88731
5	−10.2647	−10.26475

can be calculated from Eqs. (8.18) and (8.19) with boundary conditions (8.22) and (8.23) for different values of the parameters involved.

In particular, the dividing streamline $\psi = 0$ and the curve $u = \partial\psi/\partial y = 0$ intersect the wall at the stagnation point where $\tau_w = 0$. Therefore, the location of the stagnation point is given by

$$x_s = -\gamma \frac{G''(0)}{F''(0)} \tag{8.26}$$

if (8.24) is used.

8.3 Results and Discussion

Equation (8.16) subject to the boundary condition (8.20) has been solved numerically for various values of ε using the matlab function bvp4c. A description of this method can be found in [26]. To validate the accuracy of the numerical method a comparison of the obtained results corresponding to the skin friction coefficient $F''(0)$ is made with the results obtained by Wang [27] in Table 8.1 and are found to be in excellent agreement. It can also be seen from Table 8.1 that the numerical values of $F''(0)$ depend entirely on the velocity ratio parameter ε. For $0 \le \varepsilon \le 1$ one can see that as ε is increasing skin friction coefficient is decreasing and a similar result happens for $\varepsilon > 1$.

Having solved Eq. (8.16), Eq. (8.17) subject to the boundary condition (8.21) is solved numerically for various values of β where the values of α are taken from Labropulu et al. [22]. The values of $G''(0)$ for various values of ε and β are shown in Table 8.2. As it can be seen from this table, there is a good agreement between the present values and those obtained by Li et al. [28] when $\varepsilon = 0$. Table 8.2 shows that for a specified value of ε, $G''(0)$ is increasing if β is decreasing. On the other hand, for $0 \le \varepsilon \le 1$ and for given values of β, increase in ε results in an increase of $G''(0)$.

Figure 8.1 shows a comparison of skin friction coefficient $F''(0)$ between the present result and the result obtained by Wang [27] and we can see that both results are showing the same behaviour. Figure 8.2 shows the effect of ε and β on $G''(0)$, where we have similar behaviour as described in Table 8.2.

Table 8.2 Numerical values of $G''(0)$ for various values of ε and β. The values obtained by Li et al. [28] are shown in brackets

ε	$\beta = 5$	$\beta = \alpha$	$\beta = 0$	$\beta = -\alpha$
0	−4.7562 {−4.756}	0.6079 {0.6077}	1.4065 {1.4063}	2.2051 {2.2049}
0.1	−4.7177	0.6478	1.4466	2.2454
0.2	−4.6819	0.6875	1.4869	2.2864
0.5	−4.5866	0.8056	1.6084	2.4111
1	−4.4546	1.0000	1.8120	2.6241

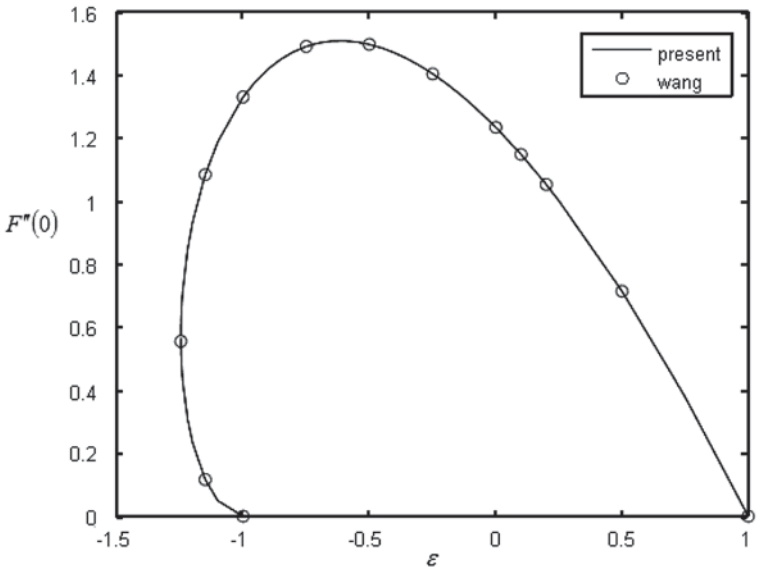


Fig. 8.1 Skin friction coefficient $F''(0)$ for several values of ε

Finally, Eqs. (8.18) and (8.19) subject to the boundary conditions (8.22) and (8.23) have been solved for various values of the Prandtl number Pr. The numerical values of $-\theta_1'(0)$, $-\theta_2'(0)$ and $\theta_1(y)$, $\theta_2(y)$ are plotted in Fig. 8.3–8.6 respectively. Figure 8.3 shows that as Prandtl number is increasing the convective heat transfer increases accordingly. Figure 8.4 depicts the variation of $\theta_1(y)$ for various values of ε when $\mathrm{Pr} = 0.5$ and shows that the temperature function $\theta_1(y)$ is increasing as ε is increasing. The variation of $-\theta_2'(0)$ for various values of Pr when $\beta = \alpha$ is given in Fig. 8.5 and it can be seen that an increase in Pr results in an increase of $-\theta_2'(0)$. Figure 8.6 is shown the variation of $\theta_2(y)$ for various values of ε when $\mathrm{Pr} = 0.5$ and $\beta = \alpha$.

Figures 8.7 and 8.8 depict the streamline patterns (Fig. 8.9).

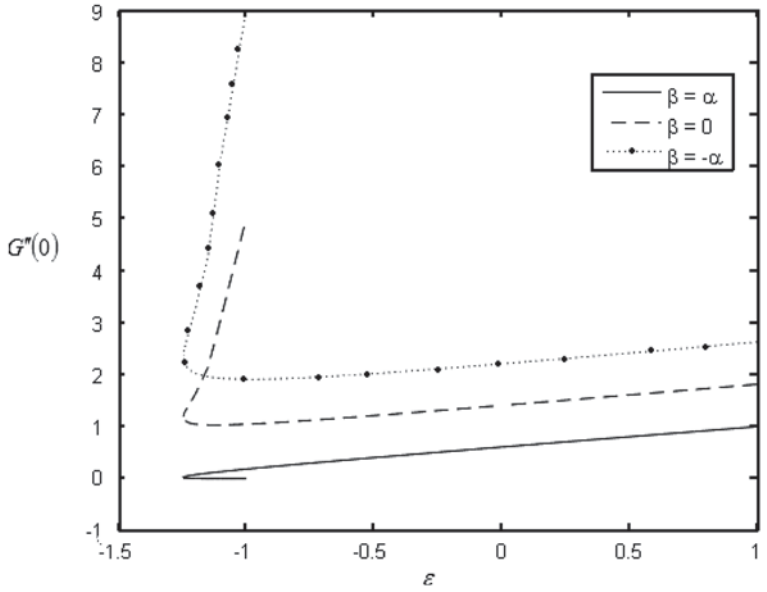


Fig. 8.2 Variation of $G''(0)$ for various values of β and ε

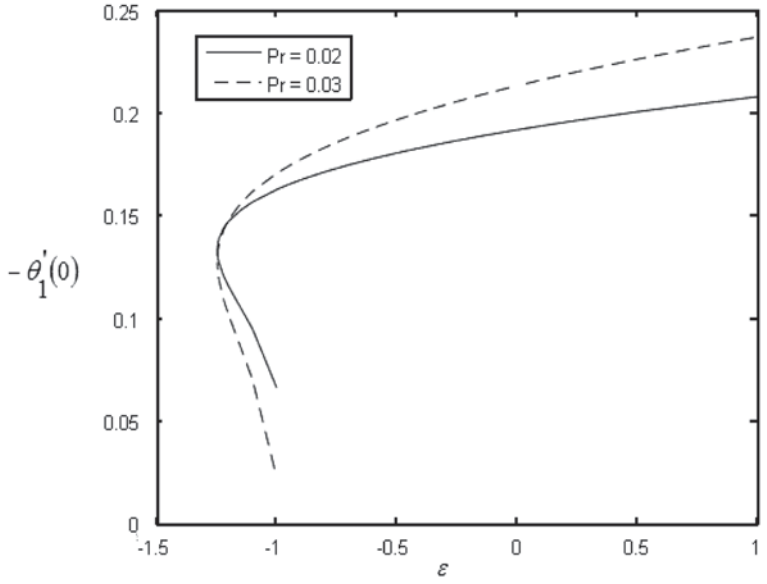


Fig. 8.3 Variation of $-\theta_1'(0)$ for various values of ε and Pr

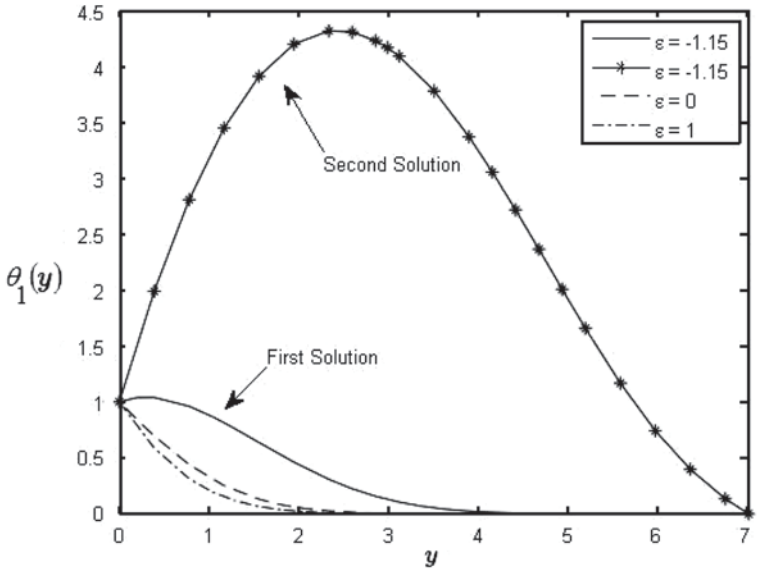


Fig. 8.4 Variation of $\theta_1(y)$ for various values of ε and Pr = 0.5

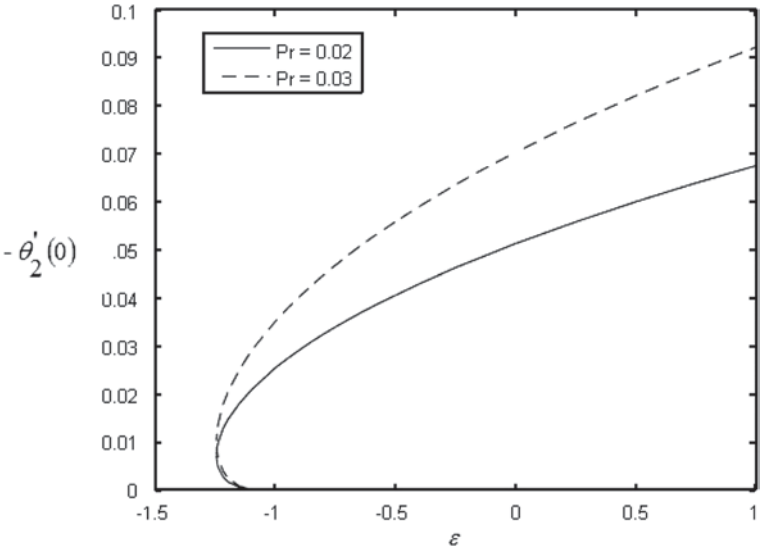


Fig. 8.5 Variation of $-\theta_2'(0)$ for various values of ε, Pr and $\beta = \alpha$

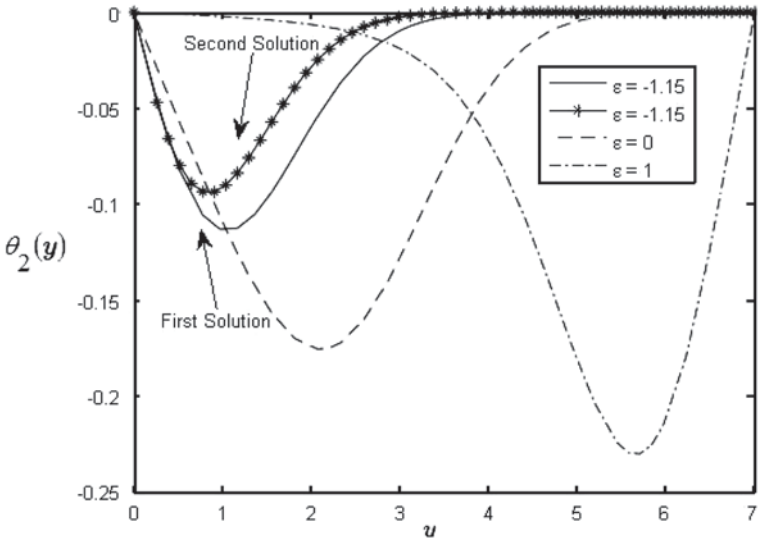


Fig. 8.6 Variation of $\theta_2(y)$ for various values of ε, $\Pr = 0.5$ and $\beta = \alpha$

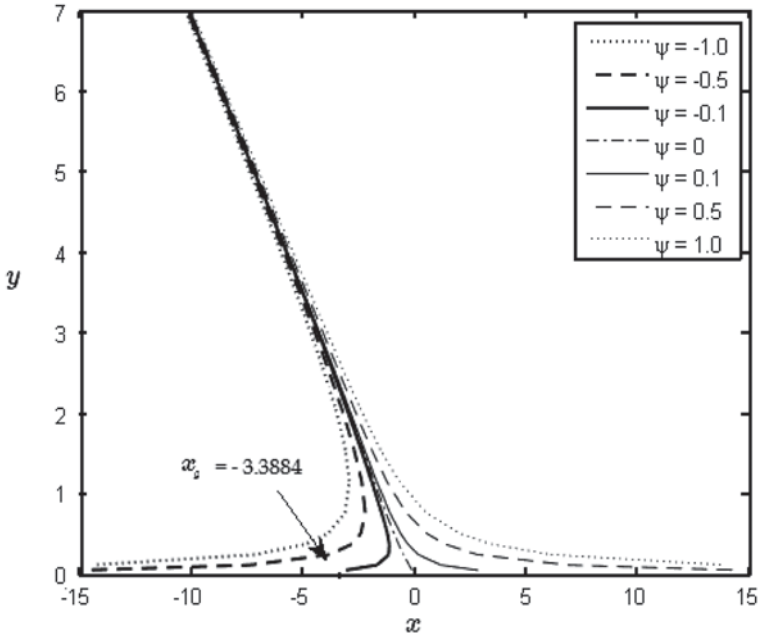


Fig. 8.7 Streamlines pattern for oblique flow when $\varepsilon = 0.5, \beta = \alpha$ and $\gamma = 3$

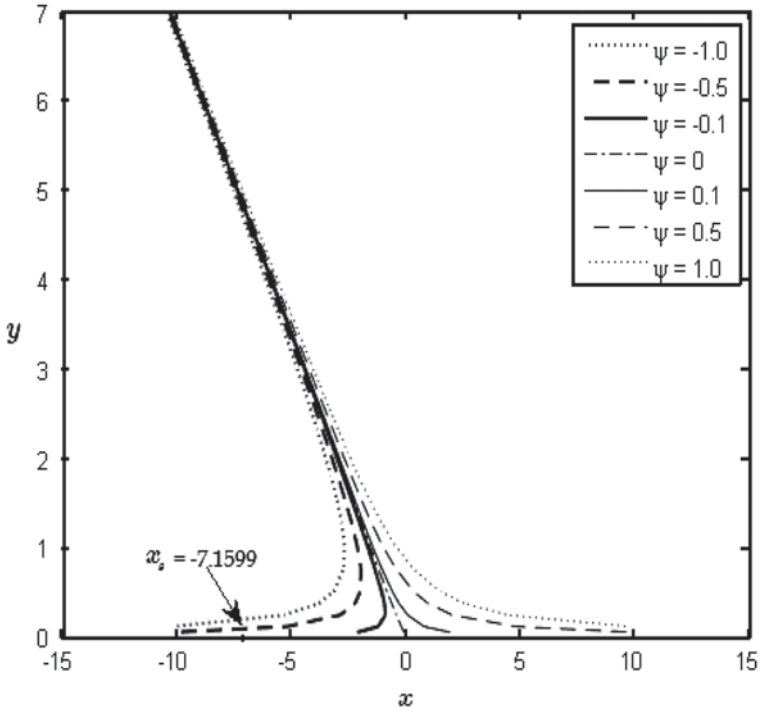


Fig. 8.8 Streamlines pattern for oblique flow when $\varepsilon = 0.75, \beta = \alpha$ and $\gamma = 3$

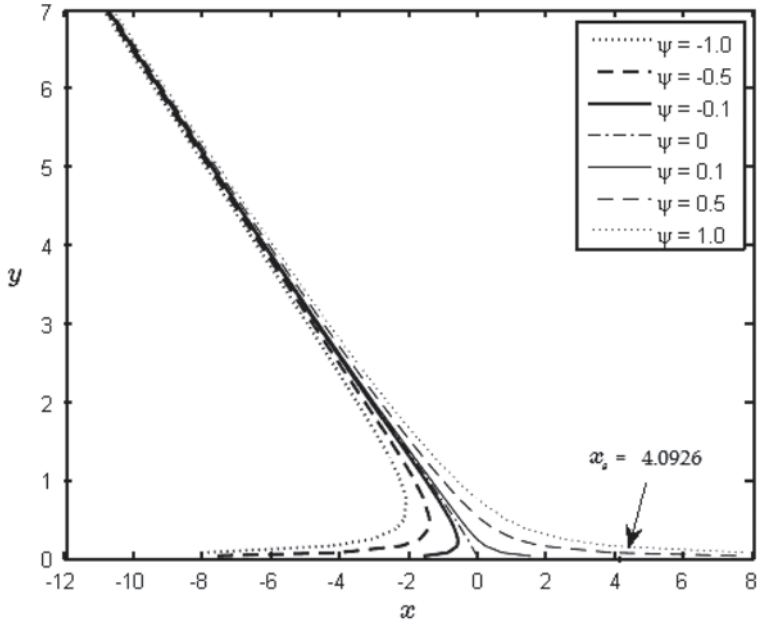


Fig. 8.9 Streamlines pattern for oblique flow when $\varepsilon = 1.5, \beta = \alpha$ and $\gamma = 3$

References

1. Nazar R, Amin N, Pop I (2004) Unsteady boundary layer flow due to a stretching surface in a rotating fluid. Mech. Res. Communications 31: 121–128
2. Crane LJ (1970) Flow past a stretching plate. Z. Angew. Math. Phys. 21: 645–647
3. Gupta PS, Gupta AS (1977) Heat and mass transfer on a stretching sheet with suction and blowing. Can. J. Chem. Eng. 55: 744–746
4. Brady JF, Acrivos A (1981) Steady flow in a channel or tube with accelerating surfaces velocity. An exact solution to the Navier-Stokes equations with reverse flow. J. Fluid Mech. 112: 127–150
5. Wang CY (1984) The three dimensional flow due to a stretching flat surface. Phys. Fluids 27: 1915–1917
6. Wang CY (1988) Fluid flow due to a stretching cylinder. Phys. Fluids 31: 466–468
7. Wang CY (1990) Liquid film on an unsteady stretching sheet. Quart. Appl. Math. 48: 601–610
8. Usha R, Sridharan R (1995) The axisymmetrical motion of a liquid film on an unsteady stretching surface. J. Fluids Eng. 117: 81–85
9. Mahapatra TR, Gupta AS (2002) Heat transfer in stagnation-point flow towards a stretching sheet. Heat Mass Transfer 38: 517–521
10. Ishak A, Nazar R, Pop I (2006) Mixed convection boundary layers in the stagnation-point flow toward a stretching vertical sheet. Meccanica 41: 509–518
11. Layek GC, Mukhopadhyay S, Samad SA (2007) Heat and Mass Transfer analysis for boundary layer stagnation-point flow towards a heated porous stretching sheet with heat absorption/generation and suction/blowing. Int. Commun. Heat and Mass Transfer 34: 347–356
12. Nadeem S, Hussain A, Khan M (2010) HAM solutions for boundary layer flow in the region of the stagnation point towards a stretching sheet. Commun. Nonlinear Sci. Numer. Simul. 15: 475–481
13. Liao S (2005) A new branch of solutions of boundary-layer flows over an impermeable stretched surface. Int. J. Heat Mass Transfer 48: 2529–2539
14. Liao S (2007) A new branch of solutions of boundary-layer flows over a permeable stretching plate. Int. J. Non-Lin Mech 42: 819–830
15. Xu H, Liao SJ (2008) Dual solutions of boundary layer flow over an upstream moving plate. Comm. Nonlinear Sci. Numer. Simulation 13: 350–358
16. Tan Y, You XC, Xu H, Liao SJ (2008) A new branch of the temperature distribution of boundary-layer flows over an impermeable stretching plate. Heat Mass Transfer 44: 501–504
17. Lok YY, Amin N, Pop I (2006) Non-orthogonal stagnation-point flow towards a stretching sheet. Int. J. Non-Lin Mech 41: 622–627
18. Stuart JT (1959) The viscous flow near a stagnation point when the external flow has uniform velocity. J. Aerospace Sci 26: 124–125
19. Tamada KJ (1979) Two-dimensional stagnation point flow impinging obliquely on a plane wall. J. Phys. Soc. Jpn. 46: 310–311
20. Dorrepaal JM (1986) An exact solution of the Navier-Stokes equation which describes non-orthogonal stagnation-point flow in two dimension. J. Fluid Mech. 163: 141–147
21. Dorrepaal JM (2000) Is two-dimensional oblique stagnation-point flow unique? Can. Appl. Math. Q 8: 61–66
22. Labropulu F, Dorrepaal JM and Chandna OP (1996) Oblique flow impinging on a wall with suction or blowing. Acta Mech. 115: 15–25
23. Pozrikidis C (1997) Introduction to Theoretical and Computational Fluid Dynamics, Oxford University Press, Oxford
24. Blyth MG, Pozrikidis C (2005) Stagnation-point flow against a liquid film on a plane wall, Acta Mech. 180: 203–219
25. Tooke RM, Blyth MG (2008) A note on oblique stagnation-point flow. Phys. Fluids 20: 033101-1-3
26. Shampine LF (2003) Singular boundary value problems for ODEs. Appl. Math and Comp. 138: 99–112
27. Wang CY (2008) Stagnation flow towards a shrinking sheet. Int. J. Nonlinear Mech. 43: 377–382
28. Li D, Labropulu F, Pop I (2009) Oblique Stagnation-point flow of a viscoe lastic fluid with heat transfer. Int. J. Non-Lin Mech 44: 1024–1030

Chapter 9
Double Allee Effects on Prey in a Modified Rosenzweig-MacArthur Predator-Prey Model

Eduardo González-Olivares and Jaime Huincahue-Arcos

Abstract In this work, a modified Rosenzweig-MacArthur predator-prey model is analyzed, which is a particular Gause type model, considering two Allee effect affecting the prey population.

This phenomenon may be expressed by different mathematical expressions; with the form here used, the existence of one limit cycle surrounding a positive equilibrium point is proved.

Conditions to the existence of equilibrium points and their local stability are established; moreover, the existence of a separatrix curve dividing the behavior of trajectories which can have different *ω-limit* sets.

Some simulations reinforced our results are given and the ecological consequences are discussed.

Keywords Predator-prey model · Functional response · Allee effect · Stability · Bifurcation · Limit cycle

9.1 Introduction

In current theory of predator-prey dynamics and as consequences of the advancement of the ecological knowledge due to theoretical, empirical, and observational research, more elements are recognized as essential to the phenomenon of predation [27], being incorporated to the study of more complex non-linear mathematical models.

E. González-Olivares (✉) · J. Huincahue-Arcos
Grupo de Ecología Matemática, Instituto de Matemáticas,
Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
e-mail: ejgonzal@ucv.cl

J. Huincahue-Arcos
e-mail: jaime.huincahue@upla.cl

J. Huincahue-Arcos
Departamento de Matemáticas, Universidad de Playa Ancha,
Valparaíso, Chile
e-mail: jaime.huincahue@upla.cl

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_9,

In this work, a Gause-type predator-prey model [16] derived from the reasonably realistic and well-known Rosenzweig-MacArthur model [27] is analyzed, incorporating the Allee effect [13, 26] on the prey growth equation also called depensation in Fisheries Sciences [10, 23].

Any mechanism leading to a positive relationship between a component of individual fitness and the number or density of conspecifics is named as a mechanism of the Allee effect [4], i.e., an Allee effect occurs in populations when individuals suffer a decrease in fitness at low densities [26].

Many ecological mechanisms producing Allee effects are known [25] and distinct causes may generate this phenomenon (Table 1 in [5] or Table 2.1 in [13]). Recent ecological research suggests the possibility that two or more Allee effects can be generated by mechanisms acting simultaneously on a single population (See Table 2 in [5]). The combined influence of some of these phenomena is known as *multiple Allee effect* [1, 5, 13].

The mathematical formalization of the Allee effect are varied [6, 12, 28], but it is possible to prove that most of them are topologically equivalent [18]. However, some of these forms may produce a change in the number of limit cycles through Hopf bifurcation surrounding a positive equilibrium point in predator-prey models [15, 20].

Many algebraic forms can be employed to describe the Allee effect [6, 12, 25, 31] but it is possible to prove that many of them are topologically equivalent [18]. One of this equations is given by

$$\frac{dx}{dt} = rx\left(1-\frac{x}{K}\right)\left(1-\frac{m+n}{x+n}\right) \tag{9.1}$$

where r scales the prey growth rate, K is the environmental carrying capacity, m is the Allee threshold, and the auxiliary parameter n with $n>0$ and $m>-n$, [6, 7, 28], affecting the overall shape of the per-capita growth curve of the prey.

We affirm that Eq. (9.1) describes double Allee effects, expressed once in the factor $m(x)=x-m$, similarly as in the most usual equation representing Allee effect [3, 12]; a second time is given by the term $r(x)=\dfrac{rx}{x+n}$ [31], which can be interpreted as an approximation of a population dynamics where the differences between fertile and non-fertile are not explicitly modelled. Then, we can assume this factor indicates the impact of the Allee effect due to the non-fertile population n [2].

As predator-prey interactions are inherently prone to oscillations [27], it is therefore obvious investigate the Allee effect as a potential mechanism for the creation of population cycles and their related limit cycles from of mathematical point of view [3, 12, 29].

An important objective in these works will be to determine the quantity of limit cycles (trajectories closed and isolated) of this class of non-linear differential equation system associated with the modified Rosenzweig-MacArthur model. We consider that this issue is a good criterion to classify these models, but we not consider this issue in our analysis.

Conditions that guarantee the uniqueness of a limit cycle [21], the global stability of the unique positive equilibrium in predator–prey systems, or non-existence of limit cycles [30], has been extensively studied over the last decades starting with the work by Cheng [8]; results on the existence and uniqueness of limit cycles have been obtained in some papers [8, 22], which can be used to explain many real world oscillatory phenomena in nature [11, 21, 30].

This paper is organized as follows: In Sect. 9.2, we present the model and a topologically equivalent is obtained; in Sect. 9.3, the main properties of this model are presented. In Sect. 9.4, some simulations for verify our results are given. Ecological consequences and a comparative study of the mathematical results are given in Sect. 9.5.

9.2 The Model

Considering the double Allee effect on prey described by (9.1) in the Rosenzweig-MacArthur model [27], the autonomous nonlinear bidimensional differential equation system of Kolmogorov type [16] is given by:

$$X_\sigma : \begin{cases} \dfrac{dx}{dt} = \dfrac{rx}{x+n}\left(1-\dfrac{x}{K}\right)(x-m) - \dfrac{qx}{x+a}y \\ \\ \dfrac{dy}{dt} = \left(\dfrac{px}{x+a} - c\right)y \end{cases} \tag{9.2}$$

where $x=x(t)$ and $y=y(t)$ indicate the prey and predator population sizes, respectively for $t \geq 0$ (number of individuals, density or biomass). The parameters are all positives, i. e. $\sigma = (r, n, K, q, a, p, c, m) \in \mathbb{R}_+^7 \times \mathbb{R}$, with $a<K$ and $-K<m<K$, having the following biological meanings:

r	is the intrinsic growth rate or biotic potential of the prey;
K	is the prey environmental carrying capacity;
$m>0$	is the minimum of viable population (threshold of Allee effect);
n	is the population size of sterile individuals on prey population;
q	is the maximum number of prey that necessary can be eaten by a predator in each time unit;
a	is the amount of prey needed to achieve one-half of q;
p	is the coefficient of biomass conversion, and
c	is the natural death rate of predators in absence of prey.

System (9.2) is defined in $\Omega = \{(x, y) \in \mathbb{R}^2 / x \geq 0,\ y \geq 0\}$.

The analysis must be made separately for the strong Allee effect ($m>0$) and weak Allee effect ($m \leq 0$), due the number of limit cycles can change with respect to this parameter [20]; in this work we consider only $m>0$.

The results will be compared with the Rosenzweig-MacArthur model in which the Allee effect is absent, and with the model studied in [19, 24], where the Allee effect is described by a simpler form, which is topologically equivalent to that used in this work [18].

9.2.1 Topologically Equivalent System

In order to simplify the calculus, we follow the methodology used in [17, 19, 20], making a reparameterization and a time rescaling of system (9.2), given by the function $\varphi : \overline{\Omega} \times \mathbb{R} \to \Omega \times \mathbb{R}$, defined as

$$\varphi(u,v,\tau) = \left(Ku, \frac{rK}{q} v, \frac{r}{\left(u + \frac{n}{K}\right)\left(u + \frac{a}{K}\right)} \tau \right) = (x,y,t)$$

with $\overline{\Omega} = \left\{ (u,v) \in \mathbb{R}^2 / u \geq 0, v \geq 0 \right\}$. As

$$\det D\varphi(u,v,\tau) = \frac{r^2 K^2}{u\left(u + \frac{n}{K}\right)\left(u + \frac{a}{K}\right)} > 0.$$

Then φ is a diffeomorphism preserving the orientation of time [9, 14]; the vector field X_μ is topologically equivalent to the vector field $Y_\eta = \varphi^\circ X_\mu$. It take the form $Y_\eta = P(u,v)\frac{\partial}{\partial u} + Q(u,v)\frac{\partial}{\partial v}$ and the associated second order differential equations system is

$$Y_\eta : \begin{cases} \dfrac{du}{d\tau} = u(1-u)(u-M)(u+A) - u(u+N)v \\ \\ \dfrac{dv}{d\tau} = B(u+N)(u-C)v \end{cases} \tag{9.3}$$

with $\eta = (B,C,A,N,M) \in \mathbb{R}^2_+ \times \left]0,1\right[^2 \times \left]-1,1\right[$, where $B = \frac{1}{r}(p-c)$, $C = \frac{ac}{K(p-c)}$, $A = \frac{a}{K}$, $N = \frac{n}{K}$ and $M = \frac{m}{K}$.

Clearly, $B > 0$ if and only if $p > c$, being a necessary condition for predator to survive; system (9.3) has no ecological sense if $B < 0$.

For the strong Allee effect it has $0<M<<1$; so, the equilibria are (0;0), (M;0), (1;0) and (C;L), where $L=\dfrac{(1-C)(C+A)(C-M)}{C+N}$.

The point (C;L) lies in the first quadrant, if and only if, $0<M<C<1$.

The Jacobian matrix of system (9.3) is

$$DY_\eta(u;v)=\begin{pmatrix} DY_\eta(u,v)_{11} & -u(u+N) \\ Bv(N-C+2u) & -B(C-u)(u+N) \end{pmatrix}$$

with $DY_\eta(u;v)_{11}=-4u^3+3(1+M-A)+2(A-M-v+AM)u-(AM+Nv)$

9.3 Main Results

For $0<M<<1$, system (9.3) has the following properties:

Lemma 1. Existence of invariant set

The set $\overline{\Gamma}=\{(u,v)\in\mathbb{R}^2 / 0\le u\le 1, v\ge 0\}$ is a region of positive invariance.

Proof: Since the system (9.3) is of Kolmogorov type [16], the coordinates axis are invariant sets. If $u=1$, then $\dfrac{du}{d\tau}=-v(1+N)<0$. Anything the sign of $\dfrac{dv}{d\tau}=B(1-C)(1+N)v$, the trajectories enter to the set $\overline{\Gamma}$.

Lemma 2. *Boundedness of solutions*. The solutions are bounded.

Proof: We use the Poincaré compactification with the change of variables given by $u=\dfrac{w}{z}$ and $v=\dfrac{1}{z}$; then,

$$Z_\eta=\begin{cases} \dfrac{dz}{d\tau}=-\dfrac{w}{z^3}\big(z(w+zN)(w-zC)\big)- \\ \qquad \dfrac{w}{z^3}(w-z)(w-zM)(w-zA)+z(w+zN) \\ \dfrac{dw}{d\tau}=-\dfrac{1}{z}(w+zN)(w-zC), \end{cases}$$

The equilibrium point (0;0) of vector field Z_η is equivalent to point (0;∞) of system (9.3). Evaluating in (0;0) of vector field Z_η, the zero matrix is obtained. Rescaling the time by the function $\phi:\overline{\Omega}\times\mathbb{R}\to\Omega\times\mathbb{R}$, defined as $\phi(w;z;z^3T)=(w;z;\tau)$, we obtain a new polynomial system given by

$$\widetilde{Z}_\eta=\begin{cases} \dfrac{dz}{d\tau}=-w\big(z(w+zN)(w-zC)\big)- \\ \quad w(w-z)(w-zM)(w-zA)+z(w+zN) \\ \dfrac{dw}{d\tau}=-z^2(w+zN)(w-zC), \end{cases}$$

The Jacobian matrix evaluated in the point (0;0) is $D\widetilde{Z}_\eta(0;0)=\theta_2$. To desingularize the point (0;0), the technique of blowing-up is used [9, 14]. Using time rescaling defined by $\kappa=\frac{1}{I^2}T$ and the directional blowing-up given by $\varphi_w(I;S)=(I;IS)=(w;z)$, we obtain

$$\widehat{Z}_\eta=\begin{cases}\dfrac{dI}{d\kappa}=-I(S+I-ASI-MSI+NS^2)+I^2(S^2\beta-AMS^3-CNS^3)\\ \dfrac{dS}{d\kappa}=S(S+I-SI-ASI-MSI)+S^3(N-AMSI+I\gamma),\end{cases}$$

with $\beta=A-C+M+N+AM$ and $\gamma=A+M+AM$. We obtain again lies in the first quadrant, and a new directional blowing-up is considered, which is given by $\phi_S(E;F)=(E;EF)=(I;S)$. Using the time rescaling defined by $\lambda=\frac{1}{E}\kappa$ we obtain:

$$\overline{\overline{Z}}_\eta=\begin{cases}\dfrac{dE}{d\lambda}=E(F+1)-FE^2(A+M-FN+F\beta E)-F^3E^4(AM+CN)\\ \dfrac{dF}{d\lambda}=2F(F+1)-F^2E(2A+2M-2FN+1)+F^3E^2(\beta+\gamma)\\ -F^4E^3(2AM+CN),\end{cases}$$

After some calculations we obtain

$$D\overline{\overline{Z}}_\eta(0;0)=\begin{pmatrix}1&0\\0&2\end{pmatrix}.$$

Thus, $\det D\overline{\overline{Z}}_\eta(0;0)>0$ and $\mathrm{tr}D\overline{\overline{Z}}_\eta(0;0)>0$; then, (0;0) is a repeller point of vector field $\overline{\overline{Z}}_\eta$. By blowing-down of φ_w and ϕ_S the point (0;0) is a non-hyperbolic repeller of vector fields $\overline{Z}_\eta$ and $\widehat{Z}_\eta$, respectively. This implies that the point (0;∞) of Y_η is a repeller point and solutions of vector field Y_η are bounded.□

9.3.1 Nature of Equilibria Over the Axis

Lemma 3. The equilibrium point (0;0) is a hyperbolic attractor for all parameter values.

Proof: Immediate evaluating the Jacobian matrix at this point, since $\det DY_\eta(0;0)=ABCMN>0$ and $\mathrm{tr}DY_\eta(u;v)=-(AM+BCN)<0$. Therefore, (0;0) is a locally stable point.□

Lemma 4. The equilibrium point $P_M=(M;0)$ is

1. a hyperbolic repeller, if and only if, $M-C>0$,
2. a hyperbolic saddle point, if and only if, $M-C<0$,
3. a non hyperbolic repeller, if and only if, $M-C=0$.

Proof: As

$$\det DY_\eta(M;0)=MB(1-M)(A+M)(M+N)(M-C)$$

and $$\mathrm{tr}DY_\eta(M;0)=B(M-C)(M+N)+M(1-M)(A+M).$$

i. If $M-C>0$, $\det DY_\eta(M;0)>0$ and $\mathrm{tr}DY_\eta(M;0)>0$. Thus, (M;0) is a hyperbolic repeller.
ii. If $M-C<0$, $\det DY_\eta(M;0)<0$; then, (M;0) is a hyperbolic saddle point.
iii. If $M-C=0$; then $(C;L)$ coincides with the point P_2, and $\det DY_\eta(M;0)=0$; using the Central Manifold Theorem [14], we can proved that point $(M;0)$ is a non hyperbolic repeller. □

Lemma 5. The equilibrium point (1;0) is

1. a saddle hyperbolic point, if and only if, $1-C>0$,
2. a hyperbolic saddle point, if and only if, $1-C<0$,
3. a non hyperbolic attractor, if and only if, $1-C=0$.

Proof: We have that

$$\det DY_\eta(1;0)=-B(A+1)(1-M)(1-C)(N+1) \quad \text{and}$$

$$\mathrm{tr}DY_\eta(1;0)=(A+1)(1-M)+B(1-C)(N+1)$$

i. If $1-C>0$, $\det DY_\eta(1;0)<0$; thus (1;0) is a saddle hyperbolic point.
ii. If $1-C<0$, then $\det DY_\eta(1;0)>0$ and $\mathrm{tr}DY_\eta(1;0)<0$; then, (1;0) is a hyperbolic attractor point.
iii. If $1-C=0$; then (C;L) coincides with (1;0), and $\det DY_\eta(1;0)=0$; using the Central Manifold Theorem [14], it follows that the point (1;0) is a non hyperbolic attractor. □

9.3.2 Existence of a Heteroclinic Curve

When the equilibria $(M;0)$ and (1;0) are saddle points, we will demonstrate the existence of a heteroclinic curve for a given condition of parameters.

Theorem 6. Assuming $0<M<C<1$, the equilibria $(M;0)$ and (1;0) are hyperbolic saddle points. Then, for a subset of parameter values there exists a heteroclinic cycle γ_h in the first quadrant containing these equilibria.

Proof: If $(M;0)$ and $(1,0)$ are both saddle points, then their corresponding invariant manifolds $W^s(M;0)$ and $W^u(1;0)$ are all one-dimensional objects. Clearly, the α-*limit* of $W^s(M;0)$ and the ω-*limit* of $W^u(1;0)$ are bounded in the direction of the *v-axis*. Neither the ω-*limit* of $W^u(1;0)$ is on the *u-axis*.

Let u^* be such that $M<u*<1$. Then, there are points $(u^*;v^s) \in W^s(M;0)$ and $(u^*;v^u) \in W^u(1,0)$, with v^s and v^u depending on the parameter values, such that $v^s=s(\eta)$ and $v^u=u(\eta)$.

Since the vector field Y_η is continuous with respect to the parameters values, then the stable manifold $W^s(M;0)$ must intersect the unstable manifold $W^u(1;0)$ for some parameter values. Hence, there exists a point $(u^*;v^*) \in \overline{\Gamma}$ such that $v^* = v_s^* = v_s^*$.

Moreover, by uniqueness of solutions of system (9.3), this intersection must occur along a whole trajectory γ_{1M}, joining the equilibria $(1;0)$ and $(M;0)$. Therefore, the equation $s(\eta)=u(\eta)$ defines a codimension-one submanifold in the parameters space, for which the heteroclinic curve γ_{1M} exists in $\mathbb{R}^2_+$, connecting the points $(1;0)$ and $(M;0)$.

Then, $\gamma_{1M} \subset W^s(M;0) \cap W^u(1;0)$ and it lies entirely on a segment of the u-axis and exists for any parameter value such that $0<M<C<1$.

It follows that a heteroclinic cycle γ_h exists for certain parameter values on the same submanifold. More precisely, $\gamma_h=(1;0) \cup \gamma_{1M} \cup (M;0) \cup \gamma_{M1}$. □

We note that a the existence of a heteroclinic curve joining the points $(1;0)$ and $(M;0)$ is a common property on models with strong Allee effect.

9.3.3 Nature of the Positive Equilibrium Point

In the following we consider $0<M<C<1$. The equilibrium point $(C;L)$ is in the first quadrant and the Jacobian matrix evaluated at point $(C;L)$ is:

$$DY_\eta(C;L) = \begin{pmatrix} (A+C)\mu & -C(C+N) \\ 0 & B(C-M)(1-C)(A+C) \end{pmatrix};$$

$$\text{with} \quad \mu(A,C,M,N) = \frac{C(1-C)(A+2C-M)}{A+C} - \frac{C(C-M)(N+1)}{C+N}$$

$$\text{and} \quad \det DY_\eta(C;L) = BC(C+N)(C-M)(1-C)(A+C) > 0.$$

Let $Q = (\text{tr}DY_\eta(C;L))^2 - 4\det DY_\eta(C;L)$; then,

$$Q = (A+C)^2\mu^2 - 4BC(C+N)(C-M)(1-C)(A+C).$$

If $Q=0$, then $B = \alpha\mu^2$ where $\alpha = \dfrac{A+C}{4C(C+N)(C-M)(1-C)}$.

With the above relations, we can establish the following theorem:

Theorem 7. Let $(u^*;v^s) \in W^s(M;0)$ and $(u^*;v^u) \in W^u(1,0)$.

7.1 Assuming $\upsilon^s > \upsilon^u$, then, $(C;L)$ is

a) a local hyperbolic attractor point, if and only if, $\mu<0$. Moreover,

a. 1 If $B < \alpha\mu^2$, is a focus attractor.

a. 2 If $B > \alpha\mu^2$, is a node attractor.

b) is a hyperbolic repeller point, if and only if, $\mu>0$. Moreover,

b. 1 If $B < \alpha\mu^2$, is a focus repeller, surrounded by a limit cycle.

b. 2 If $B > \alpha\mu^2$, is a node repeller.

c) is a weak focus, at least of order one, if and only if, $\mu=0$.

7.2 Assuming $\upsilon^s<\upsilon^u$; then, $(C;L)$ is a node repeller and (0;0) is globally asymptotically stable.

Proof: It is immediate from the evaluation of the Jacobian matrix.

If $0<M<C<1$, $\det DY_\eta(C;L) > 0$. So, the nature of $(C;L)$ will be determined by $\mathrm{tr}DY_\eta(C;L)$ and its sign is determined by μ.

i) Assuming $\upsilon^s > \upsilon^u$, it has:

If $\mu<0$, the point $(C;L)$ is a hyperbolic attractor, meanwhile if $\mu>0$, the point $(C;L)$ is a hyperbolic repeller.

If $Q<0$, then $B > \alpha\mu^2$ and $(C;L)$ is a node.

If $Q>0$, then $B < \alpha\mu^2$ and $(C;L)$ is a focus.

ii) Assuming $\upsilon^s > \upsilon^u$, by the existence and uniqueness theorem ensures that the *ω-limit* of $W^s(M;0)$ or $W^u(1;0)$ are in $\bar{\Gamma}$. As (0;0), (1;0) are saddle points, all path in $\bar{\Gamma}$ has as its *ω-limit* to (0;0) which is globally asymptotically stable.•

Remark 8. When $\upsilon^s > \upsilon^u$, the stable manifold $W^s(M;0)$, the straight line $u=1$ and the *u-axis* determines a subregion $\bar{\Lambda}$ (see left poster in Fig. 9.1), which is closed and bounded, i.e.,

$$\bar{\Lambda} = \left\{(u,v) \in \bar{\Omega} / M \le u \le 1, 0 \le v \le v^s < v^u\right\}$$

is a compact region and the Poincaré-Bendixson Theorem applies there, assuring the existence of a limit cycle. As the born of this limit cycle is through of the Hopf bifurcation, the largest is obtained when $v^s=v^u$, i.e. when the heteroclinic curve γ_{1M} is reached.

Then, the increase of the diameter of this limit cycle by change of parameters, which will increase until to attain the heteroclinic curve.

Remark 9. To determine the weakness of the focus $(C;L)$, the number of limit cycles bifurcating of a weak (fine) focus must be obtained [9]. The weakness of a focus indicates the number of limit cycles appearing by multiple Hopf bifurcation, i.e., the number of the concentric limit cycles surroumding a weak focus [9].

There exist various methods to establish this number being one of them the calculus of the Lyapunov quantities [9, 14]; however, this task that will not be assumed in this work. In Fig. 9.3 we show the existence of a unique limit cycle reinforced the result obtained in theorem 7b.1 (Fig. 9.2, 9.4, 9.5 and 9.6).

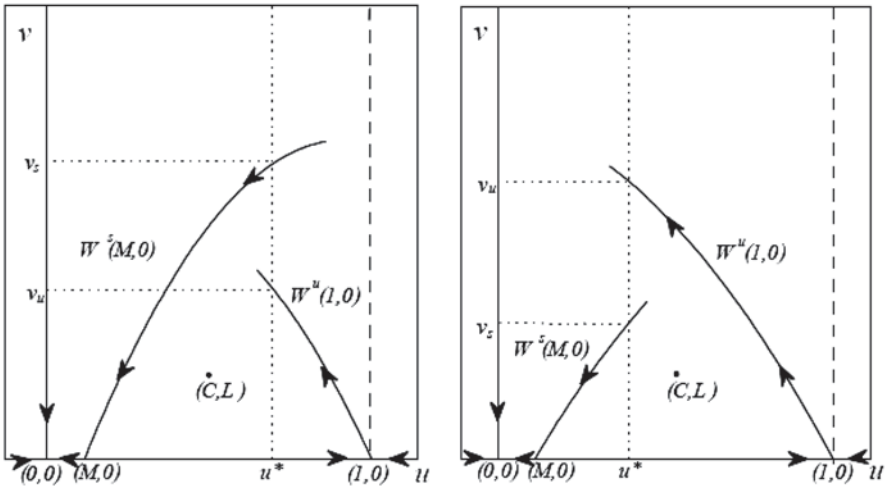


Fig. 9.1 For $0<M<C<1$, $(C;L)$ is the unique positive equilibrium point. The two possible relative positions between the stable manifold $W^s(M;0)$ of the saddle point P_M and the unstable manifold $W^u(1;0)$ of saddle point P_1 are shown. On the left side $v^s<v^u$ and on the right side $v^s>v^u$. Being the vector field Y_η continuous with respect to the parameters values, then the intersection between $W^s(M;0)$ and $W^u(1;0)$ occurs

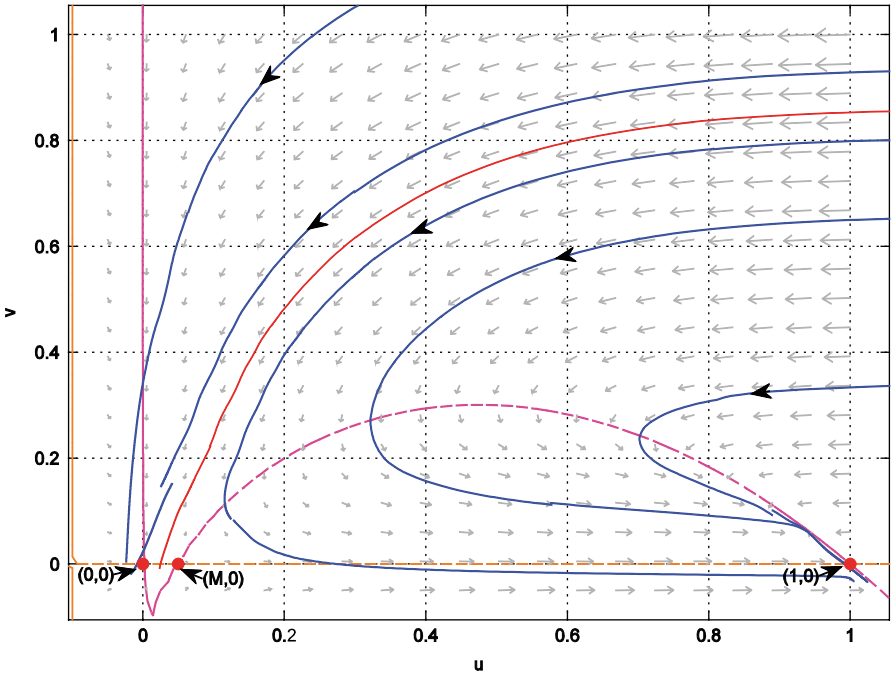


Fig. 9.2 For $A=0.3$, $B=0.2$, $C=1.2$, $M=0.05$ and $N=0.1$; there no exists positive equilibrium point. The points (1;0) and (0;0) are local attractors

9.4 Some Simulations

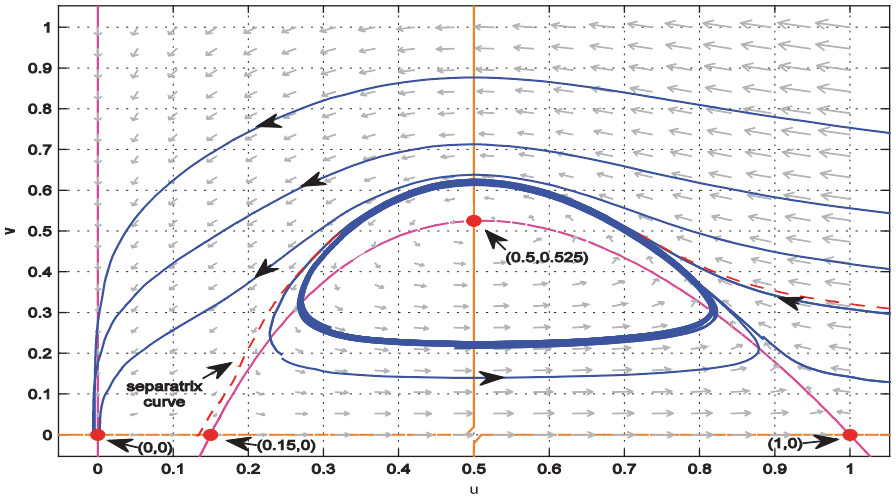


Fig. 9.3 For $A=0.2$, $B=0.5$, $C=0.5$, $M=0.15$ and $N=0.4$. The vector field Y_η has four equilibrium points in the first quadrant; (0;0) is a attractor point; (M;0) and (1;0) are a saddle point and (C;L) is a repeller, surrounded by a stable limit cycle

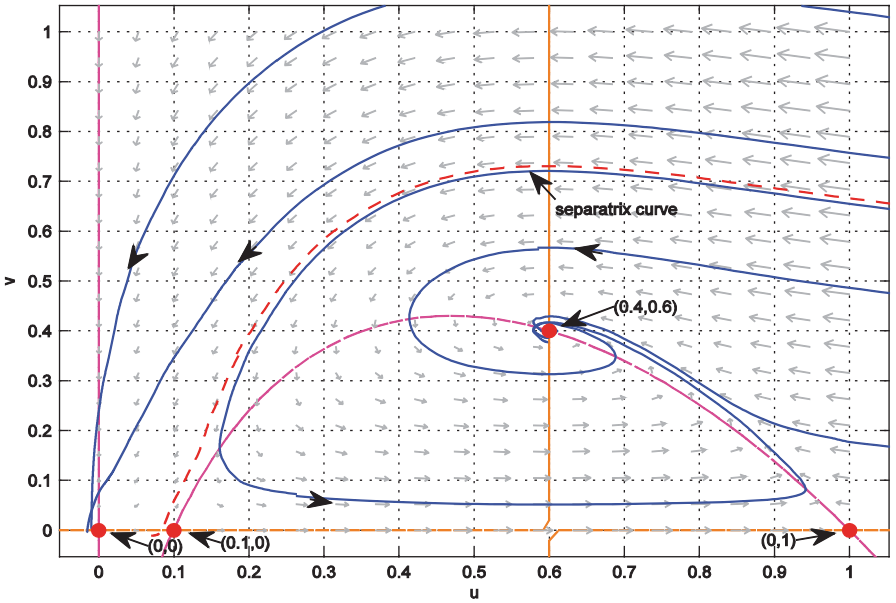


Fig. 9.4 For $A=1$, $B=0{:}5$, $C=0{:}6$, $M=0{:}1$ and $N=0{:}2$. The vector field Y_η has four equilibrium points in the first quadrant; (0;0) is a attractor point, (M;0) and (1;0) are saddle equilibrium points and (C;L) is a node attractor

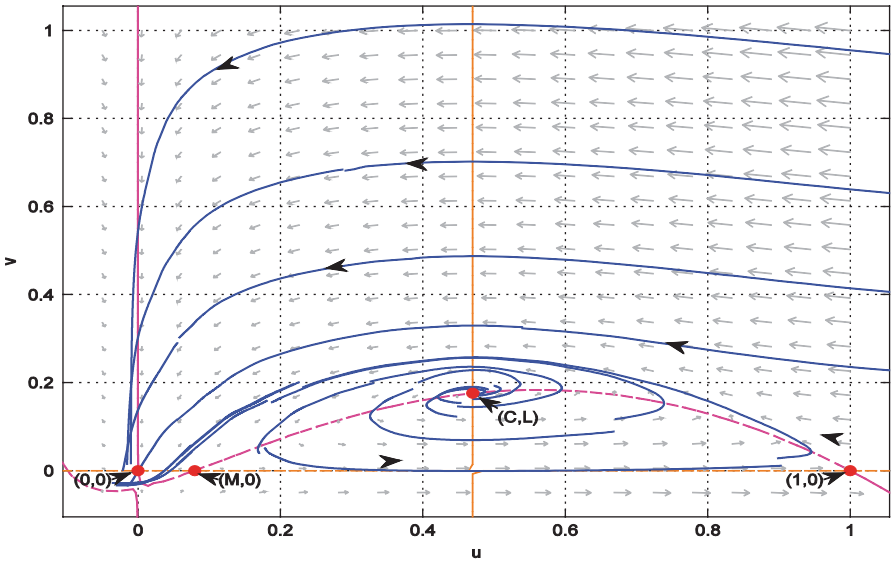


Fig. 9.5 For $A=0.1$, $B=0.3$, $C=0.47$, $M=0.08$ and $N=0.2$; the point $(C;L)$ is repeller focus and $(0;0)$ is globally asymptotically stable. In this case, $v^s<v^u$ for $(u^*;v^s) \in W^s(M;0)$ and $(u^*;v^u) \in W^u(1;0)$

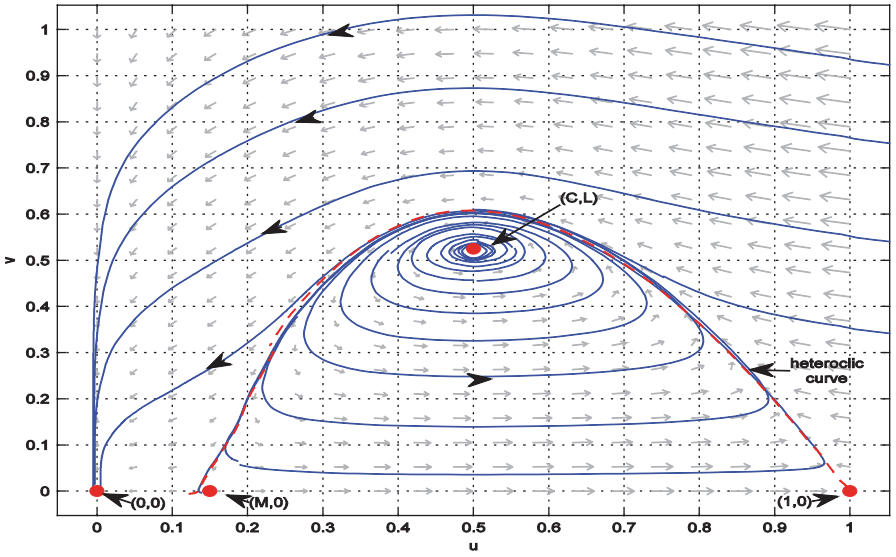


Fig. 9.6 For $A=0.1$, $B=1$, $C=0.25$, $M=0.15$ and $N=0.115$. The vector field Y_η has four equilibrium points in the first quadrant; $(0;0)$ is a attractor point; $(M;0)$ and $(1;0)$ are a saddle point and $(C;L)$ is a repeller, and the stable limit cycle collides with the heteroclinic curve

9.5 Conclusions

The existence of interesting dynamics has been shown, for a modified Rosenzweig-MacArthur model [27], a particular case of a Gause type predator-prey model, considering a double Allee effect on prey [1, 4]. The properties are established using a polynomial differential equations system (9.3) topologically equivalent to original system (9.2).

We proved that the model proposed have multiple stable equilibria for a determined set of parameter values and, therefore, different population behaviors can coexist.

As in all models considering strong Allee effect, in system (9.3) there exists a separatrix curve determined by the unstable manifold of equilibrium point $(m,0)$. Then, there are trajectories near of this separatrix, which can have different ω-*lim* it for the same set of parameter values, showing they are highly sensitive to initial conditions. So, for a fixed set of parameters, the following may happen: extinction of two populations, the coexistence for determined population sizes or oscillations of both populations.

Moreover, there are parameter constraints for which the existence of a interior equilibrium point local asymptotically stable or the existence of at least one stable limit cycle generated by Hopf bifurcation has been proved.

We affirm that Eq. (9.1) can be assumed as a paradigm to represent double Allee effect. In fact, without assuming that the population is divided into age or sex class, it can be considered that $x=x(t)$ represents the size of fertile population and n is the non-fertile population (juvenile or oldest individuals) [2]. Populations with strong Allee effects can go extinct at lower levels of mortality by predation; also, when mortality by predation increases and weaker Allee effects can drive population to extinction.

Although extinction of predator or both species are not interesting outcomes from the point of view of population dynamics, system (9.3) it capable for a complete spectrum of dynamical behaviors that can, in principle, characterize this kind of models.

We think it is important for ecologists to be aware of the kind of bistability described for system (9.3), where two potential attractors can exist: (i) the origin; (ii) a positive equilibrium point or a stable limit cycle.

Acknowledgement The authors thank the members of the Grupo de Ecología Matemática on the Instituto de Matemáticas at the Pontificia Universidad Católica de Valparaíso, for their valuable comments and suggestions. This work is partially financed by Projects Fondecyt No 1120218 and DIEA-PUCV 124.730/2012.

References

1. Angulo E, Roemer GW, Berec L, Gascoigne J, Courchamp F (2007) Double Allee effects and extinction in the island fox. Conservation Biology 21: 1082–1091.
2. Barclay H, Mackauer M, (1980) The sterile insect release method for pest control: a density-dependent model. Environmental Entomology 9: 810–817.
3. Bazykin AD (1998) Nonlinear Dynamics of interacting populations, World Scientific Publishing Co. Pte. Ltd.
4. Berec L (2007) Models of Allee effects and their implications for population and community dynamics, In: Mondaini R (Ed.) Proceedings of the 2007 International Symposium on Mathematical and Computational Biology, E-papers Serviços Editoriais Ltda., pp. 179–207.
5. Berec L, Angulo E, Courchamp F (2007) Multiple Allee effects and population management, Trends in Ecology and Evolution 22: 185–191.
6. Boukal DS, Berec L (2002) Single-species models and the Allee effect: Extinction boundaries, sex ratios and mate encounters. Journal of Theoretical Biology 218: 375–394.
7. Boukal DS, Sabelis MW, Berec L (2007) How predator functional responses and Allee effects in prey affect the paradox of enrichment and population collapses. Theoretical Population Biology 72: 136–147.
8. Cheng KS (1981) Uniqueness of a limit cycle for a predator-prey system, SIAM Journal on Applied Mathematics 12: 541–548.
9. Chicone C (2006) Ordinary differential equations with applications (2nd edition). Texts in Applied Mathematics 34, Springer.
10. Clark CW (2010) Mathematical Bioeconomics: The Mathematics of Conservation (3nd ed). John Wiley and Sons Inc.
11. Coleman CS (1983) Hilbert's 16th. Problem: How Many Cycles? In: Braun M, Coleman CS, Drew D (Eds). Differential Equations Model, Springer Verlag, pp. 279–297.
12. Conway ED, Smoller JA (1986) Global Analysis of a System of Predator-Prey Equations. SIAM Journal on Applied Mathematics 46: 630–642.
13. Courchamp F, Berec L, Gascoigne J (2008) Allee Effects in Ecology and Conservation, Oxford University Press.
14. Dumortier F, Llibre J, Artés JC (2006) Qualitative theory of planar differential systems, Springer.
15. Flores JD, Mena-Lorca J, González-Yañez B, González-Olivares E (2007) Consequences of Depensation in a Smith's Bioeconomic Model for open-access Fishery. In: Mondaini R. (Ed.) Proceedings of the 2006 International Symposium on Mathematical and Computational Biology, E-papers Serviços Editoriais Ltda., pp. 219–232.
16. Freedman HI (1980) Deterministic Mathematical Model in Population Ecology, Marcel Dekker.
17. González-Olivares E, Ramos-Jiliberto R (2003) Dynamics consequences of prey refuges in a simple model system: more prey, fewer predators and enhanced stability. Ecological Modelling 106: 135–146.
18. González-Olivares E, González-Yañez B, Mena-Lorca J, Ramos-Jiliberto R (2007) Modelling the Allee effect: are the different mathematical forms proposed equivalents? In: Mondaini R (Ed.) Proceedings of the 2006 International Symposium on Mathematical and Computational Biology, E-papers Serviços Editoriais Ltda., Río de Janeiro, pp. 53–71.
19. González-Olivares E, Meneses-Alcay H, González-Yañez B, Mena-Lorca J, Rojas-Palma A, Ramos-Jiliberto R (2011) A Gause type predator-prey model with Allee effect on prey: Multiple stability and uniqueness of limit cycle. Nonlinear Analysis: Real World Applications 12: 2931–2942.

20. González-Olivares E, González-Yañez B, Mena-Lorca J, Rojas-Palma A, Flores JD (2011) Consequences of double Allee effect on the number of limit cycles in a predator-prey model. Computers and Mathematics with Applications 62: 3449–3463.
21. Hasík K (2010) On a predator-prey system of Gause type. Journal of Mathematical Biology 60: 59–74.
22. Kuang Y, Freedman HI (1988) Uniqueness of limit cycles in Gause-type models of predator-prey systems. Mathematical Biosciences 88: 67–84.
23. Liermann M, Hilborn R (2001) Depensation: evidence, models and implications. Fish and Fisheries 2: 33–58.
24. Meneses-Alcay H, González-Yañez E (2004) Consequences of the Allee effect on Rosenzweig-MacArthur predator-prey model. In: Mondaini R (ed.) Proceedings of the Third Brazilian Symposium on Mathematical and Computational Biology BIOMAT 2003, E-papers Serviços Editoriais Ltda., Volumen 2 pp. 264–277.
25. Stephens PA, Sutherland WJ (1999) Consequences of the Allee effect for behaviour, ecology and conservation. Trends in Ecology and Evolution 14: 401–405.
26. Stephens PA, Sutherland WJ, Freckleton RP (1999) What is the Allee effect?. Oikos 87: 185–190.
27. Turchin P (2003) Complex population dynamics. A theoretical/empirical synthesis, Monographs in Population Biology 35, Princeton University Press.
28. van Voorn GAK, Hemerik L, Boer MP, Kooi BW (2007) Heteroclinic orbits indicate overexploitation in predator–prey systems with a strong Allee effect. Mathematical Biosciences 209: 451–469.
29. Wang J, Shi J, Wei J (2011). Predator-prey system with strong Allee effect in prey. Journal of Mathematical Biology 62: 291–331.
30. Xiao D, Zhang Z (2003) On the uniqueness and nonexistence of limit cycles for predator-prey systems. Nonlinearity 16: 1185–1201.
31. Zu J, Mimura M (2010) The impact of Allee effect on a predator-prey system with Holling type II functional response. Applied Mathematics and Computation 217: 3542–3556.

Chapter 10
Buckling of Plates on Rotationally and Warping Restrained Supports

V. Piscopo and A. Scamardella

Abstract The paper deals with elastic buckling of plates, having warping and elastically restrained against torsion supports, under uniaxial compression. The minimum energy principle is applied, regarding the isolated plate as part of an infinitely wide stiffened panel, reinforced by longitudinal stiffeners and transverse beams, despite of classical solutions, where two coupled transcendental equations are solved. The displacement field is developed into double sine trigonometric series and the solution convergence, in terms of buckling coefficients, is investigated. Simple design buckling formulas for isolated plate panels, as function of supporting members' torque and warping rigidity ratios, are derived by curve fitting. Finally, several stiffened panels are analysed and the proposed formulas are compared with the relevant results obtained by some FE eigenvalue buckling analyses, carried out by ANSYS.

Keywords Buckling analysis · Energy principle · Elastically restrained plates

10.1 Introduction

Ship structures, constituted by multi bay longitudinally stiffened panels supported by transverse beams, have to be checked for buckling criteria under the combined action of compressive and shear stresses, due to hull girder and local loads. From this point of view the isolated plate panel buckling analysis is generally carried out assuming, on the safety side, the simple support boundary conditions at all edges. Anyway, if the plate boundary conditions differ significantly from the simple support ones, more appropriate restraints may be applied assuming the plate is clamped

V. Piscopo (✉) · A. Scamardella
Department of Science and Technology, The University of Naples "Parthenope",
Centro Direzionale—Isola C4, 80143 Naples, Italy
e-mail: vincenzo.piscopo@uniparthenope.it

A. Scamardella
e-mail: antonio.scamardella@uniparthenope.it

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_10,

at long and/or short sides. Really, these idealized boundary conditions never occur, because longitudinal stiffeners and transverse beams have finite rotational and warping restraints, so that the elementary plate panel buckling analysis may be carried out accounting for the effective amount of supporting members' torsional and warping rigidities. Besides, it is assumed that the supporting member bending rigidities should be so as to avoid stiffened panel's overall buckling occurs before local plate buckling [7], according to the general adopted scantling design.

In the past a lot of authors, such as Lundquist and Stowell [5], Timoshenko and Gere [9], Gerard and Becker [3], Evans [2] and others investigated the buckling behaviour of uniaxially compressed plates, elastically restrained against torsion at long (short) sides and simply supported at short (long) ones, from now on SSLR (SRLS) platings. Anyway, the most extensive study is probably that one carried out by Paik and Thayamballi [6], who derived simple design buckling formulas for isolated plate panels, as function of supporting members' torsional rigidity ratio and panel aspect ratio, by solving two coupled transcendental equations, derived by the imposed boundary conditions. In the following the minimum energy principle is applied to the buckling analysis of uniaxially compressed plates, rotationally and warping restrained at long and short sides, assuming the isolated plate as part of an infinitely long stiffened panel, reinforced by longitudinal ribs and transverse beams, so accounting for the strain energy due to torque and warping rigidities of internal supporting members. The vertical displacement field is modelled as a double sine trigonometric series and new buckling formulas, as function of stiffeners' torque and warping rigidity ratios, are derived. Finally, different stiffened panels are analysed, comparing the new formulas with those ones proposed by Paik and Thayamballi and with the relevant results obtained by some eigenvalue buckling analyses carried out by ANSYS.

10.2 Theoretical Background

Most current practical design guidelines for buckling and ultimate strength of platings are mainly based on boundary conditions in which all edges are simply supported or perfectly clamped. In real platings, idealized edge conditions as simply supported or clamped ones never occur, because of finite rotational restraints. So, for a more advanced design of steel platings against buckling, it is hence important to better understand the relevant buckling strength characteristics, as a function of the torsional rigidity of supporting members along the edges.

In the classical methods the characteristic equation for buckling of platings with elastic restraints along either long or short edges, with other edges simply supported, is generally derived analytically, starting from the well-known deflection equation of platings subjected to combined in-plane loads. The solution indicates the plating deflected form under the corresponding load, which represents equilibrium but unstable position. The buckling strength is then defined by the load at the bifurcation point where beside the plane equilibrium form, a deflected but unstable

form of equilibrium occurs. Anyway, it is normally not an easy task to directly solve for platings with elastically restrained boundary conditions at all edges, or under combined axial and shear stresses, the previously proposed problem, [6]. So it seems that a more flexible technique may be derived.

In the following the more general solution based on multiple stiffened panels is adopted. From this point of view various numerical techniques are available. Hasegawa et al. [4] used the finite strip method, developing the displacement field of a finite element strip, comprised between adjacent supporting members, into a half-wave sine curve in the loaded direction and a cubic polynomial in the transverse one, so defining a displacement field with four degrees of freedom. The computational economy is superior to finite element analysis or finite difference analysis, because of rapid convergence. Furthermore the number of finite strip elements doesn't depend on the number and configuration of stiffeners in an entire multiple stiffened panel, as the location of a finite element strip may be arbitrarily defined. In the following the finite strip method is adopted, developing the displacement field into double sine trigonometric series, in order to derive the characteristic buckling formula for platings with rotationally restrained edges. The proposed method is particularly flexible and may be extended to more complex loading conditions, that cannot be easily solved by classical techniques.

10.3 Buckling Strength Analysis

10.3.1 Plates Rotationally Restrained at Long Edges

Let us consider an isolated plate panel, having dimensions $a \times b$, comprised between two adjacent transverse beams and longitudinal stiffeners, and let us assume x and y axes are taken in the long and short directions respectively, so that the panel aspect ratio α is always greater than one. The panel, simply supported at short edges, is rotationally and warping restrained at long sides by longitudinal stiffeners having the same material properties as the isolated plate panel's ones (SSLR plating). Let us also assume the longitudinal stiffeners' bending rigidity is sufficiently high to avoid overall buckling from occurring before local plate buckling, so that relative supporting members lateral deflection may be neglected, assuming the plate edges remain straight until buckling occurs [8]. As previously said, the isolated plate may be regarded as part of an infinitely wide stiffened panel, reinforced by n_s equally spaced longitudinal stiffeners, having the same geometrical and mechanical properties. The stiffened panel, having dimensions $a \times B$, is loaded by compressive forces acting in x-direction (see Fig. 10.1) and is simply supported at all edges.

The number of panels n_s has been suitably chosen to obtain consistent results in terms of buckling coefficients. The vertical displacement field may be developed into appropriate double sine trigonometric series, as follows:

$$w(x,y) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} w_{m,n} \sin \frac{m\pi x}{a} \sin \frac{n\left(n_s+1\right)\pi y}{B} \tag{10.1}$$

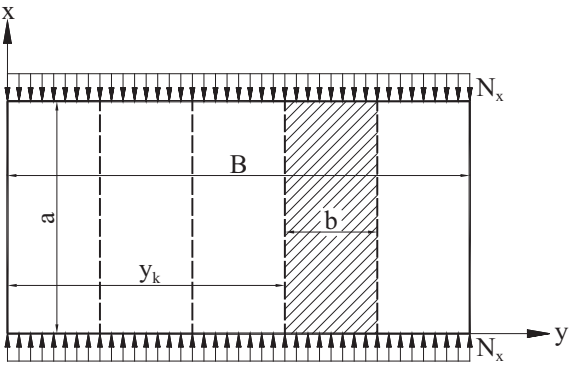


Fig. 10.1 SSLR plate scheme

so that, considering the M and N partial sums of the above trigonometric series, Eq. (10.1) may be so rewritten:

$$w(x,y) = \sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n} \sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b} \tag{10.2}$$

To evaluate the Euler stress at which buckling occurs, the energy method is applied, assuming the stiffened panel undergoes some small lateral bending, consistent with the given boundary conditions. If the in-plane forces work is smaller than the plate and attached stiffeners strain energy, the equilibrium is stable, otherwise it is unstable and buckling occurs. The general equilibrium equation is [10]:

$$\Delta U_p + \sum_{k=1}^{n_s} \Delta U^L_{s,k} = \Delta T_p \tag{10.3}$$

having denoted by ΔU_p ($\Delta U^L_{s,k}$) the strain energy due to plate bending (due to torque and warping of the k-th stiffener) and by ΔT_p the compressive forces work on the plate in x-direction. The strain energy due to plate bending is:

$$\Delta U_p = \frac{D}{2}\int_0^a\int_0^B \left[\left(\frac{\partial^2 w}{\partial x^2}\right)^2 + \left(\frac{\partial^2 w}{\partial y^2}\right)^2 + 2\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}\right] dxdy \tag{10.4}$$

finally becoming:

$$\Delta U_p = \frac{\pi^4 aD}{8b^3}(n_s+1)\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n}^2 \left(\frac{m^2}{\alpha^2}+n^2\right)^2 \tag{10.5}$$

The strain energy of the k-th stiffener, located at $y=kb$ from the edge $y=0$, is the sum of two terms due to torque and warping rigidities respectively, the last one to be accounted for only if stiffeners are restrained against warping:

$$\Delta U_{s,k}^{L} = \frac{1}{2} G J_k^L \int_0^a \left(\left. \frac{\partial \Phi_k^L}{\partial x} \right|_{y=kb} \right)^2 dx + \frac{1}{2} E I_{w,k}^L \int_0^a \left(\left. \frac{\partial^2 \Phi_k^L}{\partial x^2} \right|_{y=kb} \right)^2 dx \tag{10.6}$$

having denoted by $\Phi_k{}^L$ the *k*-th stiffener rotation around its connection to the attached plating, by $J_k{}^L$ and $I_{w,k}{}^L$ the stiffeners' St. Venant and warping moments of inertia. Imposing the deflection angle continuity condition along the junction between the plate and the longitudinal stiffeners, the following congruence condition must be verified:

$$\Phi_k^L(x, kb) = \left. \frac{\partial w}{\partial y} \right|_{y=kb} \qquad \forall k = 1 \ldots n_s \tag{10.7}$$

so that Eq. (10.6) may be simplified as follows:

$$\Delta U_{s,k}^{L} = \frac{\pi^4 G J_k^L}{4ab^2} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{q=1}^{N} m^2 n q w_{m,n} w_{m,q} \cos(n\pi k) \cos(q\pi k) \left(1 + 2 \frac{\pi^2 m^2 (1+\nu) I_{w,k}^L}{J_k^L a^2} \right) \tag{10.8}$$

The compressive forces work is:

$$\Delta T_p = -\frac{N_x}{2} \int_0^a \int_0^B \left(\frac{\partial w}{\partial x} \right)^2 dxdy \tag{10.9}$$

finally becoming:

$$\Delta T_p = -\frac{N_x \pi^2}{8\alpha} (n_s + 1) \sum_{m=1}^{M} \sum_{n=1}^{N} m^2 w_{m,n}^2 \tag{10.10}$$

By the minimum condition, the coefficients of series (10.1) can be determined by solving the following eigenvalue problem:

$$\frac{\partial}{\partial w_{j,k}} \left[\Delta U_p + \sum_{k=1}^{n_s} \Delta U_{s,k}^{L} - \Delta T_p \right] = 0 \tag{10.11}$$

that can be finally rewritten as follows:

$$w_{j,k} \left(\frac{j^2}{\alpha^2} + k^2 \right)^2 + \frac{2}{\alpha^2 (n_s + 1)} \sum_{q=1}^{N} w_{j,q} j^2 k q \sum_{r=1}^{n_s} \cos(k\pi r) \cos(q\pi r) \cdot \left(\mu_L + \psi_L \frac{\pi^2 j^2}{\alpha^2} \right) = -\frac{k_b}{\alpha^2} j^2 w_{j,k} \tag{10.12}$$

with μ_L and ψ_L torque and warping rigidity ratios:

$$\mu_L = \frac{GJ_L}{Db} = 2(1-\nu)\frac{J_L}{J_P} \tag{10.13}$$

$$\psi_L = \frac{EI_w}{Db^3} = (1-\nu^2)\frac{I_\varsigma}{I_{\varsigma-P}}\left(\frac{h_w}{t}\right)^2 \tag{10.14}$$

In Eq. (10.13) the ratio between stiffener J_L and plating J_P St. Venant moments of inertia may be assumed not greater than one, on the safety side, to account for supporting members torsional buckling. In Eq. (10.14) I_w is the stiffener sectorial moment of inertia about its connection to the attached plating, I_ς ($I_{\varsigma-p}$) is the stiffener (plating) moment of inertia respect to its vertical neutral axis, while h_w is the stiffener web height. The equation system (10.12) has been solved by a dedicated program developed in Matlab MathWorks. The number of harmonics in x-direction was taken equal to the number of half-waves into which the isolated plate buckles, while the one in y-direction equal to 299, due to the curvature of the plating buckled shape. Based on curve fitting of a large amount of data, buckling coefficients and Euler stresses may be expressed as follows:

$$\begin{aligned} k_{b-SSLR}^{new} &= \frac{6.984\mu_L^2 + 9.789\mu_L + 6.189}{\mu_L^2 + 1.732\mu_L + 1.545}\cdot\frac{\mu_L + e^{0.0246\psi_L} - 0.4e^{-6.665\psi_L}}{\mu_L + 0.6} \\ \sigma_E &= k_{b-SSLR}^{new}\frac{\pi^2 E}{12(1-\nu^2)}\left(\frac{t}{b}\right)^2 \end{aligned} \tag{10.15}$$

In Fig. 10.2 the buckling coefficient k_{b-SSLR} is plotted versus μ_L, for fixed values of ψ_L, namely 0.00, 0.05, 0.10, 0.20 and 0.30; the sixth curve refers to the formula by Paik & Thayamballi [6], while the last one (SSLC) is the limit value for plates simply supported at short sides and clamped at long ones.

10.3.2 Plates Rotationally Restrained at Short Edges

The buckling of plates under uniaxial compression, simply supported at long edges and elastically restrained at short ones (SRLS plates), may be treated as the previous case. The isolated plate may be regarded as part of an infinitely long stiffened panel, simply supported at all edges and reinforced by n_t equally spaced transverses, having the same mechanical properties as the attached plate's ones (see Fig. 10.3).

The vertical displacement field is developed into appropriate double sine trigonometric series, satisfying the simple support boundary conditions:

$$w(x,y) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} w_{m,n}\sin\frac{m(n_t+1)\pi x}{A}\sin\frac{n\pi y}{b} \tag{10.16}$$

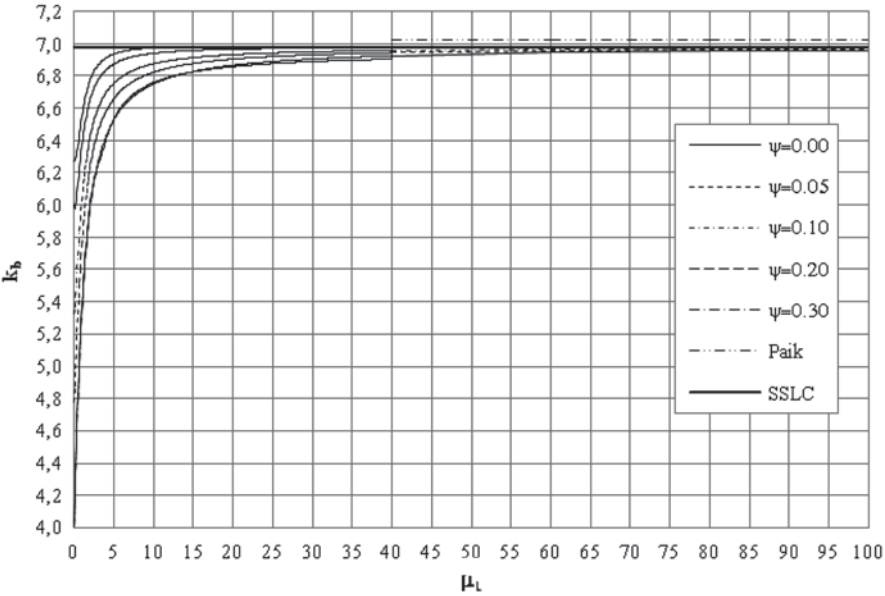


Fig. 10.2 Buckling coefficient distribution for SSLR plates

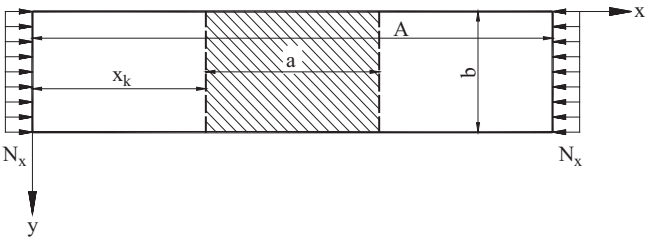


Fig. 10.3 SRLS plate scheme

The general equilibrium equation is:

$$\Delta U_p + \sum_{k=1}^{n_t} \Delta U^T_{s,k} = \Delta T_p \tag{10.17}$$

having denoted by ΔU_p ($\Delta U^T_{s,k}$) the strain energy due to plate bending (due to torque and warping of the k-th transverse) and by ΔT_p the work done by compressive forces acting in x-direction on the plate. Also in this case the strain energy due to warping may be accounted only if transverses' warping is restrained. The strain energy due to plate bending and the work done by compressive forces acting in x-direction on the plate may be expressed as in Eqs. (10.5) and (10.10), respectively. The strain

energy due to torque and warping rigidities of the *k*-th attached transverse, with similar notation, is:

$$\Delta U_{s,k}^{T} = \frac{1}{2} G J_k^T \int_0^b \left(\left. \frac{\partial \Phi_k^T}{\partial y} \right|_{x=ka} \right)^2 dy + \frac{1}{2} E I_{w,k}^T \int_0^b \left(\left. \frac{\partial^2 \Phi_k^T}{\partial y^2} \right|_{x=ka} \right)^2 dy \tag{10.18}$$

Imposing the deflection angle continuity condition along the junction between the plating and the transverses, the following congruence condition must be verified:

$$\Phi_k^T(ka, y) = \left. \frac{\partial w}{\partial x} \right|_{x=ka} \quad \forall k = 1 \ldots n_t \tag{10.19}$$

so that Eq. (10.17) may be finally rewritten as follows:

$$\begin{gathered} \Delta U_{s,k}^{T} = \frac{\pi^4 G J_k^T}{4a^2 b} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{p=1}^{M} mn^2 p w_{m,n} w_{p,n} \cos(m\pi k) \cos(p\pi k) \\ \left(1 + \frac{2\pi^2 n^2 (1+\nu) I_{w,k}^T}{J_k^T b^2} \right) \end{gathered} \tag{10.20}$$

and the following eigenvalue problem may be solved:

$$\begin{gathered} w_{j,k} \left(\frac{j^2}{\alpha^2} + k^2 \right)^2 + 2 \sum_{p=1}^{M} \frac{w_{p,k} j k^2 p}{n_t + 1} \sum_{r=1}^{n_t} \cos(j\pi r) \cos(p\pi r) \\ \left(\frac{\mu_T}{\alpha^2} + n^2 \pi^2 \psi_T \right) = -\frac{k_b}{\alpha^2} j^2 w_{j,k} \end{gathered} \tag{10.21}$$

Based on computed results, a closed-form buckling formula for SRLS plates has been derived, as a function of transverses' torque rigidity μ_T and panel aspect ratios α:

$$k_{b-SRLS}^{new} = 4 \frac{\alpha^2 + \alpha + f_1(\mu_T)}{\alpha^2 + \alpha + f_2(\mu_T)} \tag{10.22}$$

with:

$$\begin{aligned} f_1(\mu_T) &= \frac{0.6632 \mu_T^2 + 2.2390 \mu_T + 0.0400}{\mu_T^2 + 0.01106 \mu_T + 0.0001949} \\ f_2(\mu_T) &= \frac{-0.4277 \mu_T^2 + 2.1010 \mu_T + 0.1589}{\mu_T^2 + 0.05280 \mu_T + 0.0007742} \end{aligned} \tag{10.23}$$

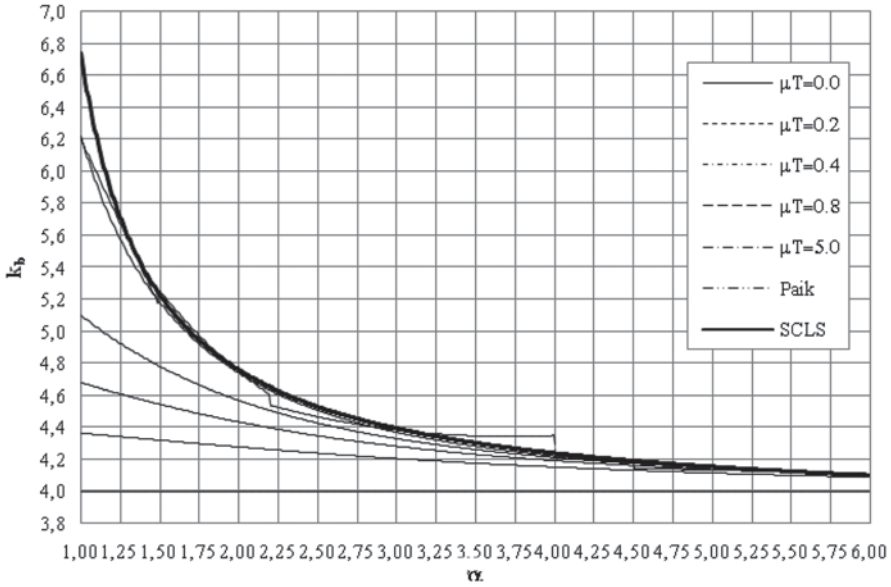


Fig. 10.4 Buckling coefficient distribution for SRLS plates

where μ_T is defined as follows:

$$\mu_T = \frac{GJ_T}{Da} = 2(1-\nu)\frac{J_T}{J_{P-T}} \tag{10.24}$$

The ratio between stiffener J_T and plating J_{P-T} St. Venant moments of inertia may be assumed not greater than one, on the safety side, to account for supporting members torsional buckling. In Fig. 10.4 buckling coefficients are plotted versus α for several values of transverses rigidity ratio μ_T, namely 0.0, 0.2, 0.4, 0.8 and 5.0. The curve by Paik & Thayamballi and the relevant one for plates clamped at short edges and simply supported at long ones (SCLS) are shown, too.

10.4 Test Examples

In the following the proposed formulas are applied and compared with the relevant results obtained by some eigenvalue buckling analyses carried out by ANSYS for different stiffened panels. All panels, as well as attached stiffeners, are within practical proportions from a design point of view and are made of high strength steel with R_{eH}=315 N/mm^2, E=206 GPa,ν=0.30. The convergence of solution has been studied, decreasing the mean shell dimensions. Three models have been built: the coarse mesh, the fine mesh and the very fine mesh ones, with mean panel dimen-

sions of 0.20, 0.10 and 0.05 m, respectively. The St. Venant supporting members moment of inertia has been evaluated by the following formula for T-sections, derived by thin-walled beam theory:

$$J_{L/T} = \frac{1}{3} h_w t_w^3 + \frac{1}{3} b_f t_f^3 \left(1 - 0.63 \frac{t_f}{b_f} \right) \tag{10.25}$$

where h_w and t_w are the stiffener web height and thickness, while b_f and t_f are the flange breadth and thickness, respectively. It has also been verified [1] that warping stresses due to restrained torsion for thin-walled beams can be evaluated by classical theories with good confidence, especially for thin- walled structures with monoconnected cross-section.

10.4.1 Plates Rotationally Restrained at Long Edges

Stiffened panels' geometrical properties, shown in Table 10.1, have been varied so that torque and warping rigidity ratios lie in the range 0.10–1.40 and 0.20–2.00, respectively. In all cases stiffener dimensions have been chosen to avoid overall buckling from occurring before local plate buckling. In Table 10.2 FE buckling coefficients are compared with the new formula and that one by Paik & Thayamballi, for the unrestrained warping case. The restrained warping case has also been analysed, comparing the FE buckling values with the relevant results obtained by the new proposed formula. Theoretical values are, in all cases, very close to FE ones.

10.4.2 Plates Rotationally Restrained at Short Edges

In Table 10.3 the geometrical properties of 33 stiffened panels are listed: also in this case stiffener dimensions, as well as plating thickness, have been suitably varied so that the torque rigidity ratio lies in the range 0.05–1.10. Stiffeners' geometrical properties have been always chosen to avoid overall buckling from occurring before local plate buckling. In Table 10.4 a comparison with the relevant results obtained by ANSYS is carried out. As in the previous case, theoretical values are very close to FE ones.

10.5 Conclusions

The minimum energy principle is applied to the buckling analysis of plates on rotationally and warping restrained supports, regarding the isolated plate as part of an infinitely wide stiffened panel. Simple design formulas have been derived by curve

Table 10.1 SSLR plates—geometrical properties

N	a (mm)	b (mm)	t (mm)	h_w (mm)	t_w (mm)	b_f (mm)	t_f (mm)	μ_L	ψ_L
1	2400	800	12	250	12	100	15	0.747	0.987
2	2400	800	15	250	12	100	15	0.382	0.632
3	2400	800	20	250	12	100	15	0.161	0.355
4	2400	800	12	250	12	100	20	1.146	0.987
5	2400	800	12	250	12	100	25	1.400	0.987
6	2400	800	12	250	15	100	20	1.400	0.987
7	2400	800	12	250	15	100	25	1.400	0.987
8	2400	800	12	150	12	80	15	0.504	0.355
9	2400	800	15	150	12	80	15	0.258	0.228
10	2400	800	12	150	12	80	20	0.809	0.355
11	2400	800	12	150	12	80	25	1.279	0.355
12	2400	800	12	150	15	80	20	1.059	0.355
13	2400	800	12	150	15	80	25	1.400	0.355
14	2400	800	12	350	12	150	15	1.093	1.935
15	2400	800	15	350	12	150	15	0.560	1.239
16	2400	800	20	350	12	150	15	0.236	0.697
17	2400	800	12	350	12	150	20	1.400	1.935
18	2400	800	12	350	12	150	25	1.400	1.935
19	2400	800	12	350	15	150	20	1.400	1.935
20	2400	800	12	350	15	150	25	1.400	1.935
21	3000	1000	12	250	12	100	15	0.598	0.987
22	3000	1000	15	250	12	100	15	0.306	0.632
23	3000	1000	20	250	12	100	15	0.129	0.355
24	3000	1000	12	250	12	100	20	0.916	0.987
25	3000	1000	12	250	12	100	25	1.400	0.987
26	3000	1000	12	250	15	100	20	1.250	0.987
27	3000	1000	12	250	15	100	25	1.400	0.987
28	3000	1000	12	150	12	80	15	0.403	0.355
29	3000	1000	15	150	12	80	15	0.206	0.228
30	3000	1000	12	150	12	80	20	0.647	0.355
31	3000	1000	12	150	12	80	25	1.023	0.355
32	3000	1000	12	150	15	80	20	0.847	0.355
33	3000	1000	12	150	15	80	25	1.224	0.355
34	3000	1000	12	350	12	150	15	0.874	1.935
35	3000	1000	15	350	12	150	15	0.448	1.239
36	3000	1000	20	350	12	150	15	0.189	0.697
37	3000	1000	12	350	12	150	20	1.381	1.935
38	3000	1000	12	350	12	150	25	1.400	1.935
39	3000	1000	12	350	15	150	20	1.400	1.935
40	3000	1000	12	350	15	150	25	1.400	1.935
41	4000	800	12	250	12	100	15	0.747	0.987
42	4000	800	15	250	12	100	15	0.382	0.632
43	4000	800	20	250	12	100	15	0.161	0.355
44	4000	800	12	250	12	100	20	1.146	0.987
45	4000	800	12	250	12	100	25	1.400	0.987
46	4000	800	12	250	15	100	20	1.400	0.987
47	4000	800	12	250	15	100	25	1.400	0.987
48	4000	800	12	150	12	80	15	0.504	0.355

Table 10.1 (continued)

N	a (mm)	b (mm)	t (mm)	h_w (mm)	t_w (mm)	b_f (mm)	t_f (mm)	μ_L	ψ_L
49	4000	800	12	150	12	80	20	0.809	0.355
50	4000	800	12	150	12	80	25	1.279	0.355
51	4000	800	12	150	15	80	20	1.059	0.355
52	4000	800	12	150	15	80	25	1.400	0.355
53	4000	800	12	350	12	150	15	1.093	1.935
54	4000	800	15	350	12	150	15	0.560	1.239
55	4000	800	20	350	12	150	15	0.236	0.697
56	4000	800	12	350	12	150	20	1.400	1.935
57	4000	800	12	350	12	150	25	1.400	1.935
58	4000	800	12	350	15	150	20	1.400	1.935
59	4000	800	12	350	15	150	25	1.400	1.935
60	4000	800	12	350	15	150	30	1.400	1.935

Table 10.2 SSLR plates—numerical comparison with FE results

N	Unrestrained warping					Restrained warping		
	kb-NEW	kb-PAIK	kb-FE	kb-NEW/ kb-FE	kb-PAIK/ kb-FE	kb-NEW -W	kb-FE -W	kb-NEW -W/ kb-FE-W
1	5.122	5.077	5.251	0.975	0.967	6.734	6.694	1.006
2	4.654	4.612	4.874	0.955	0.946	6.595	6.376	1.034
3	4.296	4.275	4.494	0.956	0.951	6.392	6.091	1.049
4	5.488	5.469	5.295	1.036	1.033	6.821	6.715	1.016
5	5.663	5.664	5.334	1.062	1.062	6.864	6.729	1.020
6	5.663	5.664	5.656	1.001	1.001	6.864	7.008	0.979
7	5.663	5.664	5.709	0.992	0.992	6.864	7.024	0.977
8	4.826	4.779	5.156	0.936	0.927	6.450	6.576	0.981
9	4.459	4.428	4.754	0.938	0.931	6.111	6.298	0.970
10	5.187	5.145	5.237	0.991	0.982	6.555	6.656	0.985
11	5.585	5.576	5.296	1.054	1.053	6.688	6.706	0.997
12	5.419	5.393	5.557	0.975	0.970	6.632	6.927	0.957
13	5.663	5.664	5.626	1.007	1.007	6.715	6.982	0.962
14	5.446	5.423	5.296	1.028	1.024	6.890	6.695	1.029
15	4.900	4.852	4.926	0.995	0.985	6.720	6.396	1.051
16	4.423	4.394	4.535	0.975	0.969	6.611	6.104	1.083
17	5.663	5.664	5.329	1.063	1.063	6.934	6.707	1.034
18	5.663	5.664	5.356	1.057	1.058	6.934	6.716	1.032
19	5.663	5.664	5.729	0.989	0.989	6.934	7.017	0.988
20	5.663	5.664	5.769	0.982	0.982	6.934	7.025	0.987
21	4.948	4.900	5.240	0.944	0.935	6.700	6.613	1.013
22	4.537	4.501	4.847	0.936	0.929	6.588	6.307	1.045
23	4.240	4.222	4.472	0.948	0.944	6.399	6.039	1.060
24	5.293	5.257	5.289	1.001	0.994	6.773	6.690	1.012
25	5.663	5.664	5.335	1.062	1.062	6.864	6.714	1.022
26	5.564	5.554	5.625	0.989	0.987	6.840	6.982	0.980
27	5.663	5.664	5.683	0.997	0.997	6.864	7.008	0.979
28	4.684	4.641	5.042	0.929	0.921	6.419	6.526	0.984

Table 10.2 (continued)

N	Unrestrained warping					Restrained warping		
	kb-NEW	kb-PAIK	kb-FE	kb-NEW/ kb-FE	kb-PAIK/ kb-FE	kb-NEW -W	kb-FE -W	kb-NEW -W/ kb-FE-W
29	4.373	4.347	4.665	0.937	0.932	6.097	6.235	0.978
30	5.008	4.960	5.144	0.974	0.964	6.499	6.598	0.985
31	5.389	5.360	5.225	1.031	1.026	6.622	6.662	0.994
32	5.226	5.186	5.448	0.959	0.952	6.567	6.864	0.957
33	5.545	5.533	5.544	1.000	0.998	6.675	6.936	0.962
34	5.253	5.214	5.278	0.995	0.988	6.852	6.708	1.021
35	4.749	4.703	4.934	0.962	0.953	6.702	6.396	1.048
36	4.344	4.319	4.533	0.958	0.953	6.620	6.084	1.088
37	5.651	5.651	5.318	1.063	1.063	6.932	6.722	1.031
38	5.663	5.664	5.353	1.058	1.058	6.934	6.734	1.030
39	5.663	5.664	5.699	0.994	0.994	6.934	7.034	0.986
40	5.663	5.664	5.747	0.985	0.986	6.934	7.044	0.984
41	5.122	5.077	5.137	0.997	0.988	6.734	6.892	0.977
42	4.654	4.612	4.807	0.968	0.960	6.595	6.566	1.004
43	4.296	4.275	4.484	0.958	0.953	6.392	6.341	1.008
44	5.488	5.469	5.167	1.062	1.058	6.821	6.871	0.993
45	5.663	5.664	5.195	1.090	1.090	6.864	6.917	0.992
46	5.663	5.664	5.517	1.027	1.027	6.864	7.134	0.962
47	5.663	5.664	5.553	1.020	1.020	6.864	7.187	0.955
48	4.826	4.779	5.080	0.950	0.941	6.450	6.785	0.951
49	5.187	5.145	5.138	1.010	1.001	6.555	6.833	0.959
50	5.585	5.576	5.187	1.077	1.075	6.688	6.883	0.972
51	5.419	5.393	5.429	0.998	0.993	6.632	7.072	0.938
52	5.663	5.664	5.488	1.032	1.032	6.715	7.128	0.942
53	5.446	5.423	5.146	1.058	1.054	6.890	6.934	0.994
54	4.900	4.852	4.820	1.017	1.007	6.720	6.618	1.015
55	4.423	4.394	4.511	0.981	0.974	6.611	6.472	1.021
56	5.663	5.664	5.168	1.096	1.096	6.934	6.966	0.995
57	5.663	5.664	5.192	1.091	1.091	6.934	6.977	0.994
58	5.663	5.664	5.547	1.021	1.021	6.934	7.269	0.954
59	5.663	5.664	5.578	1.015	1.015	6.934	7.284	0.952
60	5.663	5.664	5.611	1.009	1.009	6.934	7.295	0.951
			MEAN	*1.003*	*0.999*			*0.997*
			COV%	*4.368*	*4.607*			*3.432*

Table 10.3 SRLS plates—geometrical properties

N	a (mm)	b (mm)	t (mm)	h_w (mm)	t_w (mm)	b_f (mm)	t_f (mm)	μ_T
1	2400	800	12	250	12	100	15	0.249
2	2400	800	12	250	12	100	20	0.3819
3	2400	800	12	250	12	100	25	0.5902
4	2400	800	12	250	15	100	20	0.5209
5	2400	800	12	250	15	100	25	0.7292
6	2400	800	12	150	12	80	15	0.1679
7	2400	800	12	150	12	80	20	0.2695

Table 10.3 (continued)

N	a (mm)	b (mm)	t (mm)	h_w (mm)	t_w (mm)	b_f (mm)	t_f (mm)	μ_T
8	2400	800	12	150	12	80	25	0.426
9	2400	800	12	150	15	80	20	0.353
10	2400	800	12	150	15	80	25	0.510
11	2400	800	12	350	12	150	15	0.364
12	2400	800	12	350	12	150	20	0.575
13	2400	800	12	350	12	150	25	0.912
14	2400	800	12	350	15	150	20	0.770
15	2400	800	12	350	15	150	25	1.107
16	3000	1000	12	250	12	100	15	0.199
17	3000	1000	15	250	12	100	15	0.102
18	3000	1000	12	250	12	100	20	0.306
19	3000	1000	12	250	12	100	25	0.472
20	3000	1000	12	250	15	100	20	0.417
21	3000	1000	12	250	15	100	25	0.583
22	3000	1000	12	150	12	80	15	0.134
23	3000	1000	15	150	12	80	15	0.069
24	3000	1000	12	150	12	80	20	0.216
25	3000	1000	12	150	12	80	25	0.341
26	3000	1000	12	150	15	80	20	0.282
27	3000	1000	12	150	15	80	25	0.408
28	3000	1000	12	350	12	150	15	0.291
29	3000	1000	15	350	12	150	15	0.149
30	3000	1000	12	350	12	150	20	0.460
31	3000	1000	12	350	12	150	25	0.730
32	3000	1000	12	350	15	150	20	0.616
33	3000	1000	12	350	15	150	25	0.886

Table 10.4 SRLS plates—numerical comparison with FE results

N	kb-NEW	kb-PAIK	kb-FE	kb-NEW/kb-FE	kb-PAIK/kb-FE
1	4.232	4.242	4.100	1.032	1.035
2	4.279	4.285	4.102	1.043	1.045
3	4.314	4.323	4.104	1.051	1.053
4	4.305	4.312	4.150	1.037	1.039
5	4.326	4.343	4.152	1.042	1.046
6	4.185	4.196	4.121	1.015	1.018
7	4.242	4.251	4.125	1.028	1.031
8	4.289	4.295	4.129	1.039	1.040
9	4.271	4.278	4.166	1.025	1.027
10	4.303	4.310	4.170	1.032	1.034
11	4.274	4.281	4.090	1.045	1.047
12	4.312	4.321	4.091	1.054	1.056
13	4.337	4.347	4.092	1.060	1.062
14	4.329	4.346	4.141	1.045	1.049
15	4.345	4.347	4.142	1.049	1.049
16	4.205	4.216	4.149	1.014	1.016
17	4.141	4.138	4.107	1.008	1.008

Table 10.4 (continued)

N	kb-NEW	kb-PAIK	kb-FE	kb-NEW/kb-FE	kb-PAIK/kb-FE
18	4.256	4.264	4.151	1.025	1.027
19	4.297	4.304	4.156	1.034	1.035
20	4.287	4.293	4.179	1.026	1.027
21	4.313	4.322	4.181	1.032	1.034
22	4.161	4.169	4.116	1.011	1.013
23	4.150	4.101	4.071	1.019	1.007
24	4.215	4.226	4.121	1.023	1.025
25	4.267	4.275	4.125	1.035	1.036
26	4.247	4.256	4.162	1.020	1.023
27	4.285	4.291	4.166	1.028	1.030
28	4.250	4.260	4.102	1.036	1.038
29	4.172	4.182	4.059	1.028	1.030
30	4.295	4.301	4.104	1.047	1.048
31	4.326	4.343	4.105	1.054	1.058
32	4.316	4.327	4.153	1.039	1.042
33	4.336	4.347	4.155	1.043	1.046
			MEAN	*1.034*	*1.036*
			COV%	*1.291*	*1.387*

fitting of a large amount of data and several stiffened panels have been analysed by ANSYS, where some eigenvalue buckling analyses have been carried out. In all cases a good agreement between theoretical and FE values was found. The influence of supporting members' torsional and warping rigidity is always appreciable and it may not be neglected in a refined buckling strength check of platings under uniaxial compression.

Acknowledgements The work has been financed by the University of Naples "Parthenope", Department of Science and Technology.

References

1. Campanile A, Mandarino M, Piscopo V (2010) On the exact solution of non-uniform torsion for beams with asymmetric cross-section. World Academy of Science, Engineering and Technology, Issue 31, July 2009, pp. 46–53
2. Evans JH (1960) Strength of wide plates under uniform edge compression. Trans SNAME, 1960 n. 68, pp. 585–621
3. Gerard G, Becker H (1954) Handbook of structural stability, Part I. Buckling of flat plates, NACA Technical Note, 1954, No. 3781
4. Hasegawa A, Asce AM, Ota K, Nishino F, Asce M (1976) Buckling strength of multiple stiffened panels. Proceedings of the Specialist Conference "Methods of structural analysis"
5. Lundquist E, Stowell EZ (1942) Critical compressive stress for flat rectangular plates elastically restrained. NACA Technical Note, 1942, No. 733
6. Paik JK, Kim JY (2000) Bucking strength of steel plating with elastically restrained edges. Thin-walled Structures, 37

7. Piscopo V (2010) Refined buckling analysis of rectangular plates under uniaxial and biaxial compression. World Academy of Science, Engineering and Technology, Issue 46, October 2010, pp. 554–561
8. Piscopo V (2012) Local and Overall Buckling of Uniaxially Compressed Stiffened Panels, Sustainable Maritime Transportation and Exploitation of Sea Resources - Proceedings of the 14th International Congress of the International Maritime Association of the Mediterranean, IMAM 2011
9. Timoshenko SP, Gere JM (1985) Theory of elastic stability. Mc-Graw Hill International Book Company, 17th edition
10. Timoshenko SP, Woinowsky-Krieger S (1959) Theory of Plates and Shells. Mc-Graw Hill International Book Company

Chapter 11
Analytic Programming—A New Tool for Synthesis of Controller for Discrete Chaotic Lozi Map

R. Senkerik, Z. Kominkova Oplatkova, M. Pluhacek and I. Zelinka

Abstract In this chapter, it is presented a utilization of a novel tool for symbolic regression, which is analytic programming, for the purpose of the synthesis of a new feedback control law. This new synthesized chaotic controller secures the fully stabilization of selected discrete chaotic systems, which is the two-dimensional Lozi map. The paper consists of the descriptions of analytic programming as well as selected chaotic system, used heuristic and cost function design. For experimentation, Self-Organizing Migrating Algorithm (SOMA) and Differential evolution (DE) were used. Two selected experiments are detailed described.

Keywords Analytic programming · Symbolic regression · Chaos control · Evolutionary algorithms · Lozi map

11.1 Introduction

During the recent years, usage of new intelligent systems in engineering, technology, modeling, computing and simulations has attracted the attention of researchers worldwide. The most current methods are mostly based on soft computing, which is a discipline tightly bound to computers, representing a set of methods of special algorithms, belonging to the artificial intelligence paradigm. The most popular of these methods are neural networks, evolutionary algorithms, fuzzy logic and tools for symbolic regression like genetic programming. Currently, evolutionary algorithms are known as a powerful set of tools for almost any difficult and complex optimization problem.

R. Senkerik (✉) · Z. K. Oplatkova · M. Pluhacek
Department of Informatics and Artificial Intelligence, Tomas Bata University in Zlin,
Nad Stranemi 4511, 760 05 Zlin, Czech Republic
e-mail: senkerik@fai.utb.cz

I. Zelinka
Department of Computer Science, VŠB-Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
e-mail: ivan.zelinka@vsb.cz

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_11,

The interest about the interconnection between evolutionary techniques and control of chaotic systems is spread daily. First steps were done in [1] representing the utilization of differential evolution algorithm for the synchronization and control of chaotic systems. The papers [2, 3] were concerned to tune several parameters inside the original control technique for discrete chaotic systems. The evolutionary tuned control technique was based on Pyragas method: Extended delay feedback control—ETDAS [4]. Another example of interconnection between deterministic chaos and evolutionary algorithms represents the research focused on the embedding of chaotic dynamics into the evolutionary algorithms [5–7].

This chapter shows a possibility how to generate the whole control law by means of analytic programming (AP) (not only to optimize several parameters) for the purpose of stabilization of the selected discrete chaotic system. The synthesis of control is inspired by the Pyragas's delayed feedback control technique [8, 9].

AP is a superstructure of EAs and is used for synthesis of analytic solution according to the required behaviour. Control law from the proposed system can be viewed as a symbolic structure, which can be synthesized according to the requirements for the stabilization of the chaotic system.

This chapter represents an extension of work [10] and cumulation of experiences from the previous work [2, 11–13].

Firstly, AP is explained, and then a description of used soft-computing tools is proposed. The next sections are focused on the problem design and the description of cost function utilized within the evolutionary process. Results and conclusion follow afterwards.

11.2 Motivation

This work is focused on the expansion of AP application for synthesis of a whole robust control law instead of parameters tuning for existing and commonly used control technique to stabilize desired Unstable Periodic Orbits (UPO) of selected discrete chaotic system.

This work represents an extension of previous research [14, 15], with the application to the chaotic discrete Lozi map.

In general, this research is concerned to stabilize Lozi map chaotic system at p-1 UPO, which is a stable state, utilizing the synthesized control law.

11.3 Analytic Programming

Basic principles of the AP were developed in 2001. Until that time only genetic programming (GP) and grammatical evolution (GE) had existed. GP uses genetic algorithms (GA) while AP can be used with any EA, independently on an individual representation. Various applications of AP are described in [16–18].

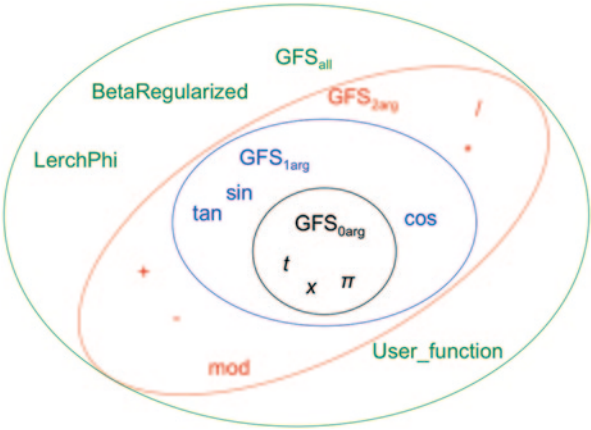


Fig. 11.1 Hierarchy in the GFS

The core of AP is based on a special set of mathematical objects and operations. The set of mathematical objects is a set of functions, operators and so-called terminals (as well as in GP), which are usually constants or independent variables. This set of variables is usually mixed together and consists of functions with different number of arguments. Because of the variability of the content of this set, it is termed the "general functional set"—GFS. The structure of GFS is created by subsets of functions according to the number of their arguments. For example, GFS_{all} is a set of all functions, operators and terminals, GFS_{3arg} is a subset containing functions with only three arguments, GFS_{0arg} represents only terminals, etc. The subset structure presence in GFS is vitally important for AP. The hierarchy of GFS is depicted in Fig. 11.1. It is used to avoid the synthesis of pathological programs, i.e. programs containing functions without arguments, etc. The content of GFS is dependent only on the user. Various functions and terminals can be mixed together [17].

The second part of the AP core is a sequence of mathematical operations used for the program synthesis. These operations are used to transform an individual of a population into a suitable program. Mathematically stated, it is mapping from an individual domain into a program domain. The mapping consists of two main parts. The first part is called Discrete Set Handling (DSH) (Fig. 11.2; [17]) and the second one stands for security procedures which do not allow synthesizing pathological programs. The method of DSH, when used, allows handling arbitrary objects including nonnumeric objects such as linguistic terms {hot, cold, dark…}, logic terms (True, False) or other user defined functions. In the AP, DSH is used to map an individual into GFS and together with security procedures creates the above-mentioned mapping, which transforms an arbitrary individual into a program.

AP needs some EA [17] that consists of a population of individuals for its run. Individuals in the population consist of integer parameters, i.e. an individual is an integer index pointing into GFS. The creation of the program can be schematically observed in Fig. 11.3. The individual contains numbers which are indices for GFS.

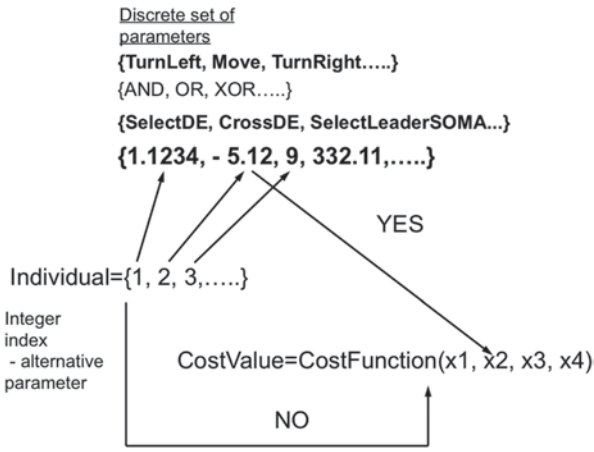


Fig. 11.2 Discrete set handling

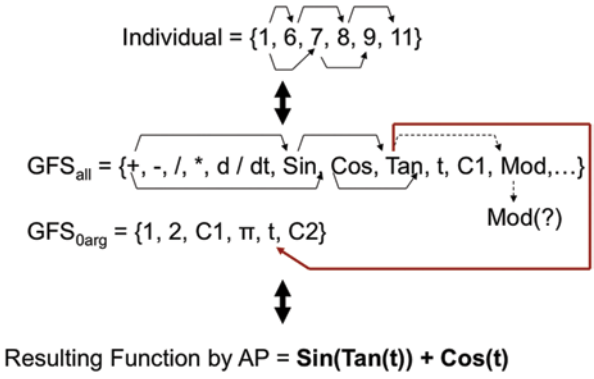


Fig. 11.3 The main principle of AP

Figure 11.3 demonstrates an artificial example as to how a final function is created from an integer individual via Discrete Set Handling (DSH).

The number 1 in the position of the first parameter means that the operator plus (+) from GFS_{all} is used (the end of the individual is far enough). Because the operator + must have at least two arguments, the next two index pointers 6 (sin from GFS) and 7 (cos from GFS) are dedicated to this operator as its arguments. The two functions, sin and cos, are one-argument functions, therefore the next unused pointers 8 (tan from GFS) and 9 (*t* from GFS) are dedicated to the sin and cos functions. As an argument of cos, the variable *t* is used, and this part of the resulting function is closed (*t* has zero arguments) in its AP development. The one-argument function tan remains, and there is one unused pointer 11, which stands for Mod in GFS_{all}. The modulo operator needs two arguments but the individual in the example has no other indices (pointers, arguments). In this case, it is necessary to employ security procedures and jump to the subset with GFS_{0arg}. The function tan is mapped on *t* from GFS_{0arg} which is on the 11th position, cyclically from the beginning.

Fig. 11.4 Schema of AP_{meta} procedures

AP exists in 3 versions—basic without constant estimation, AP_{nf}—estimation by means of nonlinear fitting package in *Wolfram Mathematica* (www.wolfram.com) environment and AP_{meta}—constant estimation by means of another evolutionary algorithms; the term "meta" implies meta-evolution.

AP_{basic} stands for the version where constant estimation is done in the same way as in genetic programming. In the case that data approximation requires estimation of coefficients in the approximated polynomial or moving the basic curve from the axes origin, the user has to assign a set of constant values into GFS. This results in a huge enlargement of the functional sets and deceleration of the evolutionary procedure. Therefore two other strategies were adopted—AP_{nf} and AP_{meta}. These two versions of AP use the constant *K,* which is indexed during the evolution (11.1). When *K* is needed, a proper index is assigned—K_1, K_2, … K_n (11.2). Numeric values to indexed *K*s are estimated (11.3) either by means of nonlinear fitting methods—AP_{nf} or by means of the second evolutionary algorithm—AP_{meta}.

$$\frac{x^2 + K}{\pi^K} \tag{11.1}$$

$$\frac{x^2 + K_1}{\pi^{K_2}} \tag{11.2}$$

$$\frac{x^2 + 3.156}{\pi^{90.78}} \tag{11.3}$$

AP_{meta} is a very time consuming process and the number of cost function evaluations, which is one of the comparative factors, is usually very high. This is given by two evolutionary processes (Fig. 11.4.) required for obtaining of a new synthesized symbolic formula.

EA_{master} represents the main evolutionary algorithm for AP, EA_{slave} is the secondary evolutionary algorithm within AP for the process of constants (coefficients) estimation. Thus the total number of cost function evaluation (CFE) required for the obtaining of solution is given by (11.4):

$$CFE = CFE(EA_{master}) \cdot CFE(EA_{slave}) \tag{11.4}$$

Despite this disadvantage of very high CFE required, because of the character of the problem, many simulations simply cannot utilize nonlinear fitting methods in the Mathematica environment or predefined huge set of possible constants.

11.4 Used Soft-Computing Tools

This section gives the brief overview and the description of used soft-computing tools. This research utilized the symbolic regression tool, which is analytic programming and two evolutionary algorithms: Self-Organizing Migrating Algorithm [19]; and Differential Evolution [20].

Future simulations expect a usage of soft computing GAHC algorithm (modification of HC12) [21] and a CUDA implementation of HC12 algorithm [22].

11.4.1 Self-Organizing Migrating Algorithm (SOMA)

Self-Organizing Migrating Algorithm is a stochastic optimization algorithm that is modeled on the basis of social behavior of cooperating individuals [19]. It was chosen because it has been proven that the algorithm has the ability to converge towards the global optimum [19] and due to the successful applications together with AP [23, 24].

SOMA works on a population of candidate solutions in loops called *migration loops*. The population is initialized randomly distributed over the search space at the beginning of the search. In each loop, the population is evaluated and the solution with the highest fitness becomes the leader L. Apart from the leader, in one migration loop, all individuals will traverse the input space in the direction of the leader. Mutation, the random perturbation of individuals, is an important operation for evolutionary strategies (ES). It ensures the diversity amongst the individuals and it also provides the means to restore lost information in a population. Mutation is different in SOMA compared with other ES strategies. SOMA uses a parameter called PRT to achieve perturbation. This parameter has the same effect for SOMA as mutation has for genetic algorithms.

The novelty of this approach is that the PRT Vector is created before an individual starts its journey over the search space. The PRT Vector defines the final movement of an active individual in search space.

The randomly generated binary perturbation vector controls the allowed dimensions for an individual. If an element of the perturbation vector is set to zero, then the individual is not allowed to change its position in the corresponding dimension.

An individual will travel a certain distance (called the PathLength) towards the leader in n steps of defined length. If the PathLength is chosen to be greater than one, then the individual will overshoot the leader. This path is perturbed randomly.

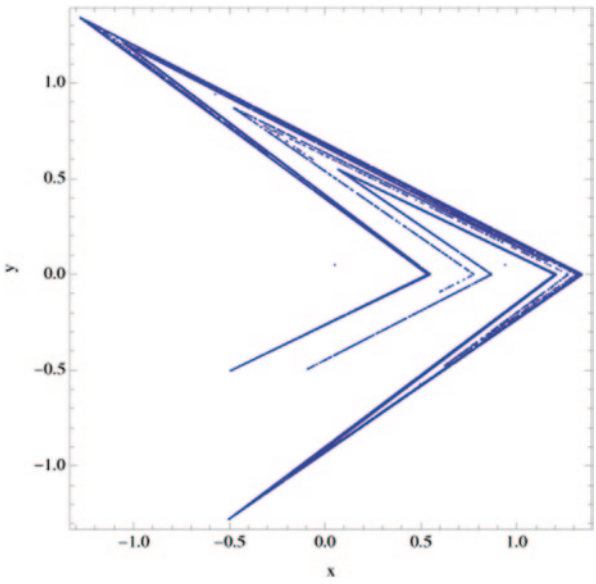


Fig. 11.5 *x*, *y* plot of the Lozi map

11.4.2 Differential Evolution

DE is a population-based optimization method that works on real-number-coded individuals [25–27]. DE is quite robust, fast, and effective, with global optimization ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Description of used DERand1Bin strategy is presented in (11.5). Please refer to [20, 27] for the description of all other strategies.

$$u_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \tag{11.5}$$

11.5 Lozi Map

Lozi map is the selected example of chaotic systems, which represents the simple discrete two-dimensional chaotic map. The *x*, *y* plot of the Lozi map is depicted in Fig. 11.5. The map equations are given in (11.6). The parameters are: $a = 1.7$ and $b = 0.5$ as suggested in [28, 29]. The chaotic behavior of the uncontrolled Lozi map is depicted in Fig. 11.6.

$$\begin{aligned} X_{n+1} &= 1 - a\left|X_n\right| + bY_n \\ Y_{n+1} &= X_n \end{aligned} \tag{11.6}$$

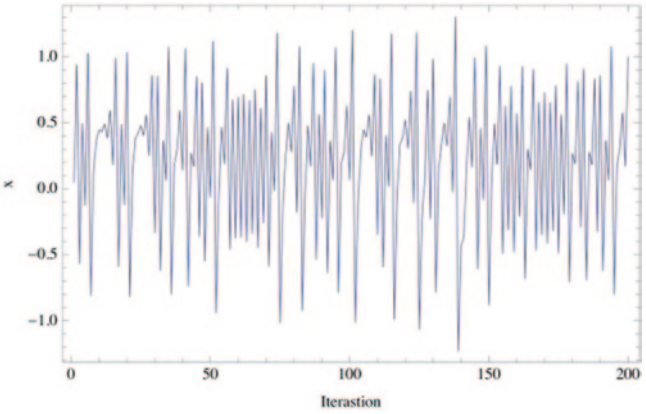


Fig. 11.6 Iterations of the uncontrolled Lozi map (variable *x*)

11.6 Original Chaos Control Method

This work is focused on explanation of application of AP for synthesis of a whole control law instead of demanding tuning of any original method control law to stabilize desired Unstable Periodic Orbits (UPO). In this research desired UPO is only p − 1 (the fixed point, which represents the stable state). Original Time-Delay-Auto-Synchronization (TDAS) delayed feedback control method was used in this research as an inspiration for synthesizing a new feedback control law by means of evolutionary techniques and for preparation of sets of basic functions and operators for AP.

The original control method—TDAS has form (11.7) and its discrete form is given in (11.8).

$$F(t) = K\left[x\left(t-\tau\right) - x(t)\right] \tag{11.7}$$

$$F_n = K(x_{n-m} - x_n) \tag{11.8}$$

Where: K is adjustable constant, F is the perturbation, τ_d is a time delay; and m is the period of m-periodic orbit to be stabilized. The perturbation F_n in Eq. (11.8) may have arbitrarily large value, which can cause diverging of the system. Therefore, F_n should have a value between $-F_{\max}$, $F_{\max}$. In this work a suitable $F_{\max}$ value was taken from the previous research.

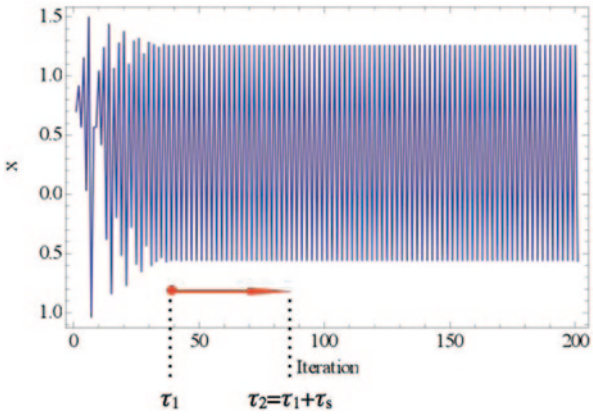


Fig. 11.7 "Floating window" for minimization

11.7 Cost Function Design

The proposal of the basic cost function (CF) is in general based on the simplest CF, which could be used problem-free only for the stabilization of p − 1 orbit. The idea was to minimize the area created by the difference between the required state and the real system output on the whole simulation interval—τ_i. This CF design is very convenient for the evolutionary searching process due to the relatively favorable CF surface. Nevertheless, this simple approach has one big disadvantage, which is the including of initial chaotic transient behavior of not stabilized system into the cost function value. As a result of this, the very tiny change of control method setting for extremely sensitive chaotic system causing very small change of CF value, can be suppressed by the above-mentioned including of initial chaotic transient behavior.

But another universal cost function had to be used for stabilizing of extremely sensitive chaotic system and for having the possibility of adding penalization rules. It was synthesized from the simple CF and other terms were added.

This CF is in general based on searching for desired stabilized periodic orbit and thereafter calculation of the difference between desired and found actual periodic orbit on the short time interval—τ_s (40 iterations for higher order UPO) from the point, where the first minimal value of difference between desired and actual system output is found (i.e. floating window for minimization—see Fig. 11.7.).

Such a design of universal CF should secure the successful stabilization of either p − 1 orbit (stable state) or higher periodic orbits anywise phase shifted. Furthermore, due to CF values converging towards zero, this CF also allows the using of decision rules, avoiding very time demanding simulations. This rule stops EA immediately, when the first individual with good parameter structure is reached, thus the value of CF is lower then the acceptable (CF_{acc}) one. Based on the numerous experiments, typically $CF_{acc} = 0.001$ at time interval $\tau_s = 20$ iterations, thus the difference between desired and actual output has the value of 0.0005 per iteration—

Table 11.1 SOMA settings for AP

SOMA Parameter	Value
PathLength	3
Step	0.11
PRT	0.1
PopSize	50
Migrations	4
Max. CF Evaluations (CFE)	5,345

i.e. successful stabilization for the used control technique. The CF_{Basic} has the form (11.9):

$$CF_{Basic} = pen_1 + \sum_{t=\tau 1}^{\tau 2} \left| TS_t - AS_t \right| \tag{11.9}$$

where:

TS target state,
AS actual state
τ_1 the first min value of difference between TS and AS
τ_2 the end of optimization interval ($\tau_1 + \tau_s$)
$pen_1 =$ 0 if $\tau_i - \tau_2 \geq \tau_s$
$pen_1 =$ 10*($\tau_i - \tau_2$) if $\tau_i - \tau_2 < \tau_s$ (i.e. late stabilization).

11.8 Results

Analytic Programming requires some EA for its run. In this paper, AP_{meta} version was used. Meta-evolutionary approach means usage of one main evolutionary algorithm for AP process and the second algorithm for coefficient estimation, thus to find optimal values of constants in the evolutionary synthesized control law.

SOMA algorithm was used for the main AP process and DE was used in the second evolutionary process. Settings of EA parameters for both processes given in Tables 11.1 and 11.2 were based on performed numerous experiments with chaotic systems and simulations with AP_{meta}.

The data set for AP required only constants, operators like plus, minus, power and output values x_n and x_{n-1}. The set of elementary functions for AP was inspired in the original delayed feedback chaos control method TDAS (See Sects. 11.6; (11.7) and (11.8)). Thus AP dataset consists only of simple functions (operators) with two arguments and functions with zero arguments, i.e. terminals (constants and system output values). Functions with one argument, e.g. Sin, Cos, etc.; were not required.

Basic set of elementary functions for AP:

GFS2arg = +, −, /, *, ^

Table 11.2 DE settings for meta-evolution

DE Parameter	Value
PopSize	40
F	0.8
CR	0.8
Generations	150
Max. CF Evaluations (CFE)	6,000

Table 11.3 Cost Function values and simple statistics

Experiment no.	CF value	Avg. error per iteration
Experiment 1	$6.2992 \cdot 10^{-15}$	$3.1496 \cdot 10^{-16}$
Experiment 2	$1.4567 \cdot 10^{-6}$	$7.2836 \cdot 10^{-8}$

GFS0arg = $data_{n-1}$ to $data_n$, K

Total number of cost function evaluations for AP was 5,345, for the second EA it was 6,000, together 32.07 millions per each simulation.

Following description of two selected experiments results contains illustrative examples of direct output from AP—synthesized control laws without coefficients estimated (11.10) and (11.12); further the notations with simplification after estimation by means of second algorithm DE (11.11) and (11.13), Table 11.3 with corresponding CF values and the average error value between actual and required system output, and finally Figs. 11.8, 11.9, 11.10 and 11.11 with simulation results.

11.8.1 Experiment 1

$$F_n = 2K_1(-x_{n-1} - 2x_n)(x_{n-1} - x_n)x_n \tag{11.10}$$

$$F_n = 1.18253(-x_{n-1} - 2x_n)(x_{n-1} - x_n)x_n \tag{11.11}$$

11.8.2 Experiment 2

$$F_n = K_1 x_{n-1} - 2(x_{n-1} + K_2)x_{n-1} + x_n x_{n-1} + x_{n-1} \tag{11.12}$$

$$F_n = 16.1492x_{n-1} - 2(x_{n-1} + 7.8473)x_{n-1} + x_n x_{n-1} + x_{n-1} \tag{11.13}$$

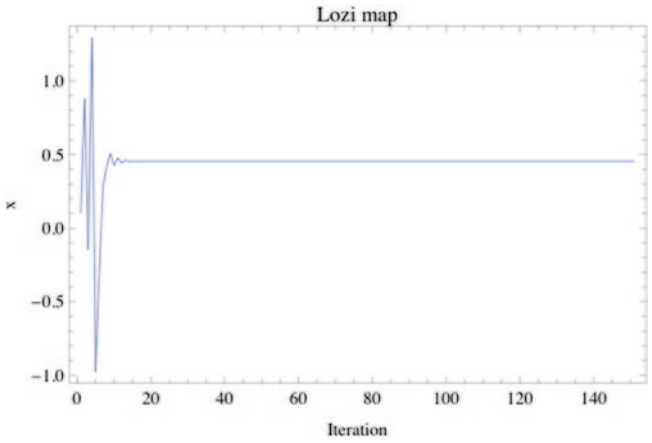


Fig. 11.8 Simulation results—Experiment 1, variable *x* of Lozi map

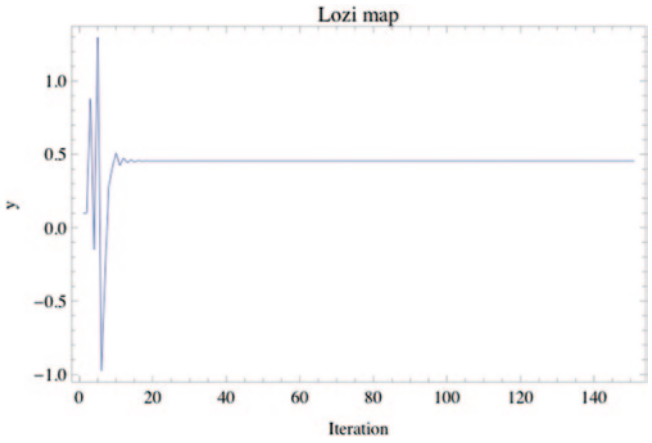


Fig. 11.9 Simulation results—Experiment 1, variable *y* of Lozi map

11.9 Conclusions

This chapter deals with a synthesis of a new universal robust control law by means of AP for stabilization of selected discrete chaotic system at fixed point. Two-dimensional Lozi map as the example of discrete chaotic systems were used in this research.

Obtained results reinforce the argument that AP is able to solve this kind of difficult problems and to produce a new robust synthesized control law in a symbolic

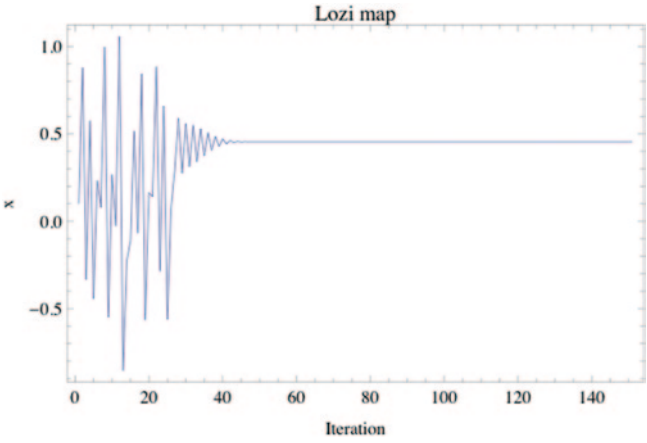


Fig. 11.10 Simulation results—Experiment 2, variable *x* of Lozi map

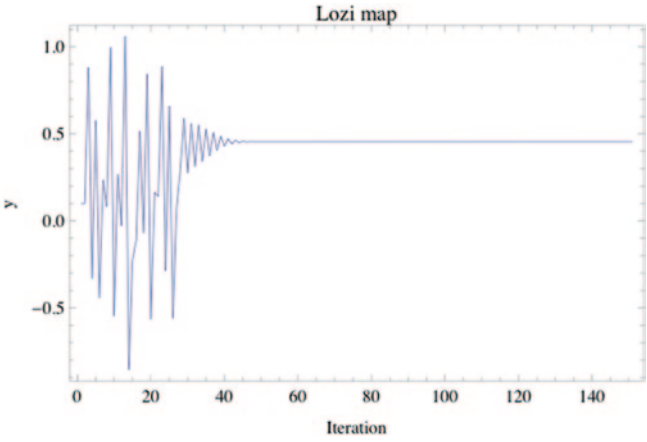


Fig. 11.11 Simulation results—Experiment 2, variable *y* of Lozi map

way securing desired behaviour and precise stabilization of the selected chaotic systems.

Presented two simulation examples show two different results. Extremely precise stabilization and simple control law in the first case and not very precise and slow stabilization and relatively complex notation of chaotic controller in the second case. This fact lends weight to the argument, that AP is a powerful symbolic regression tool, which is able to strictly and precisely follow the rules given by cost function and synthesize any symbolic formula, in the case of this research—the feedback controller for chaotic system.

The future research will include the development of better cost functions, testing of different AP data sets, and performing of numerous simulations to obtain more results and produce better statistics, thus to confirm the robustness of this approach.

Acknowledgements This work was supported by European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089, the project IT4Innovations Centre of Excellence No. CZ.1.05/1.1.00/02.0070, Grant Agency of the Czech Republic: GACR 13-08195S, and by the Development of human resources in research and development of latest soft computing methods and their application in practice project: No. CZ.1.07/2.3.00/20.0072 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic; and by Internal Grant Agency of Tomas Bata University under the project No. IGA/FAI/2014/010.

References

1. Liu B, Wang L, Jin YH, Huang DX, Tang F (2007) Control and synchronization of chaotic systems by differential evolution algorithm. Chaos, Solitons & Fractals, Volume 34, Issue 2, pp. 412–419, ISSN 0960-0779.
2. Zelinka I, Senkerik R, Navratil E (2009) Investigation on evolutionary optimization of chaos control. Chaos, Solitons & Fractals, Volume 40, Issue 1, pp. 111–129.
3. Senkerik R, Zelinka I, Davendra D, Oplatkova Z (2010) Utilization of SOMA and differential evolution for robust stabilization of chaotic Logistic equation. Computers & Mathematics with Applications, Volume 60, Issue 4, pp. 1026–1037.
4. Pyragas K (1995) Control of chaos via extended delay feedback. Physics Letters A, Volume 206, pp. 323–330.
5. Aydin I, Karakose M, Akin E (2010) Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection. Expert Systems with Applications, Vol. 37, No. 7, pp. 5285–5294.
6. Davendra D, Zelinka I, Senkerik R (2010) Chaos driven evolutionary algorithms for the task of PID control. Computers & Mathematics with Applications, Vol. 60, No. 4, pp. 1088–1104, ISSN 0898–1221.
7. Pluhacek M, Senkerik R, Davendra D, Kominkova Oplatkova Z, Zelinka I (2013) On the behavior and performance of chaos driven PSO algorithm with inertia weight. Computers & Mathematics with Applications, Vol. 66, No. 2, pp. 122–134.
8. Just W (1999) Principles of Time Delayed Feedback Control. In: Schuster H.G., Handbook of Chaos Control, Wiley-Vch.
9. Pyragas K (1992) Continuous control of chaos by self-controlling feedback. Physics Letters A, Vol. 170, pp. 421–428.
10. Senkerik R, Kominkova Oplatkova Z, Pluhacek M (2013) Analytic Programming—A Novel Tool for Synthesis of Controller for Chaotic Lozi Map, In: Proceedings of the SCSI 2013, Rhodes Island, GR, EUROPMENT, pp. 201–206, ISBN 978-1-61804-204-0.
11. Senkerik R (2013) Evolutionary Chaos Control—a Brief Survey, In Proceedings of Nostradamus 2012: International conference on prediction, modeling and analysis of complex systems, Springer Series: “Advances in Intelligent Systems and Computing”, Vol. 192, pp. 35–48, ISBN: 978-3-642-33226-5.
12. Senkerik R, Oplatkova Z, Zelinka I, Davendra D, Jasek R (2012) Application of Analytic Programming for Evolutionary Synthesis of Control Law—Introduction of Two Approaches, In: Springer Series “Studies in Computational Intelligence”—“Advances in Intelligent

Modelling and Simulation: Simulation Tools and Applications", (Aleksander Byrski, Zuzana Oplatkova, Marco Carvalho and Marek Kisiel Dorohinicki (Eds.)), pp. 253–268, 2012, ISBN: 978-3-642-28887-6.
13. Senkerik R, Oplatkova Z, Zelinka I, Davendra D, Jasek R (2013) Application of Evolutionary Techniques for Optimization of Chaos Control—Introduction of Three Approaches, In: Springer Series "Inteligent Systems"—"Handbook of Optimization", (Ivan Zelinka, Vaclav Snasel, Ajith Abraham(Eds.)), pp. 801–820, ISBN 978-3-642-30503-0.
14. Senkerik R, Oplatkova Z, Zelinka I, Davendra D (2013) Synthesis of feedback controller for three selected chaotic systems by means of evolutionary techniques: Analytic programming. Mathematical and Computer Modelling, Vol. 57, No. 1–2, 2013, pp. 57–67, ISSN 0895-7177.
15. Kominkova Oplatkova Z, Senkerik R, Zelinka I, Pluhacek M (2013) Analytic programming in the task of evolutionary synthesis of a controller for high order oscillations stabilization of discrete chaotic systems. Computers & Mathematics with Applications, Vol. 66, No. 2, pp. 177–189.
16. Oplatkova Z (2009) Metaevolution: Synthesis of Optimization Algorithms by means of Symbolic Regression and Evolutionary Algorithms. Lambert Academic Publishing Saarbrücken, ISBN: 978-3-8383-1808-0.
17. Zelinka I, Davendra D, Senkerik R, Jasek R, Oplatkova Z (2011) Analytical Programming—a Novel Approach for Evolutionary Synthesis of Symbolic Structures. In: Evolutionary Algorithms, Eisuke Kita (Ed.), InTech, 2011.
18. Zelinka I, Guanrong Ch, Celikovsky S (2008) Chaos Synthesis by Means of Evolutionary algorithms. International Journal of Bifurcation and Chaos, Vol. 18, No. 4, pp. 911–942.
19. Zelinka I (2004) SOMA—Self Organizing Migrating Algorithm, In: New Optimization Techniques in Engineering, (B.V. Babu, G. Onwubolu (eds)), Springer-Verlag, pp. 167–217, ISBN 3-540-20167X.
20. Price K, Storn RM, Lampinen JA (2005) Differential Evolution: A Practical Approach to Global Optimization. Springer.
21. Matousek R, Zampachova E (2011) Promising GAHC and HC12 algorithms in global optimization tasks. Optimization Methods & Software, Vol. 26, No. 3, pp. 405–419.
22. Matousek R (2010) HC12: The Principle of CUDA Implementation. In Proceedings of 16th International Conference On Soft Computing Mendel 2010, pp. 303–308, ISBN 978-80-214-4120-0.
23. Chramcov B, Varacha P (2013) Usage of the Evolutionary Designed Neural Network for Heat Demand Forecast. In: Proceedings of Nostradamus 2012: Modern Methods of Prediction, Modeling and Analysis of Nonlinear Systems, pp. 103–122. ISBN 978-3-642-33226-5.
24. Varacha P, Jasek R (2011) ANN Synthesis for an Agglomeration Heating Power Consumption Approximation. In: Recent Researches in Automatic Control. Montreux: WSEAS Press, pp. 239–244. ISBN 978-1-61804-004-6.
25. Lampinen J, Zelinka I, (1999) New Ideas in Optimization—Mechanical Engineering Design Optimization by Differential Evolution. London: McGraw-hill, pp. 127–146, ISBN 007-709506-5.
26. Price K (1999) An Introduction to Differential Evolution. In: (D. Corne, M. Dorigo and F. Glover, eds.) New Ideas in Optimization, London: McGraw-Hill, pp. 79–108.
27. Price K, Storn R, (2001) "Differential evolution homepage", http://www.icsi.berkeley.edu/~storn/code.html, [Accessed 01/08/2013].
28. Hilborn RC (2000) Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers. Oxford University Press, ISBN: 0-19-850723-2.
29. Sprott JC (2003) Chaos and Time-Series Analysis. Oxford University Press, 2003.

Chapter 12
A Fitter-Population Based Artificial Bee Colony (JA-ABC) Optimization Algorithm

J. Mohamad-Saleh, N. Sulaiman and A. G. Abro

Abstract Inspired by the intelligent foraging behaviour of honeybees swarm, Artificial Bee Colony (ABC) has been introduced by Karagoba in 2005. ABC algorithm has exhibited superior performance compared to other algorithms such as Genetic Algorithm (GA), Differential Evolution (DE) and Particle Swarm Optimization (PSO) algorithms. Despite its outstanding performance, ABC suffers from slow convergence rate and premature convergence. Hence, researchers have proposed various ABC variants but none among the variants could have averted both problems simultaneously. Hence, a new ABC algorithm has been proposed which aims to overcome the limitations. The proposed algorithm focuses on enhancing average fitness of population by mutating poor possible solutions around the fittest solution. The presented results show that the proposed algorithm is capable to avert local optima traps at faster convergence speed.

Keywords ABC variants · Metaheuristic · Swarm intelligence · Computational intelligence

12.1 Introduction

Computational Intelligence (CI) is a sub-branch of Artificial Intelligence (AI). CI is a study of adaptive mechanisms which adapt to new situations for facilitating intelligent behavior in dynamic environments [1]. CI techniques tend to imitate living beings abilities such as decision making, reasoning and optimizing [2]. CI

J. Mohamad-Saleh (✉) · N. Sulaiman
School of Electrical & Electronic Engineering, Universiti Sains Malaysia,
Nibong-Tebal, Penang, Malaysia
e-mail: jms@usm.my

N. Sulaiman
e-mail: noorazlizasulaiman@gmail.com

A. G. Abro
Department of Electrical Engineering, NED University of Engineering & Technology Karachi,
Karachi, Pakistan
e-mail: ghaniabro@gmail.com

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_12,

techniques have been successfully applied in various real-world applications [1]. The techniques include mainly bio-inspired algorithms such as artificial neural networks, evolutionary computation, and swarm-intelligence-based optimization algorithms [1].

Optimization basically refers to make a system or design as effective as possible. Recently, bio-inspired optimization algorithms have been the area of researchers' interest. The algorithms have been applied successfully to solve real-world problems such as real power loss minimization [3], induction motor's parameter estimation [4], controlled islanding of distribution system [5], non-smooth economic dispatch [6], reactive power optimization [7–8] and many more.

Artificial Bee Colony (ABC) is a metaheuristic swarm-intelligence-based optimization algorithm. It has been introduced by Karaboga in 2005 [9]. Since then, it has captured much attention of researchers [10–13]. Researchers have verified that it is competitive with other prominent population-based algorithms such as Genetic Algorithm (GA), Differential Evolution (DE), Particle Swarm Optimization (PSO) algorithm and few other optimization algorithms [14–16]. ABC simulates honeybees' intelligent foraging behavior. It is simple and flexible as it implies lesser number of control parameters than other prominent optimization algorithms [16]. Nevertheless, ABC algorithm also faces few problems, i.e. slow convergence rates and premature convergence [17–18]. This is due to a limitation in the solution-search equation which is focusing more on exploration but lacking in exploitation and have excessive self reinforcement. A robust optimization algorithm should be balanced in terms of exploration and exploitation in order to exhibit good convergence over a range of optimization problems [19].

Due to these problems, researchers have come out with ABC variants [17–23] aimed at overcoming the problems. However, the variants suffer from poor exploration [17, 19, 22], poor exploitation [21], converges slowly [23] and computationally intensive [18, 20]. The flaws of the ABC algorithms motivate towards ABC variant proposed in this research work.

12.2 Artificial Bee Colony (ABC)

ABC optimization algorithm is a population-based optimization algorithm that simulates the foraging behavior of honeybees. Basically, ABC requires three different phases to complete a generation. The phases are; employed-bees, onlooker-bees and scout-bee phases. The task of employed-bees is to explore the neighborhood of the assigned food sources and then share the information with onlooker-bees. The nectar amount of the food sources represents the fitness value of the possible solutions. The employed-bees repeat the process for each of the possible solutions [23] whereas the onlooker-bees select possible solutions which have higher fitness value [22]. Onlooker-bees employ fitness-proportional selection scheme for selecting possible solutions to be updated during onlooker-bees phase. Thus, onlooker-bees

do not update all possible solutions. The employed and onlooker-bees update the assigned possible solutions using the following mutation equation:

$$z_{ij} = y_{ij} + \Phi_{ij}(y_{ij} - y_{kj}) \tag{12.1}$$

where z_{ij} is the candidate solution of food sources, y_{ij} is the *j*-th dimension of the *i*-th food sources and y_{kj} is the *k*-th food sources that are randomly chosen from a neighborhood of *i*-th food sources, $k \in [1, 2\ldots.., SN]$ where SN is the number of food sources. *k* and *i* are mutually exclusive food sources, *k* and *j* are chosen randomly and $j \in [1, 2\ldots.., D]$. D represents the dimension of the search space and $\Phi_{ij}$is the control parameter that represents random number from [−1, 1], inclusively.

A mutation equation is also called search equation. The equation governs the interaction among possible solutions for emerging higher level output. The equation selects a possible solution (y_i) to be updated and also selects another possible solution (y_k) for the mutation. y_k is a randomly chosen possible solution irrespective of its fitness value. If y_k is a fitter possible solution then, there exists higher probability that the candidate solution will be even fitter and vice-versa. As, y_k is randomly selected therefore, the algorithm may consume more mutations for obtaining the optimal solutions. This will lead to slow convergence rate. ABC algorithm employs greedy selection mechanism in order to select between the existing and candidate possible solution during onlooker and employed-bees phases. If the candidate possible solution is fitter than the existing then, ABC selects the candidate possible solution otherwise retains the existing possible solution. A possible solution which does not show improvement over a preset number of generations, is to be abandoned [23]. This number of preset generation is a control parameter called *limit* [19]. The determination of the possible solution that has to be abandoned is done during the scout-bee phase of an ABC algorithm. The scout-bee is an employed-bee whose possible solution is abandoned. Employed-bee becomes scout-bee and will search the environment randomly for discovering a new possible solution to replace the abandoned possible solution [16]. Flow chart of the standard ABC algorithm is given in Fig. 12.1, except the highlighted stages. More details of ABC optimization algorithm can be found in [16].

12.3 Proposed ABC Algorithm

It has been mentioned earlier that the standard ABC algorithm converges slowly because the neighborhood of a selected possible solution is explored on the basis of randomly selected possible solution. Hence, the algorithm consumes more generations or fitness function evaluation (FFEs) for obtaining the desired objective function value. However, the mutation equation has excellent capability to explore the neighborhood for a possible solution. Therefore, if a population of fitter yet diverse possible solutions is generated then the algorithm may converge faster.

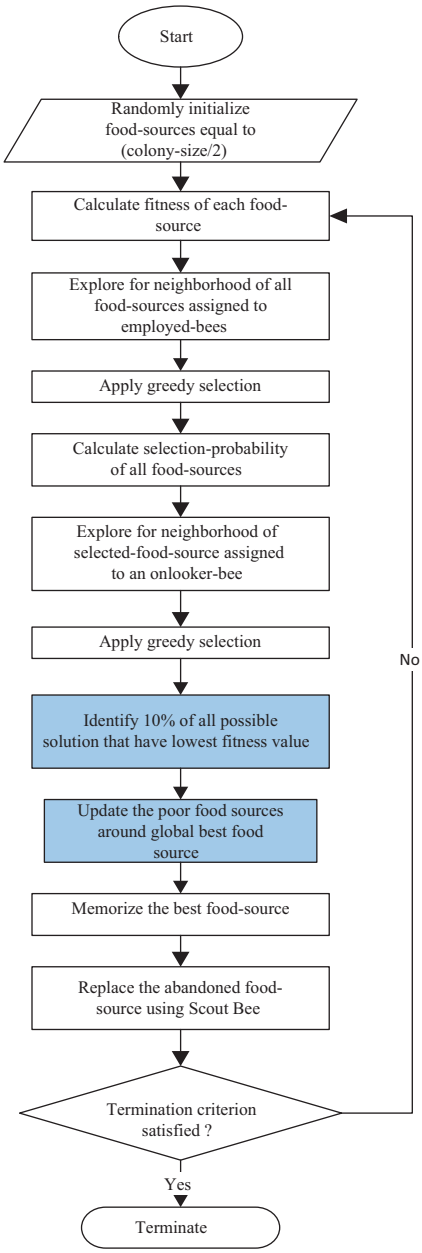


Fig. 12.1 Modified ABC (JA-ABC) algorithm

The algorithm proposed in this research works by mutating poor possible solutions around global best (gbest) possible solution, at the end of every generation. This increases average fitness value of the population. ABC algorithm having a fitter yet diverse population of possible solutions may be able to converge faster and avoid premature convergence.

The proposed algorithm has been named JA-ABC optimization algorithm. The flow chart of the proposed algorithm is shown in Fig. 12.1. The proposed phases are highlighted. In these phases, 10 % of all possible solutions are to be updated, which have the lowest fitness value. Hence, the proposed phases only update poor possible solutions. The mutation equation for the proposed phases is:

$$z_{ij} = y_{best,j} + \Phi_{ij}(y_{pj} - y_{kj}) \tag{12.2}$$

where z_{ij} is the candidate solution of new food sources, $y_{best,j}$ is the global best food source with j-th dimension, y_{pj} is the p-th food sources of j-th dimension and y_{kj} is the k-th food sources of j-th dimension. p and k are randomly chosen food sources and they are mutually exclusive. Meanwhile the parameter Φ_{ij} is a control parameter that represents random numbers within $[-1, 1]$.

As poor possible solutions are mutated around the gbest possible solution, the modified poor possible solutions would be fitter. This way, the number of fit possible solutions increases with increasing generation. Now, there exist higher probability that a selected possible solution (y_i) will be mutated with a fit possible solution (y_k) during employed and onlooker-bees phases, as fitness of every possible solution is higher in the proposed algorithm. Hence, the produced candidate solution will be fitter than the existing possible solution. Therefore, the algorithm may converge faster.

12.4 Experimental Setup

The proposed ABC algorithm (JA-ABC) has been tested on Griewank (f_1), Rastrigin (f_2), Rosenbrock (f_3) Ackley (f_4), Schwefel (f_5), Himmelblau (f_6), Sphere (f_7), Step (f_8), Bohachevsky 2 (f_9) and Schwelfel's 2.22 (f_{10}) benchmark functions. The input dimension of the benchmark functions has been set to 30. The performance of the proposed algorithm has been compared with the standard ABC (ABC) [16], gbest-guided ABC (GABC) [23] and best-so-far ABC (BsfABC) [20]. For all algorithms, the population size has been set to 50, number of generation has been limited to 1,000 and parameter *limit* has been set as 1,500. As for GABC, *C-value* has been set to 1.5. Each algorithm has been run for 30 times on each benchmark function to ensure validity of the global solution [17].

12.5 Results and Discussion

The obtained results illustrate the superiority of the proposed algorithm compared to others, on all benchmark functions particularly on f_2 and f_5. The comparison of the algorithms is illustrated by Fig. 12.2. The figures clearly illustrate better convergence rates of the proposed algorithm compared to the other ABC algorithms on the

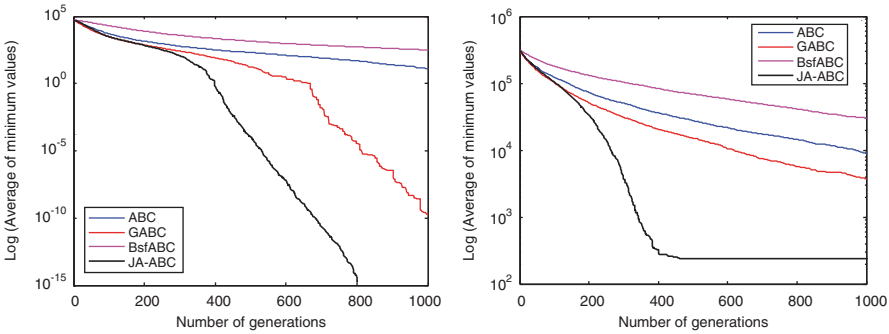


Fig. 12.2 Convergence rates optimization algorithm on *f2* and *f5*

benchmark functions. This clearly verifies the robustness of the proposed algorithm as compared to the others algorithms.

BsfABC algorithm has yielded the worst performance compared to other algorithms since mutation equation of BsfABC algorithm during onlooker-bees phase is highly local in nature. GABC algorithm seems to perform better than BsfABC and ABC algorithm. However, GABC shows inferior performance compared to the proposed algorithm since it suffers from slow convergence rate [18].

The results summarized in Table 12.1 show the average and standard deviation among 30 runs for each algorithm. The results reveal that the performance of the proposed algorithm is the best among other compared optimization algorithms. Therefore, it can be concluded that the proposed algorithm has the ability to locate global optimum compared to the other optimization algorithms.

12.6 Conclusions

A new ABC algorithm named JA-ABC has been proposed in this research work. It is based on enhancing average fitness of population by mutating poor possible solutions around the gbest possible solution. This leads to fitter yet diverse population. Thus, the algorithm possesses higher convergence rate and the ability to avoid premature convergence efficiently. The proposed algorithm has been compared with the existing ABC variants on ten commonly used benchmark functions. The presented results have shown the superior performance of the proposed algorithm to the other compared algorithms on all of the benchmark functions. Therefore, it can be concluded that theoretically the proposed algorithm can be successfully applied to any optimization problems.

Acknowledgements The authors acknowledge Universiti Sains Malaysia (USM) RU-PRGS No: 1001/PELECT/8036007 and USM Short-Term Grant No: 304/PECECT/60311038 for the financial support.

Table 12.1 Performance results of the optimization algorithms

Func.	Algorithm	Average	Std. Deviation	Func.	Algorithm	Average	Std. Deviation
f1	ABC	1.47E−13	1.65E−13	*f6*	ABC	−7.83E+01	1.55E−07
	GABC	5.59E−16	7.42E−17		GABC	−7.83E+01	5.90E−15
	BsfABC	1.69E−05	1.30E−05		BsfABC	−7.64E+01	7.09E−01
	JA-ABC	**6.11E−16**	**1.16E−16**		**JA-ABC**	**−7.83E+01**	**1.72E−01**
f2	ABC	3.77E−01	5.87E−01	*f7*	ABC	4.53E−10	3.99E−10
	GABC	6.35E−12	1.76E−11		GABC	6.07E−16	1.03E−16
	BsfABC	1.07E+01	1.71E+00		BsfABC	7.16E−02	5.62E−02
	JA-ABC	**0.00E+00**	**0.00E+00**		**JA-ABC**	**6.66E−16**	**8.21E−17**
f3	ABC	9.03E−01	1.11E+00	*f8*	ABC	4.82E−10	5.12E−10
	GABC	4.96E+00	1.03E+01		GABC	5.73E−16	1.10E−16
	BsfABC	5.27E+01	1.67E+01		BsfABC	7.79E−02	5.13E−02
	JA-ABC	**9.79E−02**	**1.95E−01**		**JA-ABC**	**6.55E−16**	**9.46E−17**
f4	ABC	1.82E−05	1.12E−05	*f9*	ABC	3.85E−08	4.32E−08
	GABC	1.76E−10	8.73E−11		GABC	5.26E−16	1.27E−16
	BsfABC	7.95E−01	3.69E−01		BsfABC	9.72E−01	4.83E−01
	JA-ABC	**5.34E−13**	**2.47E−13**		**JA-ABC**	**2.76E−16**	**1.68E−16**
f5	ABC	2.96E+02	1.20E+02	*f10*	ABC	1.07E−05	5.44E−06
	GABC	1.23E+02	1.19E+02		GABC	1.60E−10	6.40E−11
	BsfABC	1.00E+03	2.12E+02		BsfABC	1.13E−01	3.94E−02
	JA-ABC	**3.82E−04**	**8.72E−13**		**JA-ABC**	**9.77E−14**	**6.36E−14**

References

1. Engelbrecht AP (2007) Computational intelligence: an introduction. Wiley
2. Sun H-C, Huang Y-C, Huang C-M (2012) Fault Diagnosis of Power Transformers Using Computational Intelligence: A Review. Energy Procedia, 14(0): p. 1226–1231
3. Badar AQH, Umre BS, Junghare AS (2012) Reactive power control using dynamic Particle Swarm Optimization for real power loss minimization. International Journal of Electrical Power & Energy Systems, 41(1): p. 133–136
4. Abro AG, Mohamad-Saleh J (2013) Multiple-global-best based artificial bee colony algorithm for parameter estimation of induction motor Turkish Journal of Electrical Engineering & Computer Sciences, DOI: 10.3906/elk-1209-23
5. El-Zonkoly A, Saad M, Khalil R (2013) New algorithm based on CLPSO for controlled islanding of distribution systems. International Journal of Electrical Power & Energy Systems, 45(1): p. 391–403
6. Niknam T, Mojarrad HD, Meymand HZ, Firouzi BB (2011) A new honey bee mating optimization algorithm for non-smooth economic dispatch. Energy, 36(2): p. 896–908
7. Ayan K, Kılıç U (2012) Artificial bee colony algorithm solution for optimal reactive power flow. Applied Soft Computing, 12(5): p. 1477–1482
8. Zeng X-j, Tao J, Zhang P, Pan H, Wang Y-Y (2012) Reactive Power Optimization of Wind Farm based on Improved Genetic Algorithm. Energy Procedia, 14: p. 1362–1367
9. Karaboga D (2005) An Idea Based on Honey Bee Swarm For Numerical Optimization. Technical Report-TR06
10. Karaboga, D, Ozturk C, Karaboga N, Gorkemli B (2012) Artificial bee colony programming for symbolic regression. Information Sciences, **209**(0): p. 1–15

11. Karaboga N, Latifoglu F (2013) Adaptive filtering noisy transcranial Doppler signal by using artificial bee colony algorithm. Engineering Applications of Artificial Intelligence, **26**(2): p. 677–684
12. Rodriguez FJ, Lozano M, García-Martínez C, González-Barrera, JD (2013) An artificial bee colony algorithm for the maximally diverse grouping problem. Information Sciences, **230**(0): p. 183–196
13. Samanta S, Chakraborty S (2011) Parametric optimization of some non-traditional machining processes using artificial bee colony algorithm. Engineering Applications of Artificial Intelligence, **24**(6): p. 946–957
14. Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of Global Optimization, **39**(3): p. 459–471
15. Karaboga D, Basturk B (2008) On the performance of artificial bee colony (ABC) algorithm. Applied Soft Computing, **8**(1): p. 687–697
16. Karaboga D, Akay B (2009) A comparative study of Artificial Bee Colony algorithm. Applied Mathematics and Computation, **214**(1): p. 108–132
17. Abro AG, Mohamad-Saleh J (2012) Enhanced Global-Best Artificial Bee Colony Optimization Algorithm. in 2012 Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation (EMS)
18. Abro AG, Mohamad-Saleh J (2012) An Enhanced Artificial Bee Colony Optimization Algorithm. Recent Advances in Systems Science and Mathematical Modelling, ed. D.S. Nikos Mastorakis, Valeriu Prepelita: WSEAS Press
19. Gao W, Liu S, Huang L (2012) A global best artificial bee colony algorithm for global optimization. Journal of Computational and Applied Mathematics, **236**(11): p. 2741–2753
20. Banharnsakun A, Achalakul T, Sirinaovakul B (2011) The best-so-far selection in Artificial Bee Colony algorithm. Applied Soft Computing, **11**(2): p. 2888–2901
21. Gao W, Liu S (2011) Improved artificial bee colony algorithm for global optimization. Information Processing Letters, **111**(17): p. 871–882
22. Gao W-F, Liu S-Y (2012) A modified artificial bee colony algorithm. Computers & Operations Research, **39**(3): p. 687–697
23. Zhu G, Kwong S (2010) Gbest-guided artificial bee colony algorithm for numerical function optimization. Applied Mathematics and Computation, **217**(7): p. 3166–3173

Chapter 13
Modeling the Value Chain with Object-Valued Petri Nets

J. Zacek, Z. Melis and F. Hunka

Abstract A substantial part of the economic theories is based on conversion and exchange process. These processes can be arranged in a value chain, which can be considered as a cyclic model with complex attributes. There is a serious problem how to express resources and their conversions in a complex cyclic model during the simulation and how to identify these converted resources in every step of the simulation. This paper introduces the Object- valued Petri (OV-PN) modification as a new formalism to create a cyclic model of the value chain. According to the modification we had to define a new path and pass of the OV-PN. We also had to determine new properties. Properties are based on the OV-PN and reflect needs of model requirements. A new formalism is verified on a common enterprise value chain.

Keywords Value chain · Object-valued Petri nets · Cycle Petri nets · Simulation · Model validation

13.1 Introduction

The value chain is a modeling technique to formalize and monitor the competitiveness of the business. It focuses on the flow of resources between internal business processes that are interconnected to each other. A product increases its value when it passes through a flow of production chain. That is the fundamental notion in

J. Zacek (✉)
Centre of Excellence IT4Innovations, Division of the University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, 30. dubna 22, 70103 Ostrava, Czech Republic
e-mail: jaroslav.zacek@osu.cz

Z. Melis · F. Hunka
Faculty of Science, Department of Computer Science, University of Ostrava, 30. dubna 22, 70103 Ostrava, Czech Republic
e-mail: zmelis@seznam.cz

F. Hunka
e-mail: frantisek.hunka@osu.cz

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_13,

value chain analysis [7]. The REA value chain is a network of the REA exchange and conversion processes. The purpose of the network is to directly or indirectly contribute to the creation of the desired features of the final product or service, and to exchange it with other economic agents for a resource that has a greater value for the enterprise [6]. The value chain definition implies that it is important to find a suitable formalism for the simulation of the model run for practical realization. Existing theories such as state machines, Petri nets [12], or neural networks were considered while searching for a correct formalism. Value chains have specific requirements for descriptive formalism allowing their validation and simulation (according to [1]). Specific type of the value chain is supply chain [8]. State machines are not expressive enough to solve this problem. Despite the fact that neural networks are expressive enough for describing processes in the value chain there is significant complexity in simulation. Two independent neural networks must be created for the simulation of the value chain. The first network is able to validate the model and the second implements simulation steps. In both cases, the neural network must learn these properties. Therefore the process becomes time and implementation consuming. On the other hand the neural network approach is very flexible and can be used to solve multilevel problems (for example multilevel SPAM control [13]).

The Petri Net theory matches the description of the model states more closely, but its expressivity, especially for P/T Petri nets, is very limited [10]. General token is not able to capture such a complex structure, for example an object representing the resource. Therefore this article suggests using the object- valued Petri nets (OV-PN) to ensure the simulation and the validation of value chains. It also discusses some specific properties of the OV-PN and defines new properties for the value chain domain. Main advantage of using the Petri net theory is possibility to create an automatic deterministic process of the code generation [4].

13.2 The Value Chain and its Simulation Process

The value chain consists of two main parts: processes and links that form a chain with other processes (similar to supply chain described in [15]). These parts create the interconnected network of processes increasing the value of the resources. The value chain creates a cyclic bond that means all processes have their inputs and outputs connected together and form a full closed chain. Each process can have more than one input and more than one output. Multiple types of resources can form the input and the output. In Fig. 13.1 there are two significant examples of resource distribution. Resources *Plan* and *Money*, needed for purchasing the *Material*, enter to the *Purchase process*. *Sales process* produces output *Money* that enters into two another processes—*Purchase process* and *Acquisition process*.

The simulation of the value chain process is used to monitor the competitiveness of the business. Model elements show processes that increasing the value of corporate resources. Each step of the process increases the value of the company output, and therefore it can be understood as a value chain [5]. The value chain is

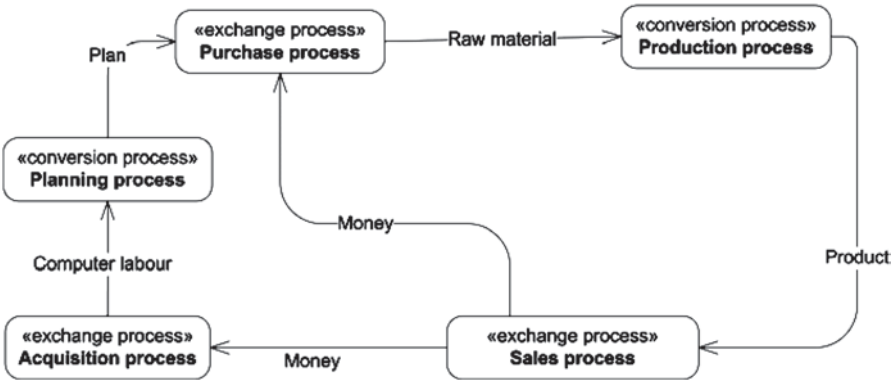


Fig. 13.1 Enterprise value chain

the set of mutually interdependent activities that are interrelated together by their inputs and outputs.

Enterprise value system shows the flow of resources between participants within the enterprise and it can be obtained by analysis of the company value chains. The value chain shows the flow of resources across business processes [14].

Figure 13.1 shows an example of the value chain depicting the flow of resources between the enterprise and business partners. The model example consists of 5 processes:

- *Purchase process* expresses the purchase of the raw material from vendors.
- *Production process* is internal conversion process creating the product.
- *Sales process* illustrates the sales of the created product to the customer.
- *Acquisition process* arranges labour for *Planning process*.
- *Planning process* prepares purchase plan.

For purposes of the planning and detailed analysis of business value chains it is necessary to record the flow of resources and their changes in time. For these cases, it is possible to use the simulation of the flow of resources across business processes. In every step of the simulation an exchange process performs exchanging of resources and the conversion process creates a new product or modifies characteristics of an existing product. In case of complex value chain the simulation can determine which links are inefficient and where is the place for the subsequent optimization. The simulation can be also used for analysis of the economic situation of the enterprise, such as stores status, financial estimates, an efficiency of the production, or logistics. You can also simulate the way of a product from an initial purchase of raw materials, through its production and finishing by sale to determine the total financial and time costs for one product.

Object-valued Petri net is an extension of P/T Petri net. This extension has been introduced in [17]. Object-valued Petri nets are used as formalism for validation and synchronization of complex object models.

Definition 3.1: Object-valued Petri net *Petri net is extended to a 6-tuple (P, T, F, V, R, C), where*:

- *P is a finite nonempty set of places*,
- *N is a finite nonempty set of transitions*,
- $T \cap P = \varnothing$ (*P and T are disjoint*),
- $F \subseteq (PxT) \cup (TxP)$ *is a finite set of arcs (flow relation)*,
- *V is a finite set of object data types*,
- *R is a finite set of transforming functions* $R : P \cup T \to \psi(V)$, *where* $\psi(V)$ *is the power of the set of object data types*.
- *C is a set of capacity function.* $C : P \to N \cup \omega, N \subseteq \mathrm{N}$ *and* ω *denotes infinite*.
- $M_0 : P \to V_{MS}$ *is the initial marking of the token.* $\forall p \in P : M_0(p) \in R(p)_{MS}$, *where R(p)MS, is the multiple set of the object data type tokens in p*.

The main idea of the Object-valued Petri net is an object-valued token that provides adequate expressivity to describe resources represented by complex object structures. The token carries basic information to identify the specific object instance. Initial marking consists of the multiple set of object data types deployed across the net. Firing of each token means change in marking of the net and also change of the token type. However token identification remains and therefore we can identify the token in every step of the simulation process. If the model is partly linked with the Object-valued Petri net theory we have to define the path of tokens. Formalism itself defines necessary basis to create the model, unfortunately that does not ensures the sequence of movements into desirable result. Object-valued Petri net realizes transition as soon as the transition is feasible. Nevertheless the real model can require other conditions to realize the transition (for instance lazy constructions). Therefore we have to state the new definition of the path and pass of the model.

Definition 3.2: Path of the OV-PN *Let OV-PN=(P, T, F, V, R, C) be an Object-valued Petri net with initial marking* M_0. *The path from the place* $u_1 \in P \cup T$ *to following place* $u_n \in P \cup T$ *is the sequence* $(u_1, u_2, ..., u_n)$, *where* (u_i, u_{i-1}), *for* $1 \le i \le n$.

Definition 3.3: Pass of the OV-PN *Object-valued Petri net OV-PN=(P, T, F, V, R, C) with initial marking* M_0 *is feasible when*:

1. *Must exist an initial place* $i \in P$ *where* $\bullet i = \varnothing$.
2. *Must exist exactly one final place* $i \in P$ *where* $o\bullet = \varnothing$.
3. *Every place* $u \in P \cup T$ *lies on the pass between initial place* ***i*** *and final place* ***o***.

In this context the first condition is understood as a marking of the input of the model that can be represented by more than one input parameter. If this condition is set to be strict to the value of input marks the model cannot realize calling of the method with more than one input parameter. The output of the model is usually one because of the standard method construction in object-oriented paradigm [11]. Third condition expresses the fact that every place and every transition exists on the path between the initial place and the final place. Therefore the Object-valued Petri net should not have blind paths and every call in the model should be reachable form initial place by passing finite number of transitions representing a flow

relation *F (similarly to* [3]*)*. Similarly to the initial place, every place in the model exists in the flow relation *F* and is able to reach the final place of the model.

Boundedness and Safeness The ordinary Petri net (P/T Petri net) defines the boudedness mechanism to limit the tokens in all reachable markings. A place in the Petri net is called *k*-bounded if it does not contain more than *k* tokens in every making in the net, including the initial marking. Moreover the Petri net is bounded if and only if its reachability graph is finite. The special case of the boundedness is safeness attribute. If the net is 1-bounded it is called *safe*.

Object model synchronized by Petri net mechanism can be bounded at the places level as the ordinary Petri net. Every method can produce more than one output during the simulation. Places may store these outputs as Object-valued tokens (similar to colour evaluation in [18]). By applying safeness rule the places in model stores only one object-valued token and the model becomes less complex.

Conservation Created model cannot be strictly conservative. In the first step in Fig. 13.1 the *Purchase process* consumes two inputs and produces one output. The *Sales process* consumes only one input and produces two outputs—two object-valued tokens parameterized as *Money*. Moreover the *Purchase process* requires a synchronization mechanism. The model cannot have a constant token count for every marking from set of the reachability set.

$$\Re(M_0): \sum_{\rho_i \in S} M(P_i) \neq \sum_{\rho_i \in S} M_0(P_i).$$

Liveness and Deadlock All methods in object-oriented paradigm can be executed more than once [16]. However by executing some method an internal state of the object can be altered. That means if we need to apply liveness property to whole model, every method must be considered as an atomic operation.

Generally the transition $t \in T$ *is alive if*:

$$\forall p \in \bullet t : M(p) \neq \varnothing \text{ and } p \in t\bullet : M(p) = \varnothing.$$

It means that transition becomes active, if there are tokens on all transition's entrances and the place that follows the transition is empty. The net is alive if there is at least one live transition in every step of simulation process otherwise a deadlock occurs. Deadlock is solved on a higher abstraction level and requires user intervention.

13.3 Object-Valued Petri Net Extensions

The main condition of the value chain is cyclicality. However the Object-valued Petri net has two definitions that limit the path of the net (definition 3.2) and pass of the net (definition 3.3). First definition says that OV-PN with a specific marking M_0 has a specific sequence from one place to another. The value chain has also specific

sequence that defines the path of the chain. Moreover the cyclic chain consists of many single paths connected to each other. To express a general value chain principle with the Petri net theory we have to define a cyclic Petri net:

Definition 4.1: Cyclic Object-valued Petri net *A marked Petri net (OV-PN, M_0) is cyclic Petri net if from every reachable marking M it is possible to return into M_0 (i.e. $M \in \Re(OV-PN, M_0) \Rightarrow M_0 \in \Re(OV-PN, M)$).*

According to [2] we must also define the inverse of an ordinary Petri net:

Definition 4.2: The inverse of an Object-valued Petri net *For a Petri net OV-PN, its inverse $\overline{OV-PN} = \left(P, \overline{T}, \overline{F}\right)$ is given by*:

- $T = \left\{\bar{t} \middle| t \in T\right\}$ and
- $\overline{F}\left(p, \bar{t}\right) = F(t, p)$ *and* $\overline{F}\left(p, \bar{t}\right) = F(t, p)$ *for every* $p \in P$ *and* $t \in T$.

The definition of the inversion of the Object-valued Petri net is presented for completeness only. In the real model of the value chain there is usually no backward path. For instance the company cannot convert the product to the raw material. On the other hand this definition gives the robust tool to verify cyclicality of the net. Algorithms to verify cyclicality could be simplified to perform the token verification. Every token in the Object-valued Petri net have the unique instantiation number. The inner value of the Object-valued token is changed during the pass of the net, however instantiation number stays unchanged despite the value transformation. The modeling tool can set up the initial marking and make finite steps of the firing. If the net is cyclic the specific instantiation returns to the initial marking.

According to the facts above we can redefine the original Object-valued Petri net tuple for modelling the value chain:

Definition 4.3: Petri net for a value chain simulation *Petri net for value chain simulation is a 5-tuple (P, T, F, S, R), where*:

- *P is a finite nonempty set of places*,
- *T is a finite nonempty set of transitions*,
- $T \cap P = \varnothing$ *(P and T are disjoint),*
- $F \subseteq (PxT) \cup (TxP)$ *is a finite set of arcs (flow relation)*,
- *S is a finite set of resources*,
- *R is a finite set of transforming functions* $R : P \cup T \to \psi(S)$, *where* $\psi(S)$ *is the power of the set of resources*.

Moreover we must claim boundedness of elements:

- $\forall p \in P \exists t \in T : F(p,t) \in F$,
- $\forall p \in P \exists t \in T : F(t,p) \in F$,
- $\forall p \in T \exists p \in P : F(p,t) \in F$,
- $\forall p \in T \exists p \in P : F(t,p) \in F$

and their connection to cyclic model:

- $\forall p \in P : M(p) \in \Re(M_0), M_0 \in \Re(M(p))$,
- *every* $t \in T$ *is reachable from any* $p \in P$ *in limited count of steps*.

Naturally we also must specify the properties of the new definition:

Boundedness and Safeness In the value chain the one resource can be transferred into the more than one process (i.e. money to buy a new material and money to fund innovations).

The safeness of the net is the matter of discussion. In fact there are two possibilities. The net can be safe and that means one place stores only one object-valued token. That property can be convenient to verify the whole conversion process and user can focus to one resource and transformation process. This simulation is similar to redefining business processes in the company. By applying the safeness property the whole model became simple to understand and verification of the process flow is much easier.

The second view on the value chain simulation is to get statistic data and optimize workflow parameters. The model must simulate the conversion process with more than one Object-valued token. A typical example is a production process creating the specific product. At the beginning of the simulation the company needs to know how many products must be created to cover money for a product development. In the short term the first view can set the margin of the seller and express the production process. The simulation of the second view takes longer and works with multiple tokens. The price of the product decreases with time and by the long term simulation the company can reveal if the price model has been set correctly. Therefore the second view cannot be safe form a Petri net point of view.

Liveness and Deadlock Object-valued Petri net must fulfill liveness property because of the object-oriented paradigm construction. Cyclic Petri nets are based on general OV-PN, but it differs on boudedness and safeness property. Therefore liveness property must be changed.

Transition $t \in T$ *is alive if*:

- $\forall p \in \bullet t : M(p) \neq \varnothing$
- $p \in t\bullet : |M(p)| + U(t) \leq K(p)$, *where* $U(t)$, *is a number of resources produced by transition t*

The net is live if there is at least one live transition in every step of the simulation process.

The value chain consists of processes and links. Process itself consists of atomic operations that can be repeated infinitely with the same result. For example: a production process is defined by precise methodology how to produce a product on the serial assembly. The parameters of the process are set at the beginning of the serial assembly (i.e. speed of the line) and usually remains unchanged during production. Therefore we naturally apply the liveness property to the Object-Valued Petri net model of the value chain. All processes remain the same despite the fact that the Object-valued token flow through the process.

The process itself can have more than one input link. In Fig. 13.1 the *Purchasing process* requires the *Money* and the *Plan*. *Plan process* takes more time to create a specific *purchasing plan* and *sales process* delivers the *Money* immediately after the product has been sold. In this specific step of the model simulation the deadlock occurs. That means the execution of the *Purchasing process* is delayed until both inputs provided with links are available. These cases can be problematic and generally can be solved on a higher abstraction level, i.e. modelling tool. If the model is validated a deadlock cannot occur because of the cyclicality property of the value chain. Moreover that implies that every process must be reachable.

Conservation The conservation property means that one object-valued token cannot be duplicated when it passes the transition and the transition has the same number of inputs and outputs. In other words the count of the Object-valued tokens is same in every step of the simulation. The basic models of the value chain can be conservative. However most models in the real world are more complex and there is big challenge to apply the conservation rule to express a chain of resources. For example money in the real world is an input to more than one process—production process, planning, development, etc. From a Petri net point of view we must duplicate tokens with specific inner attributes and sends them to other transitions. Therefore the model does not have the constant token count for every marking from the reachability set.

13.4 Value Chain Simulation Example

During the transformation the value chain elements are mapped into the modified cyclic Object-Valued Petri net elements. Similar transformation process can be found in [9]. Processes that exist in the value chain will be represented as transitions and all properties mentioned above will be applied. Links that exist in the value chain will be composed of two arcs and one place. The Arc indicates the direction of an Object-valued token and the Place carries the Object-valued token(s) that represents information.

For example we used the value chain from Fig. 13.1. The transformed result is shown in Fig. 13.2. The Petri Net consists of five transitions and six places. The simulation starts with the Token set on the place *Product* and it takes five cycles before repeating:

1. The company has product to sell. The *Product* enters into the *Sales process*. That generates *Money* for the *Purchase process* and the *Acquisition process* (token is divided into two places).
2. *Money* enters into the *Acquisition process* and creates the *Computer work*. The *Purchase process* is not executed because of insufficient *Plan* input.
3. The *Computer work* enters into the *Planning process* and generates the *Plan*.
4. The *Purchase process* transition has all needed inputs to perform firing. *Money* and *Plan* tokens are exchanged for *Material*.

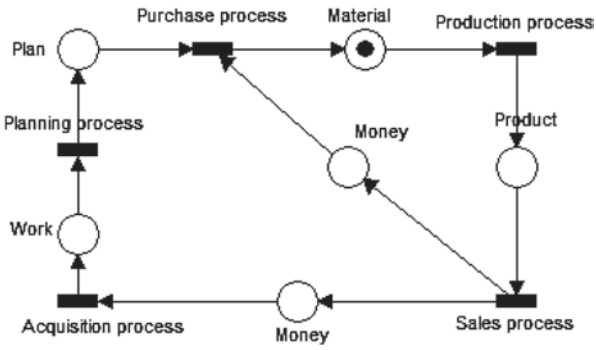


Fig. 13.2 Transformed value chain

5. In the last step the *Material* enters into the *Production process* and creates a new *Product*.

The model is cyclic. That means the cycle 1 does not have to be the first and all steps are realized infinitely. The order is only that matters. The key in the value chain simulation process, except the synchronization primitives, is the Object-valued token. In the beginning of the simulation the Object-valued token is parameterized as a *Product*. In the first step the token is transformed into the *Money* and split into the two tokens with specific ratio used to determine the inner value. The association of the tokens to the value chain can be identified through the instantiation ID, and inner values can be changed as needed. Moreover in the second cycle, one of tokens enters to *Acquisition process* and transforms (parameterizes) into the *Computer work*. That means the token has different inner values and even a data structure. Analogical transformation changes the *Computer work* into the *Plan* structure in the third cycle. In the fourth cycle two tokens are merged together by *Purchase process* and the result is the Object-valued token parameterized as a *Material*. The merge process can be performed because of the same token instantiation ID. In the last step the token is transformed to the *Product* by the *Production process* and the chain is closed.

There is only one token in Fig. 13.2 and all links are limited to 1. We can simulate the whole conversion process with more than one token and we can establish a capacity function on every place in the net. All splits and merges of tokens are identified by instantiation ID and therefore they are distinguishable. That means we can recognize the specific token as a part of the cyclic chain and the base for optimization of processes in the model.

13.5 Conclusion

The paper introduced a new formalism based on the Object-Valued Petri net to create, synchronize and manage cyclic models of the value chain. The paper described a basic theory of economic models based on conversion and exchange processes

and introduced a value chain term in the first part. The paper also described why current formalisms such as neural network and state machines are not suitable to build a value chain model. Paragraph 3 shows an Object-value Petri net theory focused on the path and pass of the net. This theory is suitable to build a value chain model because the Object-valued token can be used to express resources of the value chain and their transformations. However Object-valued Petri net are not cyclic and have strictly defined pass and path of the net. The definition of an extended Object-valued Petri net formalism—definition 4.3—solves this problem and adds the cyclicality. All basic properties are discussed and redefined for the new cyclic object-oriented model. An extended Object-Valued Petri net formalism solves all problems mentioned in the second paragraph and can be applied to any cyclic model. The proposed formalism has been verified on the ordinary value chain and basic steps of the simulation are described in paragraph 4.

Acknowledgements This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and the grant reference no. SGS08/PRF/2013 provided by Ministry of Education, Youth and Sports.

References

1. Bocewicz G., Wójcik R., Banaszak Z.: Cyclic Scheduling for Supply Chain Network, Trends in Practical Applications of Agents and Multiagent Systems, Springer Berlin Heidelberg, pp 39–47, 2012.
2. Bouziane, Z., Finkel, A.: Cyclic Petri Net Reachability Sets are Semi-linear Effectively Constructible, CONCUR'11 Proceedings of the 22nd international conference on Concurrency theory, Springer-Verlag, Berlin, ISBN: 978-3-642-23216-9.
3. Darondeau P., Demri S., Meyer R., Morvan Ch.: Petri Net Reachability Graphs: Decidability Status of First Order Properties, Logical Methods in Computer Science, Vol. 8(4:9), pp. 1–28, 2012.
4. Ding Z., Liu J., Wang J.: A Petri Net Based Automatic Executable Code Generation Method for Web Service Composition, Proceedings of the 2012 International Conference on Information Technology and Software Engineering, pp 39–48, ISBN 978-3-642-34530-2, 2013.
5. Fiala, J., Ministr, J.: Pruvodce analyzou a modelovanim procesu, VB-TU Ostrava, 2003, ISBN 20-248-0500-6.
6. Geerts, G. L., McCarthy, W. E.: Using Object Oriented Templates from the REA Accounting Model to Engineer Business Process and Tasks, Paper presented at European Accounting Congress, Gratz, Austria, 1997.
7. Hunka, F., Zacek, J., Melis, Z., Sevcik, J.: REA Value Chain versus Supply Chain, Scientific Papers of the University of Pardubice, 2011, s. 68–77, ISSN 1211-555X.
8. Kersten W., Blecker T., Ringle M. Ch.: Managing the Future of Supply Chain, Eul Verlag, ISBN 978-3-8441-0180-5, 2012.
9. Li J., Zhou M., Xianzhong D.: Reduction and Refinement by Algebraic Operations for Petri Net Transformation, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on (Volume:42, Issue: 5), pp 1244–1255, ISSN 1083-4427, 2012.
10. Martinik, I.: Methodology of object-oriented programmatic system development using theory of Object Petri nets, dissertation thesis, VB-TU Ostrava, 1999.
11. Shilling, J.: Three Steps to Views: Extending the Object-Oriented Paradigm, OOPSLA 89 Proceedings, 1989.

12. Sklenar, J., Caruana, E.: Using Timed Petri Nets in Discrete Simulation, Proceedings of the Industrial Simulation Conference, 2004 (ISC-2004), Malaga, 2004, page 7–11.
13. Sochor T., Davidova A.: Potential of Multilevel SPAM Protection in the Light of Current SPAM Trends, 10th International Conference on Networking, Sensing and Control, Paris-Evry University, France, 2013.
14. van Hee M. K., Sidorova N., van der Werf M. J.: Business Process Modeling Using Petri Nets, Transactions on Petri Nets and Other Models of Concurrency VII, pp 116–161, ISBN 978-3-642-38142-3, 2013
15. Wang W. J., Ip H. W., Muddada R. R., Huang L. J., Zhang J. W.: On Petri net implementation of proactive resilient holistic supply chain networks, The International Journal of Advanced Manufacturing Technology, ISSN 0268-3768, Springer-Verlag, 2013.
16. Zacek J., Hunka F.: CEM: Class executing modeling, World Conference on Information Technology, 2010, page 1597–1601, ISSN 1877-0509.
17. Zacek, J., Hunka, F.: Object model synchronization based on Petri net, 17th International Conference on Soft Computing MENDEL 2011, Brno University of Technology, Faculty of Mechanical Engineering, 2011, s. 523–527, ISBN 978-80-214-4302-0.
18. Zhao, X., Wei, C., Lin, M., Feng, X., Lan, W: Petri Nets Hierarchical Modelling Framework of Active Products Community, Advances in Petri Net Theory and Applications, 2010, page 153–174, ISBN 978-953-307-108-4.

Chapter 14
Combined Method for Solving of 1D Nonlinear Schrödinger Equation

Vyacheslav A. Trofimov and Evgeny M. Trykin

Abstract We propose combined method, based on using of both the conservative finite-difference scheme and non-conservative Rosenbrock method, for solving a linear or non-linear 1D Schrödinger equation. The computer simulation results, obtained by using of combined method, are compared with corresponding results obtained using the conservative finite-difference scheme or Rosenbrock method. For 2D nonlinear problem the proposed method can significantly increase a computer simulation performance due to eliminating of using an iterative process, which is necessary for the conservative finite-difference scheme realization. The efficiency of this combined method with artificial boundary conditions is demonstrated by numerical experiments.

Keywords Rosenbrock method · Conservative finite-difference scheme · Schrödinger equation · Artificial boundary conditions

14.1 Introduction

As it well known, a Schrödinger equation is widely investigated because of this equation using for describing of many physical phenomena: structure of molecules and atoms; propagation of laser beams and pulses; Bose-Einstein condensate. The main method for solving this equation is computer simulation because of nonlinearity of the Schrödinger equation in many cases. Generally, there are two approaches for finite-difference scheme construction. The first one is based on a conservatism principle. This results in nonlinearity of corresponding finite-difference schemes. For nonlinear optics problems such approach was proposed in the paper [1] and developed in [2] for various nonlinear optics problems. Other approach is based on using of split-step method for finite-difference schemes construction [3, 4]. This

V. A. Trofimov (✉) · E. M. Trykin
Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Leninskie Gory, 119992 Moscow, Russian Federation
e-mail: vatro@cs.msu.ru

E. M. Trykin
e-mail: emtrykin@gmail.com

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_14,

approach does not allow the Hamiltonian preserving for nonlinear Schrödinger equation. As a consequence, laser beam wave-front distortion or laser pulse frequency modulation distortion grows with increasing of a laser beam propagation distance. Therefore, good accuracy achieving of computer simulation results requires using a small value of mesh step on a coordinate, along which laser beam propagation takes place.

One more method for computer simulation of various differential equations was proposed by H. H. Rosenbrock in 1963 y. [5] (It should be stressed that the Rosenbrock method was applied for numerical solution of various differential equations with success [6–9]). This method is explicit one and, hence, it is non-conservative method with respect to Hamiltonian of the Schrödinger equation. Nevertheless, it has some advantages, and we believe that a conservative finite-difference scheme using together with Rosenbrock method may lead to increasing of computer simulation efficiency. It is very promising way if artificial boundary conditions [10–23] are used also at computer simulation of laser pulse propagation in nonlinear medium. That is why we develop the combined method based on using of both conservative finite-difference scheme and Rosenbrock method.

14.2 Problem Statement

Let's consider the 1D nonlinear Schrödinger equation, describing of femtosecond laser pulse propagation in photonic crystal and written below in dimensionless variables

$$\frac{\partial A}{\partial t} + iD\frac{\partial^2 A}{\partial z^2} + i\beta A + i\gamma A\left|A\right|^2 = 0, 0 < z < L_z, 0 < t < L_t \tag{14.1}$$

with initial condition

$$A\Big|_{t=0} = e^{-(z-L_c)^2 + i2\pi\chi(z-L_c)} \tag{14.2}$$

and artificial boundary condition, formulated in following way

$$\begin{aligned} &\left(\frac{\partial A}{\partial t} - 2D\Omega\frac{\partial A}{\partial z} + 2i\beta A + i\gamma A\,|\,A\,|^2\right)\Big|_{z=0} = 0, \\ &\left(\frac{\partial A}{\partial t} + 2D\Omega\frac{\partial A}{\partial z} + 2i\beta A + i\gamma A\,|\,A\,|^2\right)\Big|_{z=L_z} = 0. \end{aligned} \tag{14.3}$$

In Eq. 14.1 a function $A = A(t, z)$ is the complex amplitude; t is a time; z denotes a spatial coordinate; L_t, L_z are maximal values of time and space coordinates, L_c is a coordinate of the laser beam center at initial time. $D, \beta, \gamma, \chi, \Omega$ are real coefficients, which satisfy the conditions $D = {}^{1}\!/_{4\pi\chi}, \beta = \pi\chi, \Omega = 2\pi\chi$. For definite we choose $\chi = 1$.

14.3 Finite-Difference Scheme Based on Rosenbrock Method

To construct a finite-difference scheme based on the Rosenbrock method we represent the complex amplitude by using real and imaginary parts (note that the modern computer can calculate in complex arithmetic, and this representation is not necessary for a method implementation)

$$A(t,\ z) = u(t,z) + iv(t,z). \tag{14.4}$$

In the interval $0 \le z \le L_z$ we introduce an uniform grid

$$\omega_z = \left\{ z_j = jh,\ j = 0,...,N_z,\ h = {}^{L_z}\!/_{N_z} \right\}. \tag{14.5}$$

Let's define the grid functions

$$\begin{aligned} &\overline{A}_j = \overline{A}_h(t, z_j) = \overline{u}_j + i\overline{v}_j, \overline{u}_j = \overline{u}_h(t, z_j), \\ &\overline{v}_j = \overline{v}_h(t, z_j), 0 \le j \le N_z \end{aligned} \tag{14.6}$$

in the nodes of the grid ω_z and write the difference Laplace operator

$$\Lambda_{z\bar{z}} A_j = \frac{A_{j-1} - 2A_j + A_{j+1}}{h^2}, 1 \le j \le N_{z-1}. \tag{14.7}$$

Using Eqs. 14.4–14.7 we can write the following set of ODE for a solution of the problem Eqs. 14.1–14.3

$$\frac{d\overline{U}}{dt} = G(\overline{U}),\ j = \overline{0, N_z}, \tag{14.8}$$

where $\overline{U} = (\overline{u}_0, \overline{u}_1, ..., \overline{u}_{N_z}, \overline{v}_0, \overline{v}_1, ..., \overline{v}_{N_z})$, and a vector $G(\overline{U})$ is calculated in following way

$$G\left(\overline{U}_0\right)=\begin{pmatrix} 2D\Omega\dfrac{\overline{u}_1-\overline{u}_0}{h}+2\beta\overline{v}_0+\gamma\left(\overline{u}_0^2+\overline{v}_0^2\right)\overline{v}_0 \\ 2D\Omega\dfrac{\overline{v}_1-\overline{v}_0}{h}-2\beta\overline{u}_0-\gamma\left(\overline{u}_0^2+\overline{v}_0^2\right)\overline{u}_0 \end{pmatrix},$$

$$G\left(\overline{U}_j\right)=\begin{pmatrix} D\Lambda_{zz}\overline{v}_j+\beta\overline{v}_j+\gamma\left(\overline{u}_j^2+\overline{v}_j^2\right)\overline{v}_j \\ -D\Lambda_{zz}\overline{u}_j-\beta\overline{u}_j-\gamma\left(\overline{u}_j^2+\overline{v}_j^2\right)\overline{v}_j \end{pmatrix}, j=\overline{1,N_z-1}, \tag{14.9}$$

$$G\left(\overline{U}_{N_z}\right)=\begin{pmatrix} -2D\Omega\dfrac{\overline{u}_{N_z}-\overline{u}_{N_z-1}}{h}+2\beta\overline{v}_{N_z}+\gamma\left(\overline{u}_{N_z}^2+\overline{v}_{N_z}^2\right)\overline{v}_{N_z} \\ -2D\Omega\dfrac{\overline{v}_{N_z}-\overline{v}_{N_z-1}}{h}-2\beta\overline{u}_{N_z}-\gamma\left(\overline{u}_{N_z}^2+\overline{v}_{N_z}^2\right)\overline{u}_{N_z} \end{pmatrix}.$$

The next step in finite-difference scheme constructing is discretization of the time interval. For this purpose we introduce uniform grid along the time

$$\omega_t=\left\{t_m=m\tau,\ m=0,...,N_t,\ \tau={}^{L_t}\!/_{N_t}\right\}. \tag{14.10}$$

After that we define the grid functions

$$\begin{aligned} &A_{m,j}=A_h(t_m,z_j)=u_{m,j}+iv_{m,j}, u_{m,j}=u_h(t_m,z_j),\\ &v_{m,j}=v_h(t_m,z_j), 0\le j\le N_z, 0\le m\le N_t,\\ &\hat{U}=(\overline{u}_{m+1,0},\overline{u}_{m+1,1},...,\overline{u}_{m+1,N_z},\overline{v}_{m+1,0},\overline{v}_{m+1,1},...,\overline{v}_{m+1,N_z}),\\ &U=(\overline{u}_{m,0},\overline{u}_{m,1},...,\overline{u}_{m,N_z},\overline{v}_{m,0},\overline{v}_{m,1},...,\overline{v}_{m,N_z}). \end{aligned} \tag{14.11}$$

Below for brevity, we omit the index h in notation of the mesh functions.

According to Rosenbrock method, the solution on the next layer on time is calculated in following way

$$\hat{U}=U+\tau Rek, j=\overline{0,N_z}, \tag{14.12}$$

where *Re k* is a real part of the vector k, which is a solution of linear equations

$$\left(E-\beta\tau G_U\right)k=G(U),\ j=\overline{0,N_z}. \tag{14.13}$$

Above E is the identity matrix, G_U is the Jacobian matrix for set of equations Eq. 14.8 and β is a complex parameter, which takes one of two values: complex $\beta=0.5+0.5\imath$ or real $\beta=0.5$.

To write the set of equations, corresponding to Eq. 14.12 let's represent the vector k and coefficient β in the form

$$k = \begin{pmatrix} k_u \\ k_v \end{pmatrix} + i \begin{pmatrix} \tilde{k}_u \\ \tilde{k}_v \end{pmatrix}, \beta = \beta_{Re} + i\beta_{Im}. \tag{14.14}$$

Using these notations one can rewrite Eq. 14.13 in the form

$$\begin{aligned} &\left(E - \beta_{Re}\tau G_U\right)\begin{pmatrix} k_u \\ k_v \end{pmatrix} + \beta_{Im}\tau G_U \begin{pmatrix} \tilde{k}_u \\ \tilde{k}_v \end{pmatrix} = \begin{pmatrix} G(U)_{Re} \\ G(U)_{Im} \end{pmatrix}, \\ &-\beta_{Im}\tau G_U \begin{pmatrix} k_u \\ k_v \end{pmatrix} + \left(E - \beta_{Re}\tau G_U\right)\begin{pmatrix} \tilde{k}_u \\ \tilde{k}_v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \end{aligned} \tag{14.15}$$

This set of equations can be rewritten in matrix form

$$\begin{aligned} &C_0\hat{Y}_0 - B_0\hat{Y}_1 = F_0(U_0), \quad -Q_j\hat{Y}_{j-1} + C_j\hat{Y}_j - B_j\hat{Y}_{j+1} = F_j(U_j), \\ &-Q_{N_z}\hat{Y}_{N_z-1} + C_{N_z}\hat{Y}_{N_z} = F_{N_z}(U_{N_z}), \\ &\hat{Y}_j = \left(k_{u_j}, \quad k_{v_j}, \quad \tilde{k}_{u_j}, \quad \tilde{k}_{v_j}\right)^T, \quad 0 \le j \le N_z. \end{aligned} \tag{14.16}$$

As it well known, an effective method for solving this set of equations is Thomas algorithm, with accordance to which the solution of Eq. 14.16 is represented as

$$Y_j = \alpha_{j+1}Y_{j+1} + \zeta_{j+1}, \quad j = N_z - 1, ..., 0, \quad Y_{N_z} = \zeta_{N_z+1}, \tag{14.17}$$

where α_j is a matrix, ζ_j is a vector and they are calculated in following way

$$\begin{aligned} &\alpha_{j+1} = (C_j - Q_j\alpha_j)^{-1}B_j, \quad j = 1, ..., N_z - 1; \quad \alpha_1 = C_0^{-1}B_0. \\ &\zeta_{j+1} = (C_j - Q_j\alpha_j)^{-1}(F_j + Q_j\alpha_j), \quad j = 1, ..., N_z; \quad \zeta_1 = C_0^{-1}F_0. \end{aligned} \tag{14.18}$$

It should be stressed that this method is conditionally conservative one, has second order approximation in spatial coordinate and first order approximation in time coordinate and it is explicit.

14.4 Conservative Finite-Difference Scheme

Let's use the same grids, which are written above for the Rosenbrock method, and define the mesh functions

$$\begin{aligned} &A = A_j = A_{m,j} = A(t_m, z_j), \quad \hat{A} = \hat{A}_j = \hat{A}_{m,j} = \hat{A}(t_m + \tau, z_j), \\ &\overset{0.5}{A} = 0.5(\hat{A} + A), \quad |\overset{0.5}{A}|^2 = 0.5(|\hat{A}|^2 + |A|^2). \end{aligned} \tag{14.19}$$

For the problem Eqs. 14.1–14.3 we write the following conservative finite-difference scheme

$$\begin{aligned}
&\frac{\hat{A}-A}{\tau}+iD\Lambda_{z\bar{z}}\overset{0.5}{A}+i\beta\overset{0.5}{A}+i\gamma|\overset{0.5}{A}|^2\overset{0.5}{A}=0,\ j=\overline{1,N_z-1},m=\overline{1,N_t},\\
&\frac{\hat{A}_0-A_0}{\tau}-2D\Omega\frac{\overset{0.5}{A}_1-\overset{0.5}{A}_0}{h}+2i\beta\overset{0.5}{A}_0+i\gamma|\overset{0.5}{A}_0|^2\overset{0.5}{A}_0=0,\\
&\frac{\hat{A}_{N_z}-A_{N_z}}{\tau}+2D\Omega\frac{\overset{0.5}{A}_{N_z}-\overset{0.5}{A}_{N_z-1}}{h}+2i\beta\overset{0.5}{A}_{N_z}+i\gamma|\overset{0.5}{A}_{N_z}|^2\overset{0.5}{A}_{N_z}=0.
\end{aligned}\tag{14.20}$$

The problem Eq. 14.20 is nonlinear one, that is why for its solution we use an iteration process. For example, this process can be implemented in following way

$$\begin{aligned}
&\frac{\overset{s+1}{\hat{A}}-A}{\tau}+iD\Lambda_{z\bar{z}}\overset{s+1}{\overset{0.5}{A}}+i\beta\overset{s+1}{\overset{0.5}{A}}+i\gamma|\overset{s}{\overset{0.5}{A}}|^2\overset{s}{\overset{0.5}{A}}=0,\ j=\overline{1,N_z-1},m=\overline{1,N_t},\\
&\frac{\overset{s+1}{\hat{A}}_0-A_0}{\tau}-2D\Omega\frac{\overset{s+1}{\overset{0.5}{A}}_1-\overset{s+1}{\overset{0.5}{A}}_0}{h}+2i\beta\overset{s+1}{\overset{0.5}{A}}_0+i\gamma|\overset{s}{\overset{0.5}{A}}_0|^2\overset{s}{\overset{0.5}{A}}_0=0,\\
&\frac{\overset{s+1}{\hat{A}}_{N_z}-A_{N_z}}{\tau}+2D\Omega\frac{\overset{s+1}{\overset{0.5}{A}}_{N_z}-\overset{s+1}{\overset{0.5}{A}}_{N_z-1}}{h}+2i\beta\overset{s+1}{\overset{0.5}{A}}_{N_z}+i\gamma|\overset{s}{\overset{0.5}{A}}_{N_z}|^2\overset{s}{\overset{0.5}{A}}_{N_z}=0.
\end{aligned}\tag{14.21}$$

The value of the mesh function on the upper layer at zero iteration (s=0) is chosen as

$$\overset{s=0}{\hat{A}}=A.\tag{14.22}$$

The iteration process is stopped if the following condition, for example, is valid

$$\max_{z_j}|\overset{s+1}{\hat{A}}-\overset{s}{\hat{A}}|\le\varepsilon_1\max_{z_j}|\overset{s}{\hat{A}}|+\varepsilon_2,\quad \varepsilon_1,\varepsilon_2>0.\tag{14.23}$$

Let's rewrite the equation Eq. 14.21 in a matrix kind

$$\begin{aligned}
&C_0\hat{Y}_0-B_0\hat{Y}_1=F_0(A_0),\\
&-Q_j\hat{Y}_{j-1}+C_j\hat{Y}_j-B_j\hat{Y}_{j+1}=F_j(A_j,A_{j-1},A_{j+1},\overset{s}{\hat{A}}_j),\quad \overset{s+1}{\hat{A}}_j=\begin{pmatrix}\overset{s+1}{A_j^R}\\ \overset{s+1}{A_j^I}\end{pmatrix},0\le j\le N_z,\\
&-Q_{N_z}\hat{Y}_{N_z-1}+C_{N_z}\hat{Y}_{N_z}=F_{N_z}(A_{N_z}).
\end{aligned}\tag{14.24}$$

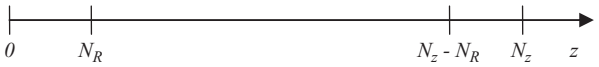


Fig. 14.1 Template of Rosenbrock method using for combined method

As well as for solution of the equations Eq. 14.16 we use the Thomas algorithm for solution of the set of equations Eq. 14.24.

The finite-difference scheme developed above is conservative on energy invariant, Hamiltonian and invariant of laser beam impulse, has the property of symmetry and second order approximation in any coordinate, and it is implicit one. The last circumstance requires the iterative procedure using.

14.5 Combined Finite-Difference Scheme

The main idea of a combined method consists in using of Rosenbrock method in the domain near the boundaries. It means that we introduce two subdomains near left and right boundaries (see Fig. 14.1). These domains belong to nodes $[0, N_R]$ and $[N_z - N_R, N_z]$ correspondingly.

In order to obtain numerical solution on the next time layer we need to find a solution in domains $[0, N_R]$ and $[N_z - N_R, N_z]$ using Rosenbrock method (Eqs. 14.8–14.16) with artificial conditions (Eq. 14.3) in boundary points of spatial grid. Let's define this solution as $\hat{A}_R$. Then obtained solution in boundary points of spatial grid we use as boundary conditions for conservative finite-difference scheme (Eqs. 14.20, 14.21)

$$\hat{A}_{C_0} = \hat{A}_{R_0}, \hat{A}_{C_{N_z}} = \hat{A}_{R_{N_z}}, \tag{14.25}$$

where $\hat{A}_c$ denotes as solution obtained by the conservative finite-difference scheme use solution $\hat{A}_c$ on the next time layer as initial condition for Rosenbrock method.

Thus, we obtain an explicit combined method for solving of the nonlinear Schrödinger equation.

14.6 Computer Simulation Results

We consider three schemes: *Ros*—Rosenbrock method with parameter $\beta = 0.5 + 0.5\imath$, *C*—conservative finite-difference scheme, *R&C*—combined method. The results of computer simulation using a combined method will be compared with the results of calculations made using the conservative finite-difference scheme or the Rosenbrock method or exact solution (*Ex*) of a linear Schrödinger equation.

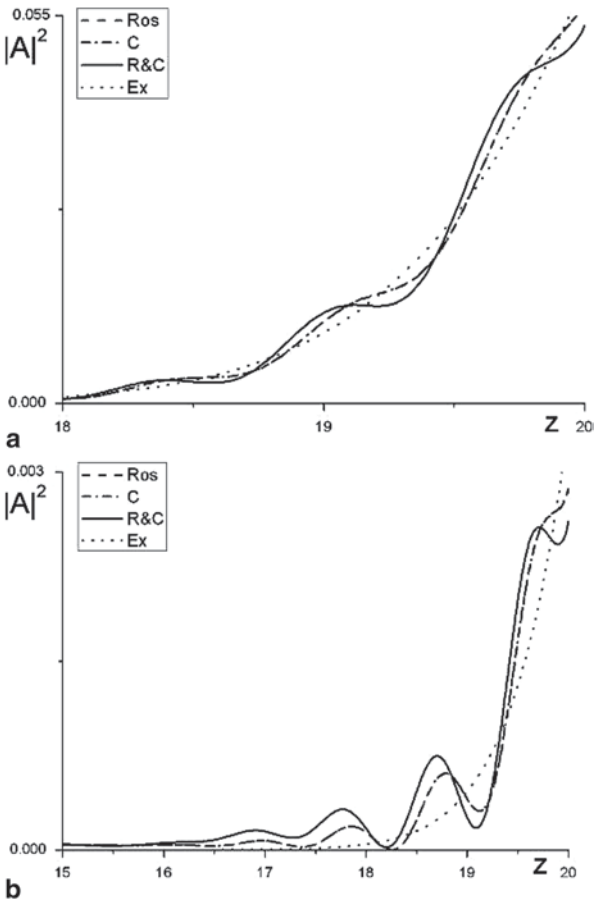


Fig. 14.2 Intensity profile $|A|^2$ for time moment $t = 7.5(a), 10.0(b)$

Let's define the maximum value of reflected amplitude as

$$A_{Ref}(S) = \max_{t_m} \max_{z_j} \left| |A_S(t_m, z_j)|^2 - |A_{Ex}(t_m, z_j)|^2 \right|, \quad 0 \le j \le N_z, 0 \le m \le N_t,$$
$$S = \{R, C, R\&C\} \tag{14.26}$$

and parameter values $L_z = 20.0, h = 0.01,\ D = 0.0796, \beta = 3.14,$ for which a computer simulation is made.

- Linear problem

Comparison of the intensity distributions, obtained using the Rosenbrock method or the conservative finite-difference scheme or combined method for solution of a linear problem $\gamma = 0$ with parameters $\tau = 0.01, N_R = 100,10$ is shown in Figs. 14.2 and 14.3 correspondingly.

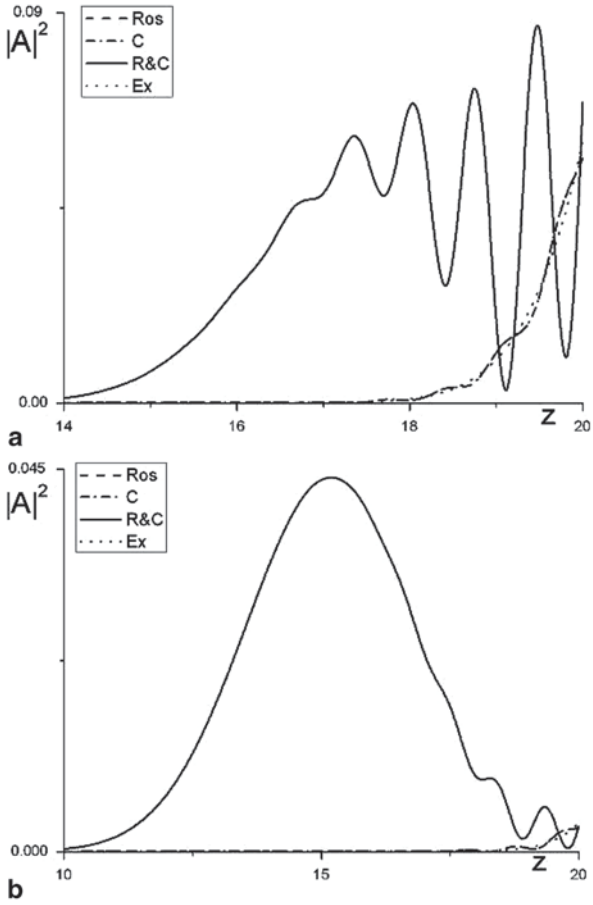


Fig. 14.3 Intensity profile $|A|^2$ for time moment $t = 7.5(a), 10.0(b)$

As it follows from Figs. 14.2 and 14.3, the deviation of the solutions obtained on the base of the combined method from the corresponding solution obtained using other schemes becomes significant with decreasing a number of nodes for subdomains, in which Rosenbrock method is used if a time step is constant. Therefore, one should to decrease a mesh step on time coordinate. The corresponding comparison, which is a similar to Fig. 14.3, is depicted in Fig. 14.4 for parameters $N_R = 10, \tau = 0.001$.

As we can see in Fig. 14.4, in this case the solution obtained using the combined method tends to the corresponding solution obtained using other schemes.

For an estimation of efficiency of proposed combined method let's write the first invariant for the problem Eqs. 14.1–14.3.

$$I_1 = \int_0^{L_z} |A|^2 \, dz = const \tag{14.27}$$

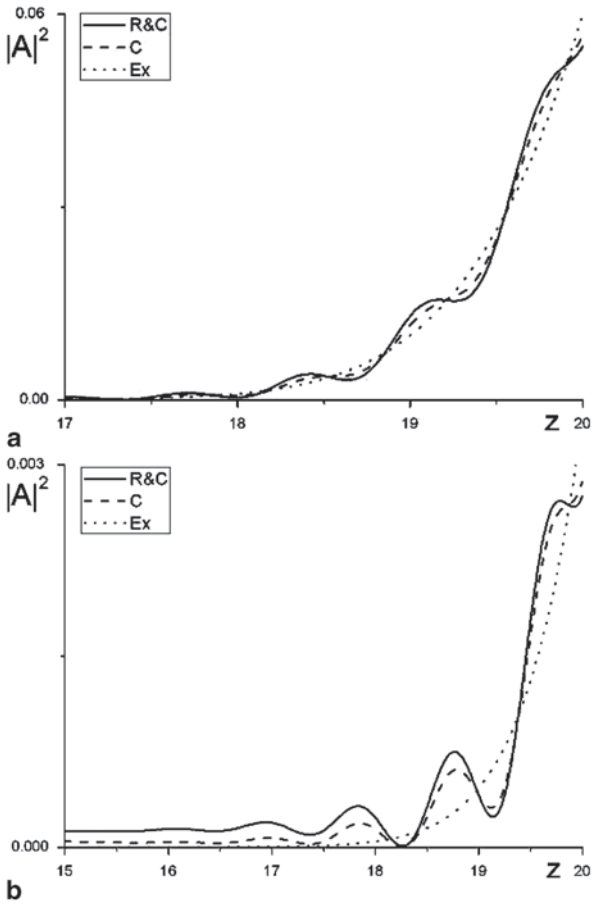


Fig. 14.4 Intensity profile $|A|^2$ for time moment $t = 7.5(a), 10.0(b)$

In Table 14.1 we present the difference of the first invariant values calculated for used finite-difference schemes and exact solution. We see that for combined method this difference is two times more than corresponding values for other schemes. However, its value does not exceed the theoretical estimation.

Let's compare the maximum value of reflected beam amplitude for combined method and corresponding amplitude calculated using conservative finite-difference scheme or Rosenbrock method (see Fig. 14.5).

Very important question consists in a possibility of getting the coinciding results at using of both combined method and conservative finite-difference scheme. In Fig. 14.6 the corresponding dependence is shown.

From Fig. 14.6 we can see that decreasing a number of mesh step on spatial coordinate N_R we have to decrease the step on time coordinate for combined method to achieve the accuracy, which coincides with accuracy of the conservative finite-difference scheme.

Table 14.1 Difference between the first invariant values computed using considered methods and the exact solution of a linear problem with parameters $L_z = 20.0, t = 10, h = 0.01, \tau = 0.01$

Method	First invariant difference value
R	0.000548643
C	0.000556224
R&C	0.001128941

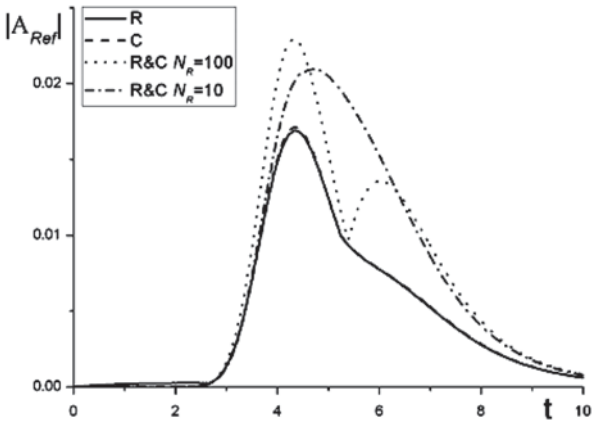


Fig. 14.5 Evolution of the maximum amplitude of reflected beam for time interval $0 \le t \le 10$

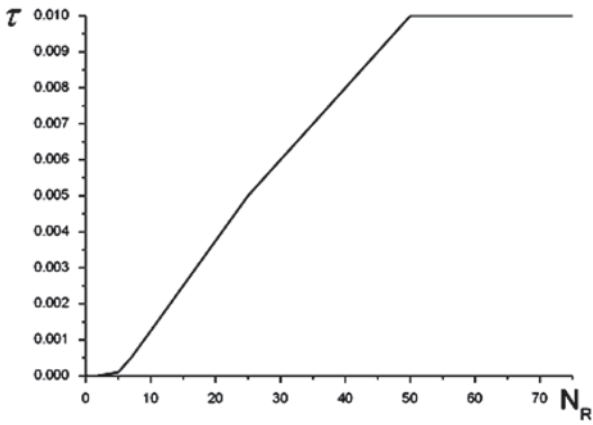


Fig. 14.6 Dependence of mesh step for time coordinate on mesh number for subdomains, in which Rosenbrock method is used. Time interval is equal $t = 10.0$

- Self-focusing of laser beam

Comparison of the intensity distributions, calculated using the Rosenbrock method or the conservative finite-difference scheme or the combined method for nonlinear problem $\gamma = 1.0$, is depicted in Fig. 14.7. It should be stressed that we take the solu-

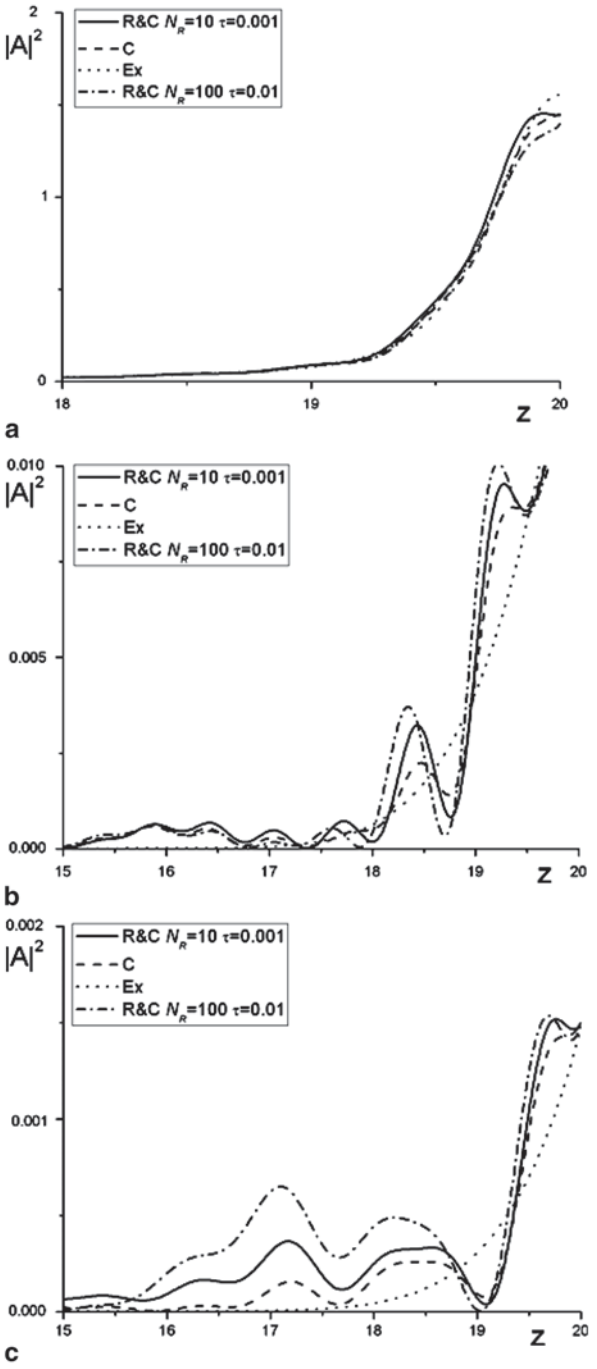


Fig. 14.7 Intensity profile $|A|^2$ for time moment $t = 5.0(a), 7.5(b), 10.0(c)$ computed for mesh step and number of nodes for subdomains $\tau = 0.01, 0.001, N_R = 100, 10$

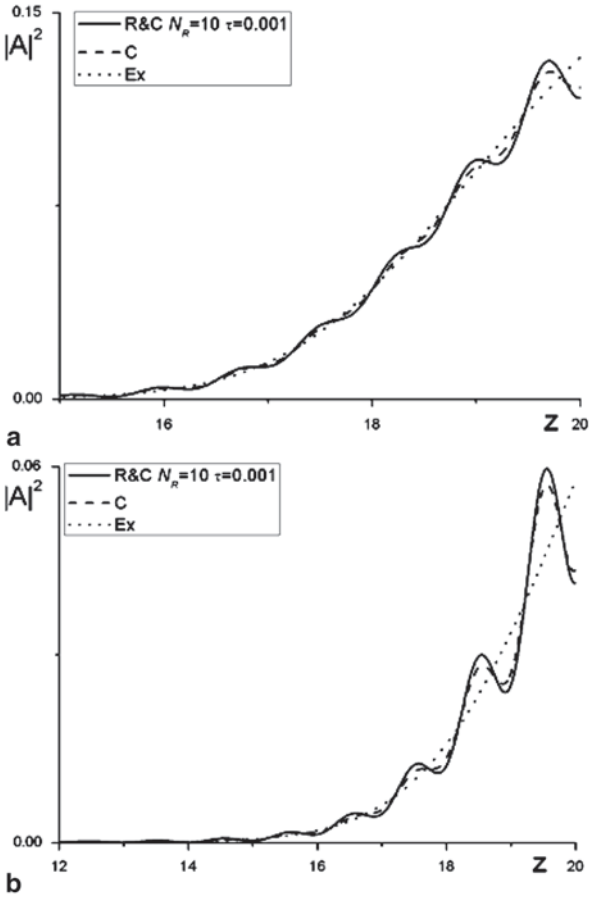


Fig. 14.8 Intensity profile $|A|^2$ for time moment $t = 7.5(a),10.0(b)$

tion, obtained by the conservative finite-difference scheme with zero-value boundary condition for sufficient bigger size of area as exact solution in nonlinear case.

As we can see in Fig. 14.7a, the difference between solution obtained by the combined method and exact solution has the order 0.1, but this error is not preserved in time. However, the wave, reflected from artificial boundary condition, is absent at all. At the next time moments (Fig. 14.7b, c) this difference becomes essentially less. It is approximately about 10^{-3} and 10^{-4} correspondingly. Also the solutions, calculated by combined method and conservative finite-difference scheme, have the same approximation order and the maximum amplitude value of reflected beam has the order about 10^{-4}.

- Defocusing of laser beam

Comparison of the intensity distributions, obtained using Rosenbrock method or the conservative finite-difference scheme or the combined method for nonlinear problem $\gamma = -1.0$ with parameters $N_R = 10, \tau = 0.001$, is shown in Fig. 14.8.

As we can see in Fig. 14.8, the difference between the solution obtained using the combined method tends to the corresponding solution obtained using the conservative finite-difference scheme. It should be stressed that the maximum amplitude value of reflected beam in Fig. 14.8b has the order about 10^{-2} and its worst than in self-focusing case. It is takes place because of the parameter Ω is constant in time and nonlinear distortion of wave-front of laser beam, propagating in defocusing medium, is stronger than for previous case.

14.7 Conclusion

In this paper we have proposed a new method for solving of 1D nonlinear Schrödinger equation with artificial boundary conditions. This method is explicit. It means that the solution on next time layer can be found without iterative process using.

It was shown above that the computer simulation results obtained using the combined method has the same accuracy in comparison with the corresponding results obtained using the conservative finite-difference scheme if mesh steps are chose in certain way.

The time of computer simulation using combined method is less than a time of computer simulation using Rosenbrock method and more than a time of computer simulation using the conservative finite-difference scheme.

Also was considered the dependency of invariant values deviation for three methods. So, we can say that the combined method is more suitable for solving the Schrödinger equation than Rosenbrock method.

Acknowledgements This work was supported in part by the Russian Foundation for Basic Research under Grant 12-01-00682-a.

References

1. Karamzin YuN (1974) Difference schemes for computations of three-frequency interactions of electromagnetic waves in a nonlinear medium with quadratic polarization. Zh Vychisl Mat Mat Fiz 14:4, pp 1058–1062
2. Karamzin YuN, Sukhorukov AP, Trofimov VA (1989) Mathematical modeling in nonlinear optics. Moskva: Izd-vo Moskovskogo universiteta
3. Strang G (1968) On the construction and comparison of differential schemes. SIAM J Numer Anal 5:506–517
4. Fleck JA, Morris JR, Feit MD (1977) Time-dependent propagation of high energy laser beams through the atmosphere. Appl Phys 14:99–115
5. Rosenbrock HH (1963) Some general implicit processes for the numerical solution of differential equations. The Computer Journal 5:4, pp 329–330
6. Dnestrovskaya EYu, Kalitkin NN, Ritus IV (1991) The solution of partial differential equations by schemes with complex coefficients. Journal of Mathematical Models and Computer Simulations 3:9, pp 114–127

7. Cao Xue-nian, Li Shou-fu, Liu De-gui (2002) Modified parallel Rosenbrock methods for stiff differential equations. Journal of Computational Mathematics 20:1, pp 23–34
8. Gerisch A, Chaplain MAJ (2006) Robust numerical methods for taxis–diffusion–reaction systems: Applications to biomedical problems. Mathematical and Computer Modelling 43:49–75
9. Verwer JG, Hundsdorfer WH, Blom JG (1998) Numerical time integration for air pollution models. Modelling, Analysis and Simulation MAS-R9825, pp 1–60
10. Antoine X, Arnold A, Besse C, Ehrhardt M, Schadle A (2008) A review on transparent and artificial boundary conditions technique for linear and nonlinear Schrödinger equations. Commun Comput Phys 4:729–796
11. Alpert B, Greengard L, Hagstram J (2002) Nonreflecting boundary conditions for the time-dependent wave equation. J Comput Phys 180:270–296
12. Antoine X, Besse C, Mouysset V (2004) Numerical schemes for the simulation of the two–dimensional Schrödinger equation using non–reflecting boundary conditions. Math Comput 73:1779–1799
13. Arnold A, Ehrhardt M, Sofronov I (2003) Discrete transparent boundary conditions for the Schrödinger equation: fast calculation, approximation, and stability. Commun Math Sci 1:501–556
14. Han H, Yin D, Huang Z (2006) Numerical solutions of Schrödinger equations in R3. Numer Meth Partial Different Equations 23:511–533
15. Jiang S, Greengard L (2007) Efficient representation of nonreflecting boundary conditions for the time-dependent Schrödinger equation in two dimensions. Commun Pure Appl Math 61:261–288
16. Luchini P, Tognaccin R (1996) Direction-adaptive nonreflecting boundary conditions. J of Comput Phys 128:121–133
17. Schadle A (2002) Non–reflecting boundary conditions for the two dimensional Schrödinger equation. Wave Motion 35:181–188
18. Szeftel J (2006) Absorbing boundary conditions for one-dimensional nonlinear Schrödinger equations. Numer Math 104:103–127
19. Xu Z, Han H, Wu X (2007) Adaptive absorbing boundary conditions for Schrödinger-type equations: application to nonlinear and multi-dimensional problems. J Comput Phys 225:1577–1589
20. Zheng C (2006) Exact nonreflecting boundary conditions for one–dimensional cubic nonlinear Schrödinger equations. J Comput Phys 215, pp:552–565
21. Xuand Z, Han H (2006) Absorbing boundary conditions for nonlinear Schrödinger equations. Phys Rev 74:037704
22. Tereshin EB, Trofimov VA, Fedotov MV (2006) Conservative finite difference scheme for the problem of propagation of a femtosecond pulse in a nonlinear photonic crystal with non-reflecting boundary conditions. Computational Mathematics and Mathematical Physics 46:1, pp 154–164
23. Trofimov VA, Dogadushkin PV (2007) Boundary conditions for the problem of femtosecond pulse propagation in absorption layered structure. Farago I, Vabishevich P, Vulkov L, Proceedings of Fourth Intern Conf Finite Difference Methods: Theory and applications, Rouse University Angel Kanchev, Lozenetz, Bulgaria, pp 307–313

Chapter 15
Towards Real Time Implementation of Sparse Representation Classifier (SRC) Based Heartbeat Biometric System

W. C. Tan, H. M. Yeap, K. J. Chee and D. A. Ramli

Abstract Implementation of the heartbeat biometric system consists of four main stages which are heartbeat data acquisition, pre-processing and feature extraction, modeling and classification. In this study a new approach for classification method based on Sparse Representation Classifier (SRC) is proposed. By introducing kernel trick into SRC, the classification performance of the classifier can be further improved by implicitly map features data into a high-dimensional kernel feature space. Based on heart sound data, experimental results have shown a promising performance of KSRC with 85.45 % of accuracy has been achieved and a better performance has been observed by this classifier compared to Support Vector Machines (SVM), SRC and K-Nearest Neighbor (KNN). This achievement has proved the possibility of heartbeat as a biometric trait for human authentication system. Due to this, an extension in term of heartbeat data acquisition toward real time implementation is then proposed in this paper. Here, a wrist-mounted heartbeat sensor to sense the heartbeat signal is designed. This developed sensor is an electrometer which is capable to measure the properties of electrocardiogram (ECG) signal. The developed hardware has also shown its viability toward execution of heartbeat data acquisition in real time.

Keywords Biometrics · Heartbeat · ECG · Kernel trick · Sparse representation classifier

W. C. Tan (✉) · H. M. Yeap · K. J. Chee · D. A. Ramli
Intelligent Biometric Research Group, School of Electrical and Electronic, Engineering Campus, Universiti Sains Malaysia, 14300, Nibong Tebal, Penang, Malaysia
e-mail: zero_0317@hotmail.com

H. M. Yeap
e-mail: hmy_mtk@hotmail.com

K. J. Chee
e-mail: rabenastre@live.com

D. A. Ramli
e-mail: dzati@eng.usm.my

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_15,

15.1 Introduction

Automated security is one of the major concerns of modern time where secure and reliable authentication is in great demand. However, traditional authentication methods such as password and smart card are now outdated due to they can be lost, stolen and shared. In this project, biometric system based on heartbeat signal is proposed. Heartbeat is chosen as modality due to an individual's heart sound parameters cannot be faked. Compared to fingerprint, it can be fooled with fake fingers, face can be extracted using user's photo and voice can be imitated easily. Besides, as heart sound is reflection of the mechanical movement of the heart and cardiovascular system, these features contains both physiological and pathological information.

Recent research [2, 12] have been proved that heartbeat or heart sound can be used as the biometric trait for human authentication. Human heart sounds are noises generated by the beating heart and the resultant flow of blood through it. Two heart sounds are normally produced during each cardiac cycle namely S1 and S2. The first heart sound S1 is normally longer, low-pitch tone and sound like "lup" whereas the second heart sound S2 is shorter, high-pitch and sound like "dup". These natural signals have been applied in auscultation by doctors for health monitoring and diagnosis.

Since heart sounds contain information about an individual's physiology, it can be potentially used as a biometric traits and provide unique identity for each person. Besides, heart sound is very difficult to counterfeit or imitate by others and therefore reduces falsification in authentication systems. In 2006, the possibility of using heart sound as biometric trait for human identification is investigated and a preliminary results indicate an identification rate of up to 96 % for a database consists of 7 individuals, with heart sounds collected over a period of 2 months [12]. Their system is based on the cepstral analysis with a specified configuration called Linear Frequency Bands Cepstral (LFBC) as feature extraction method, combined with Gaussian Mixture Modeling (GMM) and Vector Quantization (VQ) as classifier.

In 2007 [2], a heart sound biometric system is proposed by the authors using a feature extraction method called chirp-Z transform (CZT) and K-Nearest Neighbor (KNN) based on Euclidean distance as the classifier. Their system achieved 0 % false rejection rate (FRR) and 2.2 % false acceptance rate using a database containing heart sound recorded from 20 different people. The weakness of the CZT feature extraction method is that the locations of the S1 and S2 heart sounds have to be well aligned for each sample.

In 2010 [7], three different types of features are extracted which are auto-correlation, cross-correlation and cepstrum. The classifiers applied in their systems are Mean Square Error (MSE) and KNN. KNN classifier achieved 93 % identification rate evaluated using a database of 400 heart sound that were recorded from 40 individuals by 10 heart sound recordings for each individuals.

In 2013 [17], a new feature set called marginal spectrum is extracted from the heart sounds and classifier VQ based on Linde-Buzo-Gray algorithm (LBG-VQ)

is used for classification the heart sounds. The identification rate of their system achieved 94.40% evaluated using a database of 280 heart sounds from 40 participants.

In this paper, a heart sound authentication system based on Mel Frequency Cepstral Coefficient (MFCC) and Sparse Representation Classifier (SRC) are used as feature extraction and classification method respectively. Thus, this research aims to develop a robust and reliable heart sound authentication system which can work well with noisy heart sound sample. The proposed system is composed of four main phases; data acquisition, signal pre-processing, feature extraction, and training and classification phases. Consequently, heartbeat data acquisition toward real time implementation of heartbeat biometric system is then proposed in this paper. Here, a wrist-mounted heartbeat sensor to sense the heartbeat signal is designed.

This paper is organized as the following order. Section 15.2 briefly explains the methodology in extracting the features of heartbeat data. Then, Sect. 15.3 presents the classification process using SRC and KSRC. The proposed design of wrist-mounted heartbeat sensor is then given in Sect. 15.4. Consequently, the result and discussion are described Sect. 15.5. Finally, Sect. 15.6 sums up the overall conclusion.

15.2 Methodology

15.2.1 Database

An open heart sounds database HSCT-11 collected by the University of Catania Italy is applied to evaluate the performance of proposed heart sound authentication system. This database is a collection of heart sounds to be used for biometric research purpose and freely available on the internet [13]. It contains heart sounds collected from 206 people, i.e. 49 female and 157 male. Only 10 female and 5 male heart sounds are randomly selected have been used in this research. The heart sounds recordings are recorded in WAV format at a sampling frequency of 11.025 kHz, near the pulmonary valve and contains only sequences recorded in resting condition.

15.2.2 Pre-processing, Segmentation and Feature Extraction

The recorded heart sound signals corrupted by various types of noise can reduce the accuracy of identification. To overcome this problem, a fifth order Chebyshev type I lowpass filter with cutoff frequency at 880 Hz is applied on the signals. In this context, background noise or sound with frequency that higher than the filter cut off frequency will be eliminated and the signals are then normalized before the segmentation process takes place.

Table 15.1 Threshold value for heart sound segmentation process

Segmentation threshold parameter	Initialize value
Upper short-term amplitude threshold, STA1	3
Lower short-term amplitude threshold, STA2	0.5
Zero-crossing rate threshold, ZCR	5

The heart produces two strong and audible sounds namely S1 and S2. These two heart sounds contain important features for human identity verification. Therefore, the heart sound segmentation is the first step of this automatic heart sound biometric system [15].

The segmentation technique employed in this system is based on zero-crossing rate (ZCR) and short-term amplitude (STA). First, the noise-filtered and normalized signal is blocked into frames of 5 m length with 66.7% overlapped. Next, the short-term amplitude and zero-crossing rate of each frame are calculated. This simple feature can be used for detecting silent part in audio signals which is especially helpful for detecting speech from noisy background and for start and end point detection. Then, the values of upper and lower of STA and ZCR thresholds are set as given in Table 15.1.

These threshold values are obtained from trial and error process. After defining the threshold values, the frames of the signal is evaluated by the following rules;

Rule 1: the frame's STA greater than STA1 threshold is considered as a part of heart sound and the starting point of the heart sound will be calculated.

Rule 2: the frame's with STA greater than STA2 threshold or ZCR greater than ZCR threshold is considered as possible heart sound signal and will be further evaluated next frames. If the next frame matches rule 2, the evaluation will be repeated until it matches rule 1 or rule 3. Once the next frame matches rule1, the starting point of the heart sound will be equal to the starting point of the very first frame which matches rule 2. The ending point of this heart sound is evaluated when the following frame matches rule 3. If the next frames does not matches rule 1 and directly matches rule 3, this sequence of frames will not consider as a heart sound.

Rule 3: the frame's with STA lower than STA2 threshold and ZCR lower than ZCR threshold is considered as not a heart sound. The flow chart of the segmentation technique is shown in Fig. 15.1.

The extraction of the best parametric representation of acoustic signals is an important task in designing of any sound-based biometric recognition system so that a better identification performance can be produced. Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction method in speech recognition. MFCC takes human hearing perception sensitivity with respect to frequencies into consideration [9, 11]. After the heart sound signal is segmented, framed, and windowed, MFCC is used to extract meaningful parameter in the heart sound signal.

The steps to implement MFCC in this system are summarized as in Fig. 15.2. The result from first and second derivatives is also added as new features. Hence, a 39-dimensional MFCC features per frames is extracted from the digitized heart sound signal. The features extracted forms S2 heart sound will be appended to the

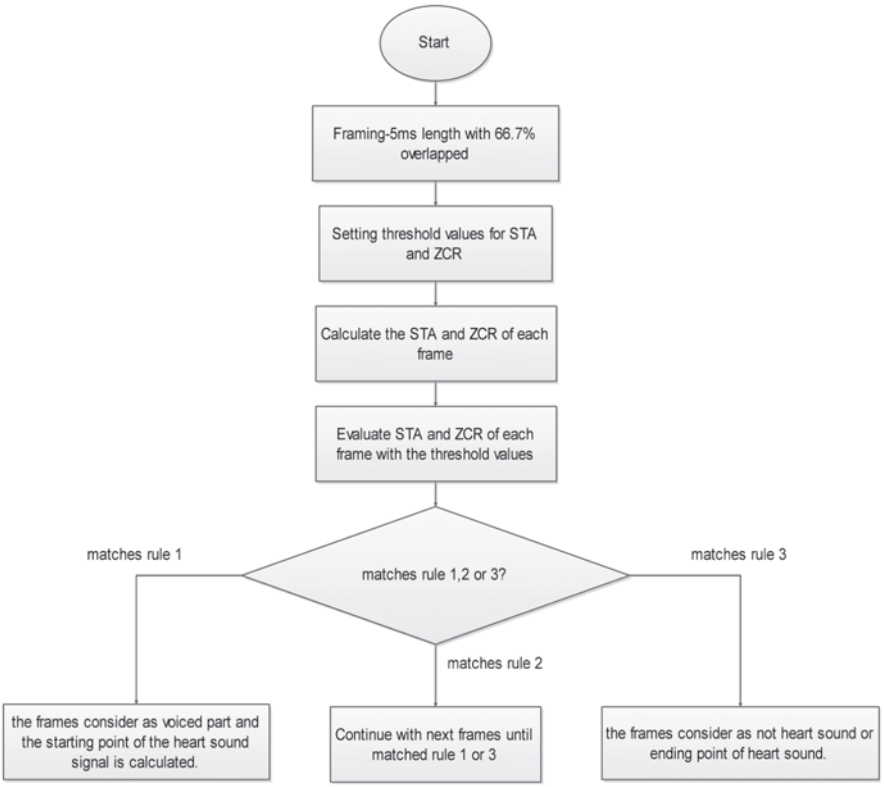


Fig. 15.1 Heart sound segmentation flow chart based on ZCR and STA

features extracted from S1 sound so that features of a complete heart sound cycle are used in classification process.

15.3 Sparse Representation Classifier (SRC)

Sparse representation is originally applied for signal representation and reconstruction. Sparse representation of signal is an expression of the signal as a linear combination of atoms in an overcomplete dictionary in which many of the coefficients are zero. The original goal of sparse representation is to represent and compress a signal using lower sampling rates than the Shannon-Nyquist rate [5]. Thus, the performance of the compress algorithm is based on the degree of sparsity of the representation to the original signal. Among all the atoms in an overcomplete dictionary, the sparse representation selects the subset of the atoms which most compactly expresses the input signal and rejects all other less compact representation.

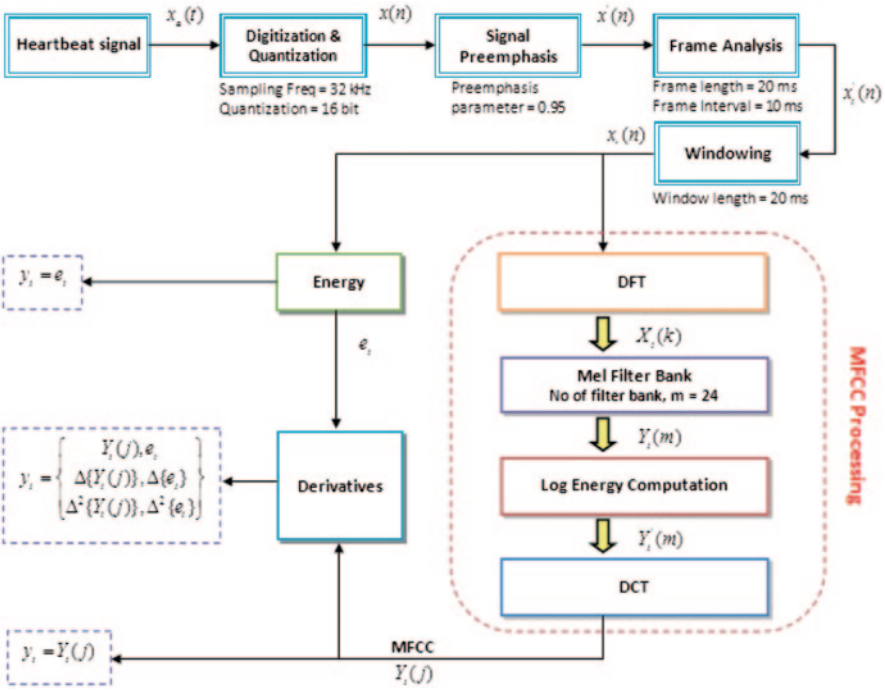


Fig. 15.2 Heartbeat feature extraction flow chart based on MFCC processing

Therefore, the sparsest representation of a signal is naturally discriminative and can be developed for signal classification purpose. Equation 15.1 shows the input signal y is linearly represented by a dictionary, D and sparse representation, x.

$$y_{m\times 1} = D_{m\times n} x_{n\times 1} \tag{15.1}$$

Sparse representation classifier is a nonparametric learning method which can directly predict or assign a class label to a test sample based on dictionary composed of training samples. This method is similar to Nearest Neighbor and Nearest Subspace classifier which do not have a training process for classification process. Sparse representation for classification is first introduced in 2009 in face recognition research [14]. Experimental results proved that sparse representation classifier (SRC) has better classification performance than nearest neighbor and nearest subspace.

In sparse representation classifier, the dictionary is constructed from training samples from various classes. The j^{th} class training samples are arranged as column of a matrix D_j as shown in Eq. 15.2. The columns of dictionary is refered as atoms.

$$D_j = \left[d_{j,1}, \ldots, d_{j,n_j} \right] \in R^{m\times n_j} \tag{15.2}$$

where $d_{j,i}$ denotes the training sample belonging to the j^{th} class, and n_j is the number of the training samples for j^{th} class. The dictionary, D is form using all the dictionary from each class as shown in Eq. 15.3.

$$D = \left[D_1, D_2, \ldots, D_c\right] \in R^{m \times n} \tag{15.3}$$

where $n = \sum_{j=1}^{c} n_j$. and c is the number of class.

Based on the Eq. 15.1, for SRC problem, the sparse representation x, is the vector of coefficients associated with the training sample in the dictionary matrix. The entries of x that is corresponding to the class which the test sample y belongs to is expected to be nonzero while the entries of x that corresponding to other classes is expected to be zero.

$$x = \left[0, \ldots, 0, x_{j,1}, \ldots, x_{j,n_j}, 0, \ldots., 0\right]^T \tag{15.4}$$

where $x_{j,i} \in R$ is the coefficient corresponding to the training sample $d_{j,i}$. The sparse representation based classification method looks for the sparsest representation by solving the following l_0 minimization problem.

$$\min \| x \|_0 \quad s.t. \quad y = Dx \tag{15.5}$$

where $\| x \|_0$ denotes the l_0 norm, which count the number of nonzero elements of sparse representation, x. Equation 15.5 is known as NP (nondeterministic polynomial) hard problem and difficult to approximate. The developed theory from sparse representation and compressive sensing research reveals that the sparsest solution from Eq. 15.5 can be obtained by replacing the l_0 norm with the l_1 norm given that the solution, x is sparse enough [3, 4, 6].

$$\min \| x \|_1 \quad s.t. \, y = Dx \tag{15.6}$$

where $\| x \|_1$ denotes the l_1 norm, which sum the absolute values of all elements in the sparse representation, x. The advantage of sparse representation based classification is their ability to deal with corrupted or noisy data within the same framework. This property of sparse representation classifier provide the advantage for heart sound biometric authentication system because usually the heart sound contains noise signal. To deal with noisy data, Eq. 15.1 is modified as

$$y_{m \times 1} = D_{m \times n} x_{n \times 1} + \xi_{m \times 1} \tag{15.7}$$

where $\xi_{m \times 1} \in R^m$ denotes the noise vector with bounded energy $\| \xi_{m \times 1} \|_2 < \varepsilon$, where $\| \cdot \|_2$ denotes the l_2 norm. While Eq. 15.6 can be modified as

$$\min \| x \|_1 \quad s.t. \quad \| y - Dx \|_2 \leq \varepsilon \tag{15.8}$$

Equation 15.15 is one standard formulation for sparse reconstruction problems in compressive sensing, called the quadratically constrained l_1 minimization problem [1, 8].

Both the l_1 minimization problems are solved using spectral projected gradient method, SPGL1 toolbox in this research. The minimum of the representation error or the residual error of class c is calculated by keeping the coefficients associated with that class and while setting the other entries to zero. This is done by introducing a characteristic function, ζ as follow.

$$r_c(y) = \| y - D\zeta_i x \|_2 \tag{15.9}$$

where $r_c(y)$ denotes the residual error. The vector ζ has value one at locations associated to the class *i* and zero for other entries. The class, *d* of the test signal, *y* is computed as the one that produces smallest residual error.

$$d = \min_i r_i(y) \tag{15.10}$$

The algorithm summarizes the complete classification procedure of SRC is described below.

Algorithm 1: Sparse Representation Classifier (SRC)

1. The input for SRC are a matrix of training samples form a dictionary $D = [D_1, D_2, \ldots, D_c] \in R^{m \times n}$ for c classes, a test sample $y \in R^m$ and an optional error tolerance $\varepsilon > 0$.
2. Normalize the atoms of D to have unit l_2 norm.
3. Solve the l_1 minimization problem in Eq. 15.13 or 15.15 using SPGL1 toolbox:
4. Compute the residuals $r_i(y) = \left\| y - D\zeta_i x \right\|_2$ *for* $i = 1, \ldots \ldots, c$.
5. The class of the given test sample, y is determined by $identity(y) = \min_i r_i(y)$.

After implementing algorithm 1, kernel tricks is then applied to the classifier to change the distribution of samples. This can be done by mapping it into a high dimensional kernel feature space [16] in order to change the linear inseparable samples in the original feature space into linear separable in the high dimensional feature space. This means a test sample can be represented as linear combination of training samples from same class accurately by applying kernel trick into SRC. The classification performance of SRC will be improved as the nonzero entries of sparse representation, *x* of the test sample are more associated with training samples from same class as itself. In this work, radial basis function (RBF) kernel is employed in KSRC as follows.

$$k(x, y) = e^{-t\|x-y\|^2} \tag{15.11}$$

where $t < 0$ is the parameter for RBF kernels. KSRC classification method is also able to overcome the disadvantages of SRC which cannot classify samples in the same direction which belong to different classes [10].

In KSRC, kernelized dictionary and testing sample is computed by the following equation.

$$D_{kernel} = \left[k\left(d_i, d_j\right)\right]_{n\times n} \tag{15.12}$$

$$y_{kernel} = \left[k\left(y, d_j\right)\right]_{n\times 1} \tag{15.13}$$

where $i, j = 1, \ldots, n$, and n, is the number of training samples. The kernelized dictionary dimension of the training samples is reduced by kernel mapping if the number of atoms is smaller than feature dimension in original dictionary and vice-versa. In KSRC, the classification task is executed by replacing D and y in SRC problem with D_{kernel} and y_{kernel}. Hence, the l_1 minimization problem for KSRC is expressed as following equations.

$$\min \|x\|_1 \quad s.t.\ y_{kernel} = D_{kernel} x \tag{15.14}$$

$$\min \|x\|_1 \quad s.t. \quad \|y_{kernel} - D_{kernel} x\|_2 \leq \varepsilon \tag{15.15}$$

The algorithm summarizes the complete classification procedure of KSRC is shown as follows.

Algorithm 2: Kernel Sparse Representation Classifier (KSRC)

1. Step 1: The input for KSRC are a matrix of training samples form a dictionary $D = \left[D_1, D_2, \ldots, D_c\right] \in R^{m\times n}$ for c classes, a test sample $y \in R^m$ and an optional error tolerance $\varepsilon > 0$.
2. Kernelize of D and y to yield D_{kernel} and y_{kernel}
3. Normalize the atoms of D_{kernel} and y_{kernel} to have unit l_2 norm.
4. Solve the l_2 minimization problem in equation 21 or 22 using SPGL1 toolbox:
5. Compute the residuals $r_i\left(y_{kernel}\right) = \left\|y_{kernel} - D_{kernel}\zeta_i x\right\|_2$ *for* $i = 1, \ldots\ldots, c$.
6. The class of the given test sample, y is determined by $identity(y_{kernel}) = \min_i r_i(y_{kernel})$.

15.4 Wrist-mounted ECG Sensor Design

In order to measure the heartbeat signal from part of the body, this study proposes two sensors in contact with the skin hence by touching one sensor electrode with each hand. The design of the proposed heartbeat data acquisition is illustrated as in Fig. 15.3 as a wrist-mounted device containing two electrodes. The first electrode is laid on the back of the proposed device which will permanently contact with user's wrist whereas the second electrode is front-facing. This electrode will contact with finger of the other hand. By touching the second electrode, the data is then measured and will be passed to mobile device or PC via Bluetooth.

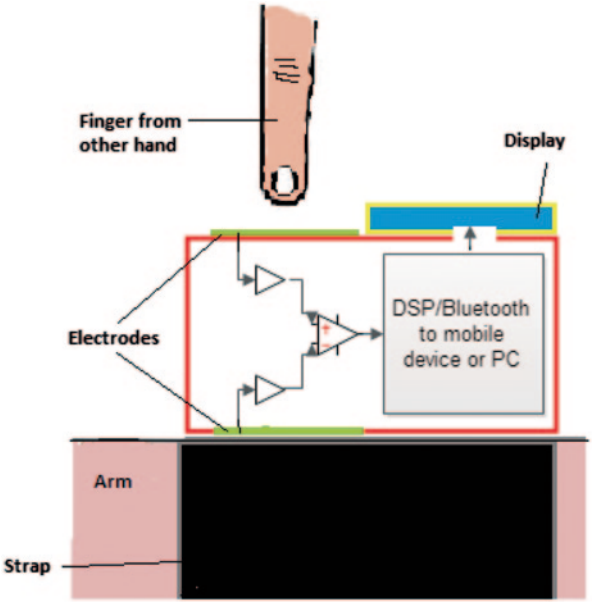


Fig. 15.3 Architecture of wrist-mounted heartbeat data acquisition

After the heartbeat data is collected, the digitization process is executed and the same steps as discussed in the methodology part i.e. pre-emphasis, and segmentation are followed. Consequently, instead of using MFCC techniques, any frequency transform methods such as wavelet transform can be used to obtain the ECG parameters. Sample of one ECG signal obtained from the proposed heartbeat data acquisition is shown in Fig. 15.4 below.

15.5 Result and Discussion

In this section, performances of the system based on KSRC and the feasibility of using the wrist-mounted sensor design are evaluated and discussed. The result of segmentation of S1 and S2 heart sounds is illustrated in Fig. 15.5. According to the figure, the STA and ZCR based segmentation method is able to correctly segment out S1 and S2 heart sounds.

The heart sounds used in this paper consists of heart sounds from 15 participants which are randomly selected from HSCT11 open database. A total of 775 heart sounds samples are divided in to two groups which are training sample and testing samples. Twenty samples of heart sounds from each participant are used as training samples while the rest are used as testing samples. Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are also implemented so as to validate the performances of SRC and KSRC method. In this work, the value of $k = 3$ is used for kNN and polynomial kernel is adopted for SVM.

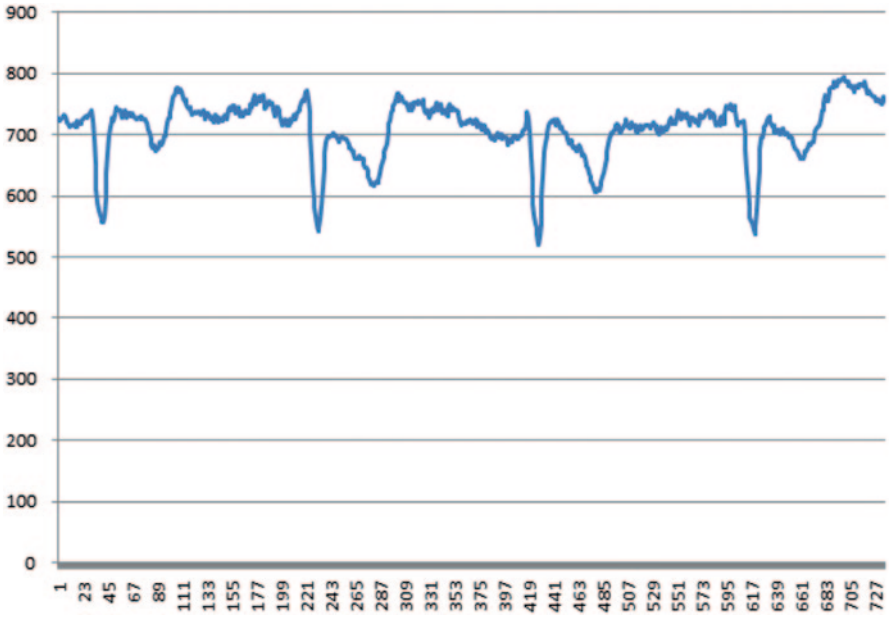


Fig. 15.4 ECG signal obtained from the wrist-mounted heartbeat data acquisition

The recognition methods have been tested as a function of two parameters i.e., the number of training samples per participant and the feature dimension of heart sound samples. In first experiment, *N* numbers of training samples are randomly selected from the heart sound data while the remaining is used for testing. The classification performances of heart sounds for various numbers of training samples using SVM, SRC, KSRC and KNN are shown in Table 15.2.

In second experiment, the effect of feature dimensions is also investigated by using three different feature lengths i.e., 200, 392 and 450. The classification performances of heart sounds for various feature dimensions of heart sound using SVM, SRC, KSRC and KNN are shown as in Table 15.3.

The results in Table 15.2 and 15.3 reveal that KSRC achieves the highest recognition rate compared to the other classifier for 20 training samples and 392 feature dimensions. KSRC shows very good results which have proved its ability in classification application not only on image data but heart sound (mono-dimensional) data as well.

Consequently, Fig. 15.6 below presents the performances of heartbeat biometric system based on ECG data which has been collected by using the suggested waist-mounted heartbeat sensor. A promising result has been observed for the preliminary data collected using the proposed design where 2 out of 5 subjects are able to be verified with accuracy almost 100 %.

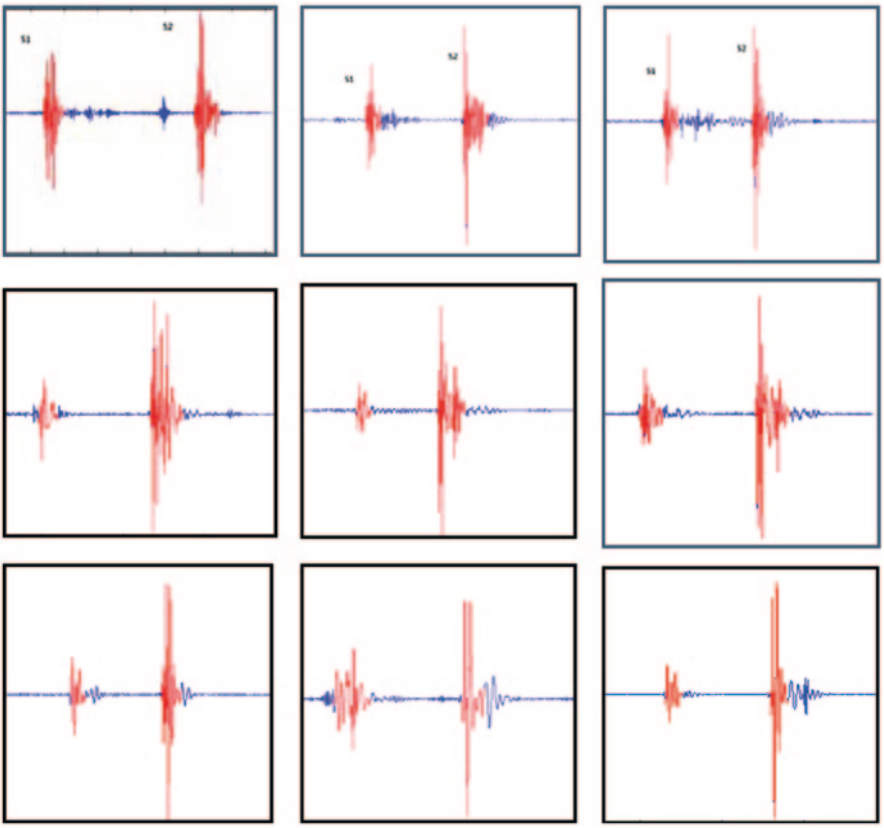


Fig. 15.5 Segmentation of S1 and S2 heart sounds from three different participants. Each row represents each participant

Table 15.2 Classification performances based on various numbers of training samples and various classifiers

Classifier	Number of training samples from each participant		
	10 (%)	15 (%)	20 (%)
SVM	76.96	81.09	84.87
SRC	75.52	81.18	84.45
KSRC	78.72	82.00	85.45
KNN (K = 3)	70.40	69.82	78.78

15.6 Conclusions

Two new techniques i.e. KSRC classifier and wrist-mounted sensor have been successfully implemented in this study. A good result obtained by KSRC in classifying heart's sound data proves that its capability as classifier to be used in heartbeat

Table 15.3 Classification performances based on various feature dimensions of heart sound features and various classifiers

Classifier	Feature Dimension of Heart Sound Features		
	200 (%)	392 (%)	450 (%)
SVM	82.98	84.87	84.24
SRC	82.14	84.45	84.03
KSRC	79.41	85.45	82.35
KNN (K = 3)	77.31	78.78	75.42

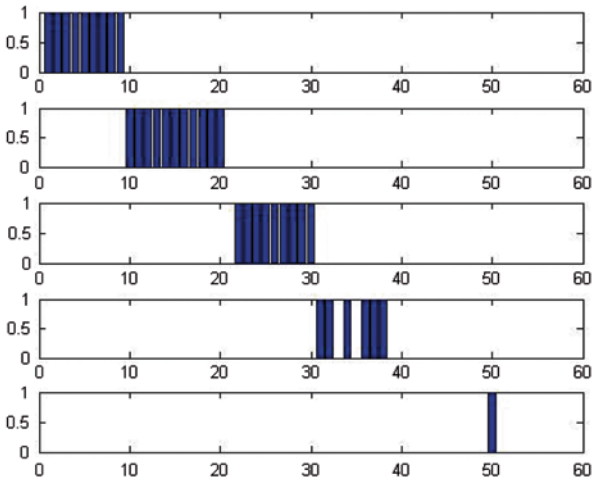


Fig. 15.6 Identification performance of selected participants using data collected from the proposed wrist-mounted heartbeat sensor

based biometric system. Due to this promising performance, a proposed wrist-mounted sensor to measure ECG signals has been designed and evaluated in this study. Further work toward the feasibility of system to be used in real time implementation has been proved in this study.

Acknowledgements This work was supported by Universiti Sains Malaysia and Fundamental Research Grant Scheme (6071266).

References

1. Becker S, Bobin J, Candès EJ (2011) NESTA: a fast and accurate first-order method for sparse recovery. SIAM Journal on Imaging Sciences 4:1–39
2. Beritelli F, Serrano S (2007) Biometric identification based on frequency analysis of cardiac sounds. Information Forensics and Security, IEEE Transactions on 2:596–604
3. Candès EJ, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics 59:1207–1223
4. Candes EJ, Tao T (2006) Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? Information Theory, IEEE Transactions on 52:5406–5425

5. Candès EJ, Wakin MB (2008) An introduction to compressive sampling. Signal Processing Magazine, IEEE 25:21–30
6. Donoho DL (2006) For most large underdetermined systems of linear equations the minimal Communications on Pure and Applied Mathematics 59:797–829
7. El-Bendary N, Al-Qaheri H, Zawbaa HM et al. (2010) HSAS: Heart Sound Authentication System. In: Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on. p 351–356
8. Figueiredo MaT, Nowak RD, Wright SJ (2007) Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. Selected Topics in Signal Processing, IEEE Journal of 1:586–597
9. Kim S, Eriksson T, Kang H-G et al. (2004) A pitch synchronous feature extraction method for speaker recognition. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP' 04). IEEE International Conference on. p I-405–408 vol. 401
10. Li Z, Wei-Da Z, Pei-Chann C et al. (2012) Kernel Sparse Representation-Based Classifier. Signal Processing. IEEE Transactions on 60:1684–1695
11. Memon S, Lech M, Ling H (2009) Using information theoretic vector quantization for inverted MFCC based speaker verification. In: Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on. p 1–5
12. Phua K, Dat TH, Chen J et al. (2006) Human identification using heart sound. In: Second International Workshop on Multimodal User Authentication, Toulouse, France
13. Spadaccini A, Beritelli F (2012) Performance Evaluation of Heart Sounds Biometric Systems on An Open Dataset. In: Proceedings of the 5th IAPR International Conference on Biometrics
14. Wright J, Yang AY, Ganesh A et al. (2009) Robust Face Recognition via Sparse Representation. Pattern Analysis and Machine Intelligence. IEEE Transactions on 31:210–227
15. Xiaoling Y, Baohua T, Jiehua D et al. (2010) Comparative Study on Voice Activity Detection Algorithm. In: Electrical and Control Engineering (ICECE), 2010 International Conference on. p 599–602
16. Yu K, Ji L, Zhang X (2002) Kernel nearest-neighbor algorithm. Neural Processing Letters 15:147–156
17. Zhao Z, Shen Q, Ren F (2013) Heart Sound Biometric System Based on Marginal Spectrum Analysis. Sensors 13:2530–2551

Chapter 16
Laminar and Turbulent Simulations of Several TVD Schemes in Two-Dimensions—Part I—Results

E. S. G. Maciel

Abstract This work, first part of this study, describes five numerical tools to perform perfect gas simulations of the laminar and turbulent viscous flow in two-dimensions. The Van Leer, Harten, Frink, Parikh and Pirzadeh, Liou and Steffen Jr. and Radespiel and Kroll schemes, in their first- and second-order versions, are implemented to accomplish the numerical simulations. The Navier–Stokes equations, on a finite volume context and employing structured spatial discretization, are applied to solve the supersonic flow along a ramp in two-dimensions. Three turbulence models are applied to close the system, namely: Cebeci and Smith, Baldwin and Lomax and Sparlat and Allmaras. The convergence process is accelerated to the steady state condition through a spatially variable time step procedure. The results have shown that, with the exception of the Harten scheme, all other schemes have yielded the best result in terms of the prediction of the shock angle at the ramp.

Keywords Laminar and turbulent flows · TVD algorithms · Cebeci and Smith turbulence model · Baldwin and Lomax turbulence model · Sparlat and Allmaras turbulence model

16.1 Introduction

Conventional non-upwind algorithms have been used extensively to solve a wide variety of problems [1]. Conventional algorithms are somewhat unreliable in the sense that for every different problem (and sometimes, every different case in the same class of problems) artificial dissipation terms must be specially tuned and judicially chosen for convergence. Also, complex problems with shocks and steep compression and expansion gradients may defy solution altogether.

Upwind schemes are in general more robust but are also more involved in their derivation and application. Some upwind schemes that have been applied to the Euler equations are, for example, [2–6]. To comments about these methods and to

E. S. G. Maciel (✉)
Department of Energy Engineering, Foundation University of Great Dourados, Rodovia Dourados—Itahum, km 12, Postal Box 364, Dourados, MS, 79804-970 Brazil
e-mail: edisavio@edissonsavio.eng.br

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,* Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_16,

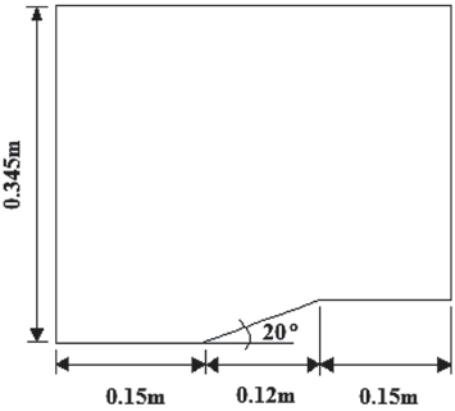


Fig. 16.1 Ramp configuration

the motivation of this study the reader is encouraged to read the first part of this study, THEORY, in [7].

This work, first part of this study, describes five numerical tools to perform perfect gas simulations of the laminar and turbulent viscous flow in two-dimensions. The [2–6] schemes, in their first- and second-order versions, are implemented to accomplish the numerical simulations. The Navier–Stokes equations, on a finite volume context and employing structured spatial discretization, are applied to solve the supersonic flow along a ramp in two-dimensions. Three turbulence models are applied to close the system, namely: [8–10]. On the one hand, the second-order version of the [2, 4–6] schemes are obtained from a "MUSCL" extrapolation procedure, whereas on the other hand, the modified flux function approach is applied in the Harten [3] scheme for the same accuracy. The convergence process is accelerated to the steady state condition through a spatially variable time step procedure, which has proved effective gains in terms of computational acceleration (see [11, 12]). The results have shown that the [2, 4–6] schemes have yielded the best results in terms of the prediction of the shock angle at the ramp. Moreover, the wall pressure distribution is also better predicted by the [3] scheme. This work treats the laminar first- and second-order and the [8–10] second-order results obtained by the five schemes.

16.2 Results

One problem was studied in this work, namely: the viscous supersonic flow along a ramp geometry. The ramp configuration is detailed as also the type of boundary contours. These characteristics are described in Figs. 16.1 and 16.2.

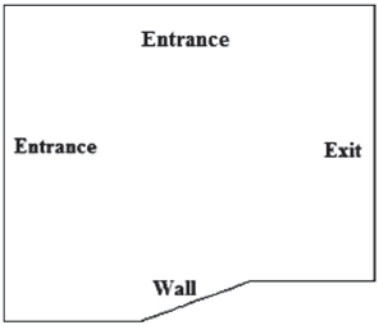


Fig. 16.2 Ramp computational domain

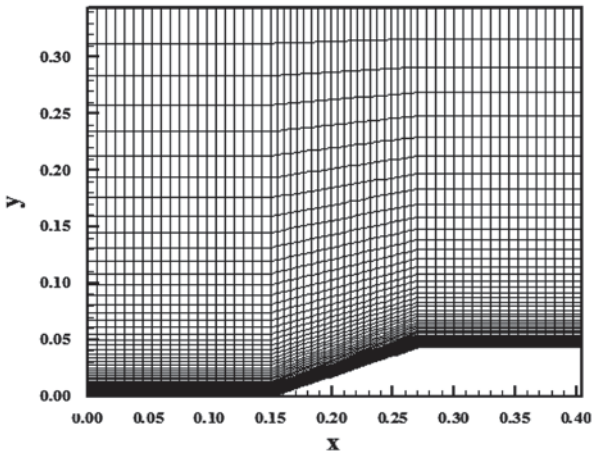


Fig. 16.3 Ramp viscous mesh

Numerical experiments were run on a Notebook computer with Intel Core i7 processor of 2.3 GHz of clock and 8.0 GBytes of RAM. The criterion adopted to reach the steady state was to consider a reduction of three (3) orders of magnitude in the value of the maximum residual in the calculation domain, a typical CFD community criterion. The maximum residual is defined as the maximum value obtained from the discretized equations in the overall domain, considering all conservation equations. The initial conditions to the ramp problem are described in Table 16.1. Figure 16.3 exhibits the mesh employed in the calculation of the viscous flow to the ramp problem. An exponential stretching of 10.0 % was applied close to the wall, in the η direction, to capture the viscous phenomena. A total of 3,540 rectangular cells and 3,660 nodes, which is equivalent to a mesh of 61×60 nodes on a finite difference context, is employed.

The Reynolds number is equal to 1.613×10^5, a turbulent flow. Three turbulence models were studied, namely: [8–10]. Two algebraic and an one-equation models are implemented.

Table 16.1 Initial conditions to the studied problem

Problem	Property	Value
Ramp	Freestream Mach, M∞	2.0
	Attack angle, °	0.0
	Ratio of specific heats, γ	1.4

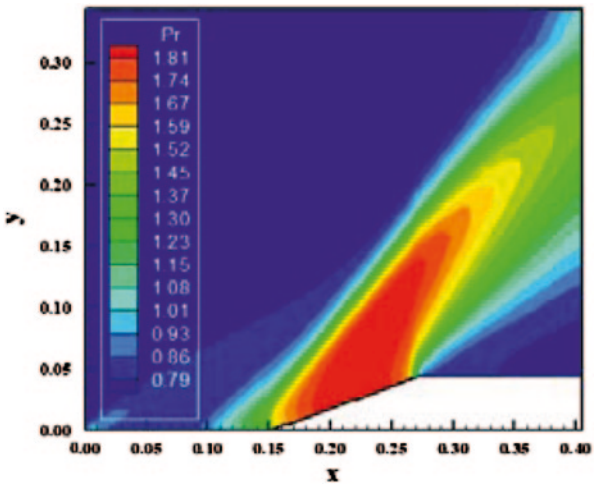


Fig. 16.4 Pressure contours (VL-1st Order)

16.2.1 Laminar Viscous Results

The laminar viscous results are divided in two solution groups: the first order and the second order solutions. The first order results are presented here to serve as a benchmark to compare the second order viscous results, aiming to distinguish the excessive diffusion characteristics resulting from the former, as referenced by the CFD literature.

16.2.1.1 First-Order Results

Figures 16.4, 16.5, 16.6, 16.7 and 16.8 presents the pressure contours obtained by the [2–6] schemes, respectively. All schemes capture a strong viscous interaction typical of viscous flow simulations, at the ramp entrance. A weak shock wave is formed ahead of the ramp due to the boundary layer detachment. The [3] scheme captures the biggest detachment region of the boundary layer, resulting in the biggest circulation bubble formation. Moreover, the [2] scheme captures the most severe pressure field, characterizing this one as more conservative than the others schemes.

Figure 16.9 presents the wall pressure distributions of all schemes. They are compared with the oblique shock wave theory results and with the Prandtl–Meyer

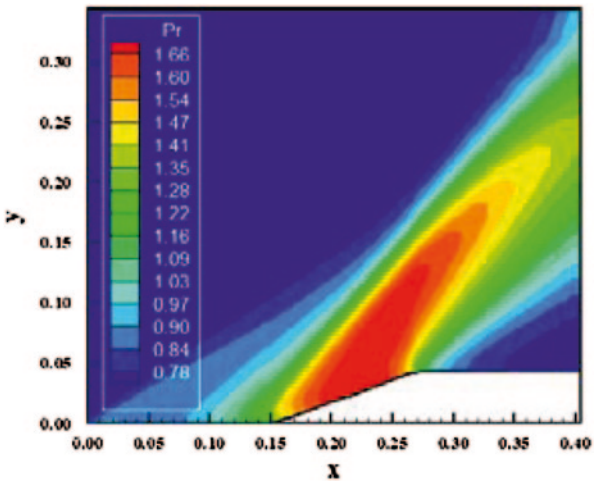


Fig. 16.5 Pressure contours (H-1st Order)

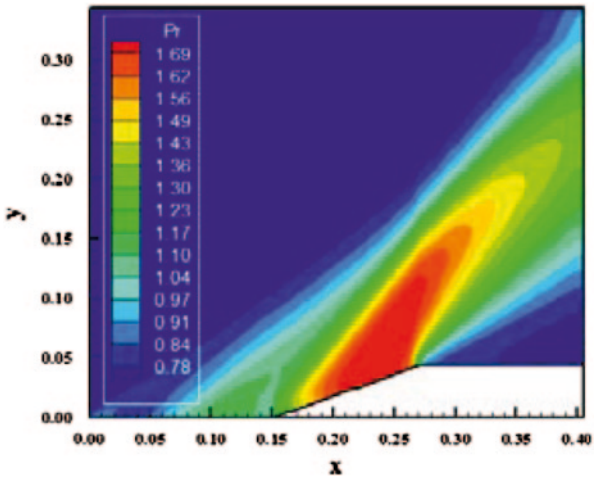


Fig. 16.6 Pressure contours (FPP-1st Order)

expansion fan results. It is important to observe that this theoretical profile is the correct to be obtained in a viscous simulation, because of the pressure gradient in the normal direction from the wall is equal to zero, according to the boundary layer theory. Hence, the pressure at the boundary layer edge is imposed to the wall pressure.

As can be seen, the [2] solution is closer to the pressure profile than the other solutions. The [3] scheme predicted the smallest severe shock than the others schemes. The expansion fan is better captured by the [5] scheme. Finally, the circulation bubble closes to the ramp corner is exhibited in Figs. 16.10, 16.11, 16.12, 16.13 and 16.14. The [4–5] solutions show bigger circulation bubbles than the other solutions.

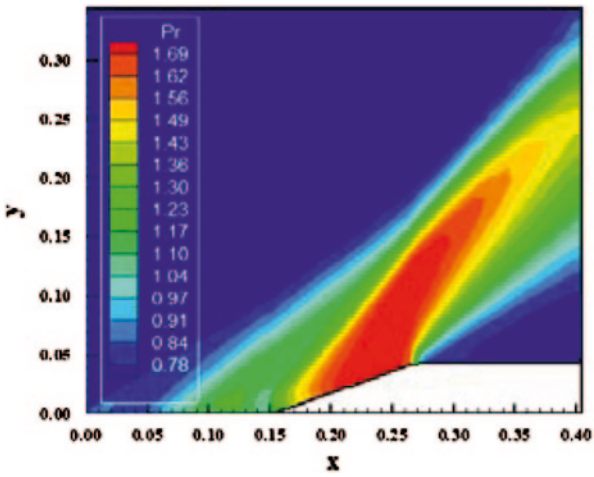


Fig. 16.7 Pressure contours (LS-1st Order)

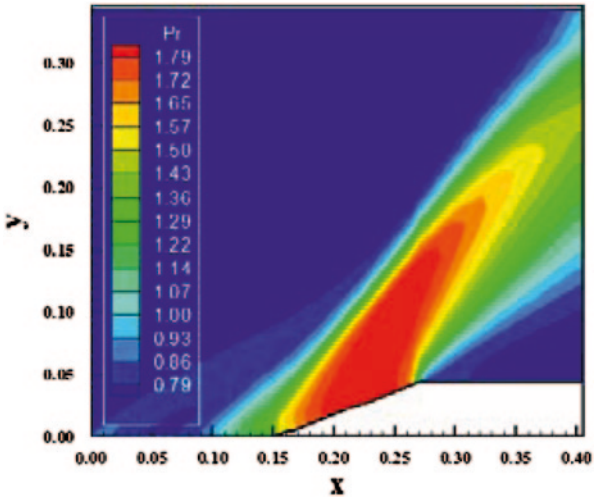


Fig. 16.8 Pressure contours (RK 1st Order)

16.2.1.2 Second-Order Results/TVD

For the second order results, a minmod non-linear limiter was employed in the [2–6] schemes. The [5–6] schemes did not present converged results. Figures 16.15, 16.16 and 16.17 exhibit the pressure contours obtained by the [2–4] schemes. All solutions present a weak shock ahead of the ramp corner. This shock wave is formed far ahead the ramp corner. The pressure field is also more severe in the solution obtained by the [2] scheme, indicating this one as the most conservative.

Figure 16.18 shows the wall pressure distributions generated by the [2–4] schemes in their TVD versions. All solutions capture the circulation bubble formation, resulting from the boundary layer detachment. The [2] solution presents a

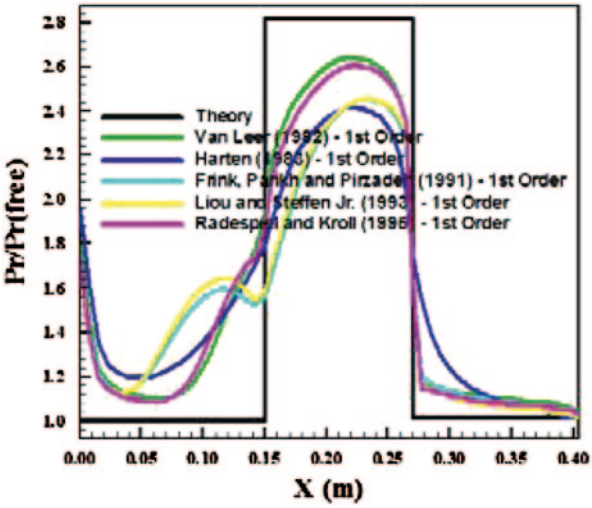


Fig. 16.9 Wall pressure distributions

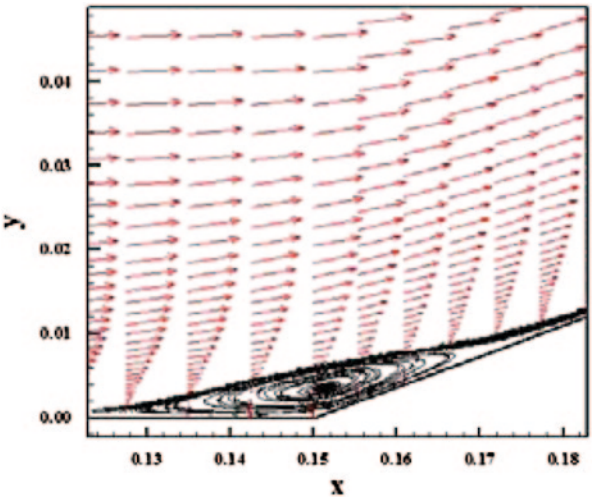


Fig. 16.10 Circulation bubble (VL-1st Order)

pressure distribution closer to the pressure plateau, whereas the [4] solution shows a more extent separation region.

Figures 16.19, 16.20 and 16.21 presents the formation of circulation bubble closes to the ramp corner obtained by [2–4] schemes. The circulation bubbles obtained by the [3], and [4] schemes are larger than the respective of the [2] scheme. As a resume of the present simulations, the [2] scheme was more conservative and more correct in physical terms, representing accurately the flow physics.

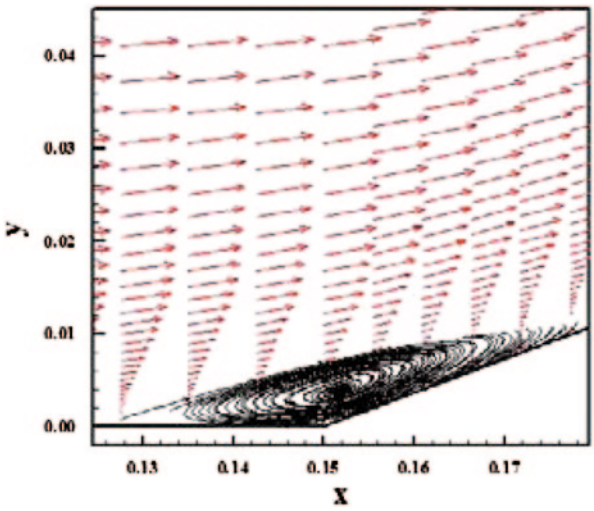


Fig. 16.11 Circulation bubble (H-1st Order)

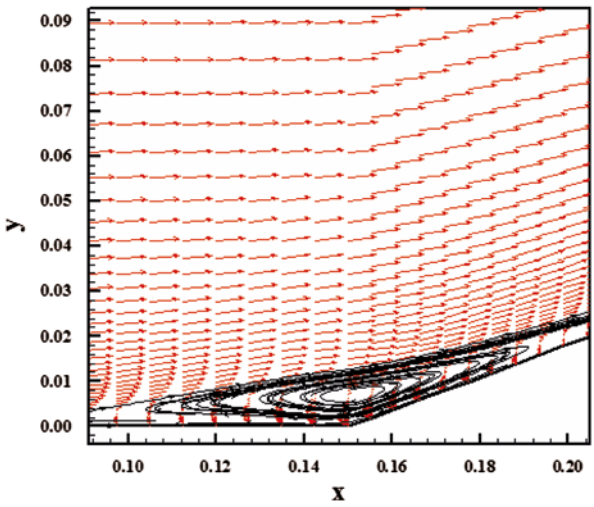


Fig. 16.12 Circulation bubble (FPP-1st Order)

16.2.2 Turbulent Viscous Results

16.2.2.1 Cebeci and Smith Results/TVD

Figures 16.22, 16.23, 16.24, 16.25 and 16.26 show the pressure contours obtained by the [2–6] schemes, respectively, as using the [8] turbulence model. All solutions practically ignore the existence of the weak shock ahead of the ramp corner. It indicates that the boundary layer detachment is negligible in all solutions and that the

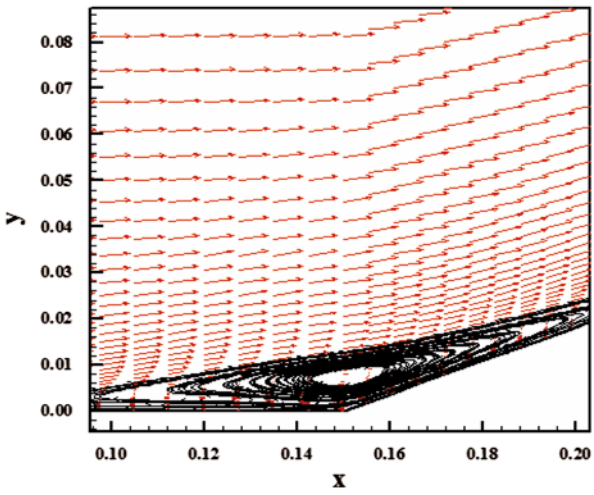


Fig. 16.13 Circulation bubble (LS-1st Order)

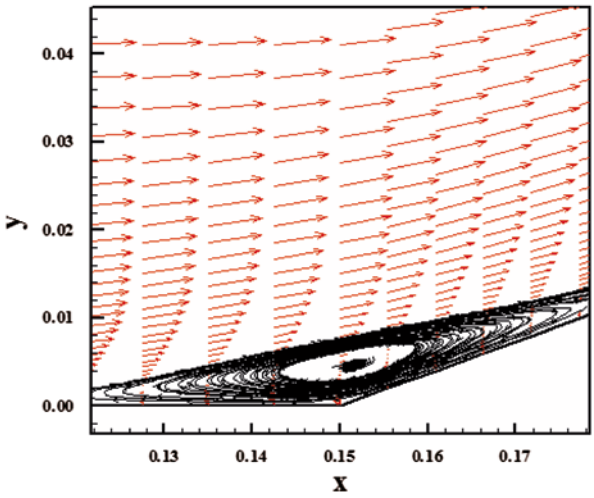


Fig. 16.14 Circulation bubble (RK-1st Order)

circulation bubble is reduced in size. The pressure field generated by the [3] scheme is the most severe in relation to those generated by the other schemes.

Figure 16.27 exhibits the wall pressure distributions obtained by the [2–6] algorithms, as using the [8] turbulence model. As can be observed, all solutions are very similar and agree better with the theoretical solution than in the laminar cases. The expansion fan pressure is better predicted by the [6] algorithm.

Figures 16.28, 16.29, 16.30, 16.31 and 16.32 show the circulation bubble formation close to the ramp corner. All solutions predicted a small circulation bubble, although that generated by the [2] scheme is larger than those generated by the other schemes. In resume, as can be observed the [8] turbulence model predicts a more

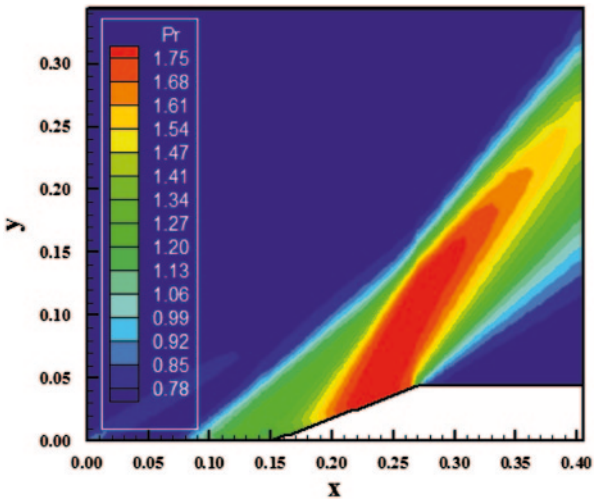


Fig. 16.15 Pressure contours (VL-TVD)

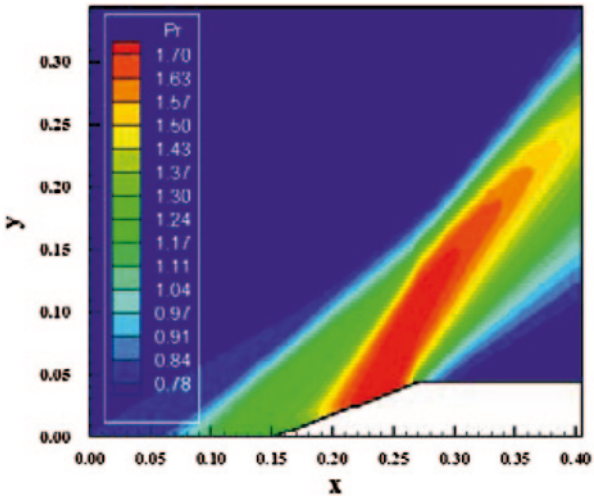


Fig. 16.16 Pressure contours (H-TVD)

energized boundary layer. With it, the weak shock wave ahead of the ramp corner is negligible and the circulation bubble presents a discrete formation.

16.2.2.2 Baldwin and Lomax Results/TVD

In this case, only the [2–4] schemes have presented converged results. Figures 16.33, 16.34 and 16.35 exhibit the pressure contours obtained by the [2–4] schemes, respectively, as using the [9] turbulence model. A weak shock wave is formed ahead of the ramp corner in all solutions. It is important to remember that such weak shock

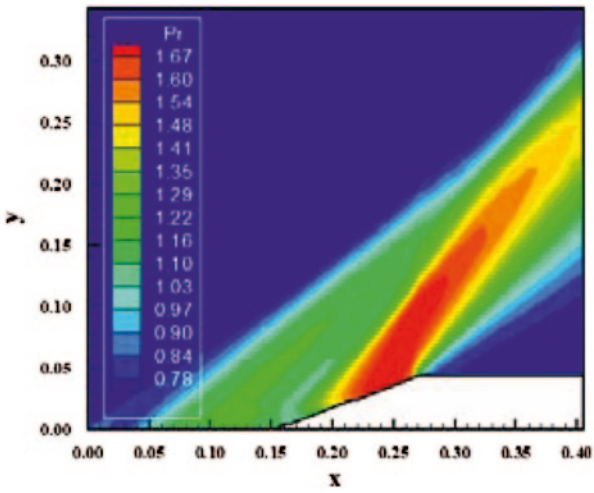


Fig. 16.17 Pressure contours (FPP-TVD)

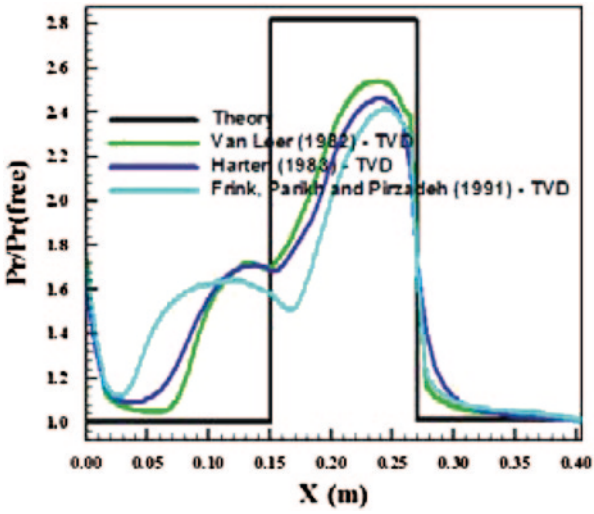


Fig. 16.18 Wall pressure distributions

wave is due to the boundary layer detachment which induces a false thick geometry at the ramp and the flow only see this thick geometry, originating the oblique shock wave. So, it is possible to distinguish that the effect of increasing boundary layer thickness is more pronounced in the [4] solution than in the other solutions. It also induces the expected behavior of a larger circulation bubble formed in the [4] solution. In terms of the pressure field, the [2] scheme presents the most severe pressure field, characterizing this algorithm as more conservative.

Figure 16.36 presents the wall pressure distributions generated by all algorithms. As noted, all solutions capture the circulation bubble formation close to the ramp corner, but all solutions differs from the theoretical solution (all under-predict

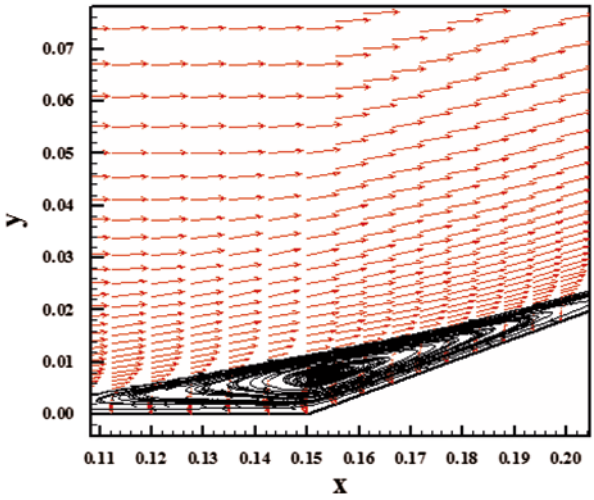


Fig. 16.19 Circulation bubble (VL-TVD)

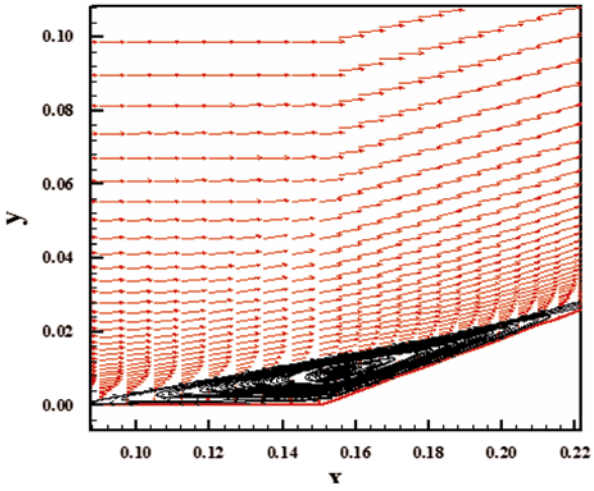


Fig. 16.20 Circulation bubble (H-TVD)

the shock plateau). Figures 16.37, 16.38 and 16.39 exhibit the circulation bubble formed close to the ramp corner generated by the [2–4] algorithms. The [4] scheme presents a bigger circulation bubble in extent and size than the others.

16.2.2.3 Sparlat and Allmaras Results/TVD

Only the [5] scheme did not present converged results. Figures 16.40, 16.41, 16.42 and 16.43 present the pressure contours obtained by the [2–4, 6] schemes, respectively, as using the [10] turbulence model. The [2] solution captures a small bound-

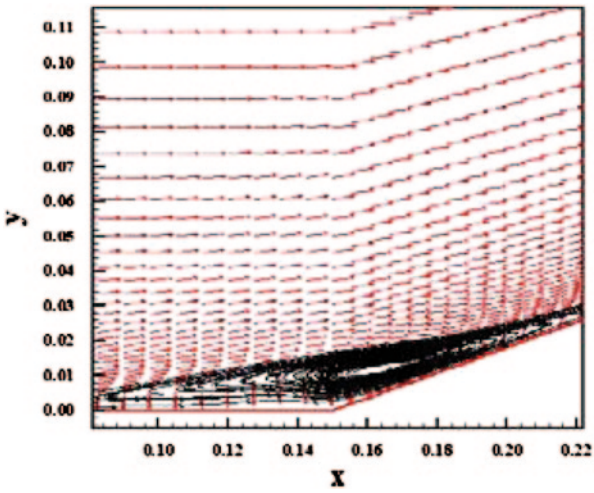


Fig. 16.21 Circulation bubble (FPP-TVD)

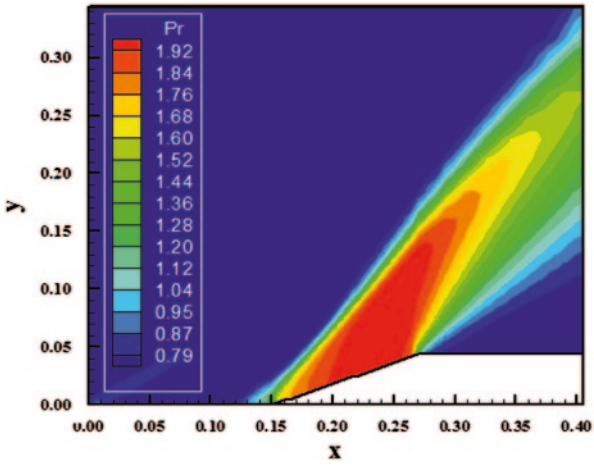


Fig. 16.22 Pressure contours (VL-CS)

ary layer detachment, which results in a less intense weak shock wave. The [4] solution captures the biggest boundary layer detachment, which results in a more intense weak shock wave. The pressure field generated by the [2] scheme is again the most severe in relation to those generated by the others schemes.

Figure 16.44 shows the wall pressure distributions obtained by the [2–4, 6] algorithms. All solutions capture the circulation bubble at the ramp corner. Moreover, the [2] pressure peak is close to the theoretical pressure plateau.

It is important to be mentioned here that the best behaviour to the pressure plateau was obtained by the [8] turbulence model in spite of the loss of physical meaning of the flow (loss of the circulation bubble formation).

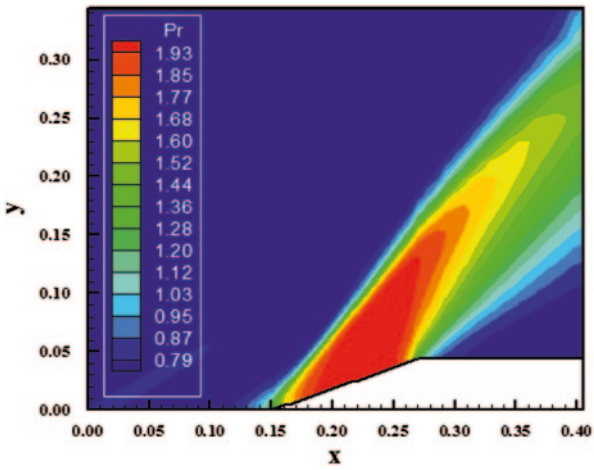


Fig. 16.23 Pressure contours (H-CS)

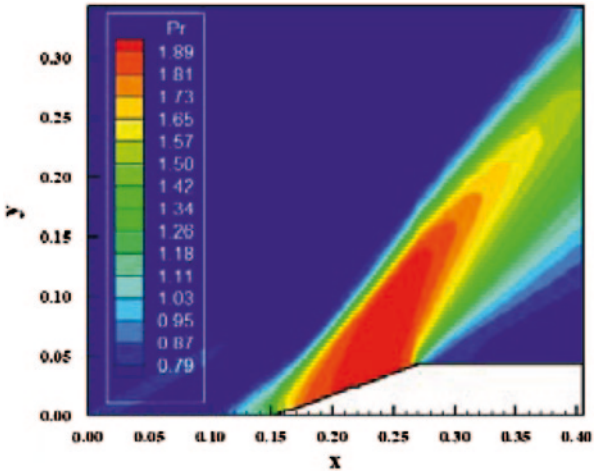


Fig. 16.24 Pressure contours (FPP-CS)

Figures 16.45, 16.46, 16.47 and 16.48 exhibit the circulation bubble captured by the [2–4, 6] schemes, respectively, as using the [10] turbulence model. As can be seen, the [4] solution generates the largest bubble region than the other solutions.

16.2.3 Quantitative Analysis

One way to quantitatively verify if the solutions generated by each scheme are satisfactory consists in determining the shock angle of the oblique shock wave, β, measured in relation to the initial direction of the flow field. [13] (pages 352 and 353) presents a diagram with values of the shock angle, β, to oblique shock waves.

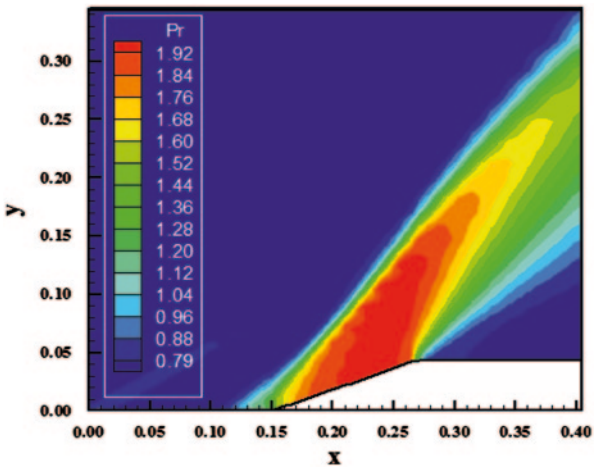


Fig. 16.25 Pressure contours (LS-CS)

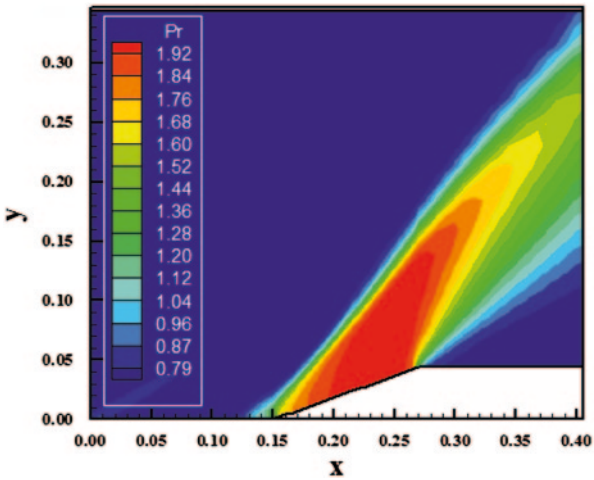


Fig. 16.26 Pressure contours (RK-CS)

The value of this angle is determined as function of the freestream Mach number and of the deflection angle of the flow after the shock wave, ϕ. To $\phi = 20°$ (ramp inclination angle) and to a freestream Mach number equals to 2.0, it is possible to obtain from this diagram a value to β equals to 53.0°. Using a transfer in Figs. 16.4, 16.5, 16.6, 16.7 and 16.8 (laminar, first order), Figs. 16.15, 16.16 and 16.17 (laminar, second order), Figs. 16.22, 16.23, 16.24, 16.25 and 16.26 (CS), Figs. 16.33, 16.34 and 16.35 (BL), and Figs. 16.40, 16.41, 16.42 and 16.43 (SA), it is possible to obtain the values of β to each scheme and to each studied case, as well the respective errors, shown in Table 16.2. It is possible to distinguish that only the [3] scheme did not capture the exact value of the oblique shock wave angle. All other schemes capture this exact value in a particular case. The [9] turbulence model was the most exact because allows the [2] and [4] schemes to capture accurately the shock angle.

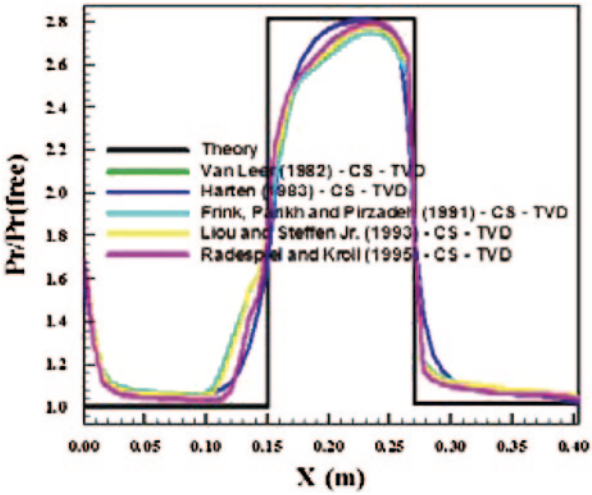


Fig. 16.27 Wall pressure distributions

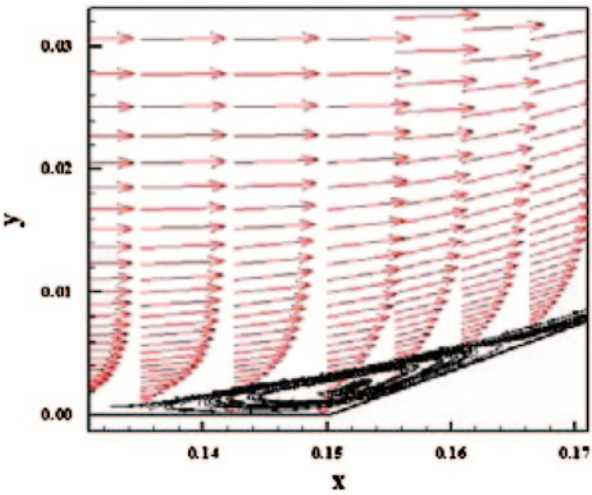


Fig. 16.28 Circulation bubble (VL-CS)

Table 16.2 Values of the oblique shock wave angle and percentage errors

Case	Lam., 1st	Lam., 2nd	CS, TVD	BL, TVD	SA, TVD
VL	51.0	56.4	51.0	53.0	51.6
Error (%)	3.77	6.42	3.77	0.00	2.64
H	49.3	55.0	52.5	55.0	51.7
Error (%)	6.98	3.77	0.94	3.77	2.45
FPP	52.4	51.6	51.4	53.0	55.0
Error (%)	1.13	2.64	3.02	0.00	3.77
LS	53.0	–	52.0	–	–
Error (%)	0.00	–	1.89	–	–
RK	51.0	–	51.2	–	53.0
Error (%)	3.77	–	3.40	–	0.00

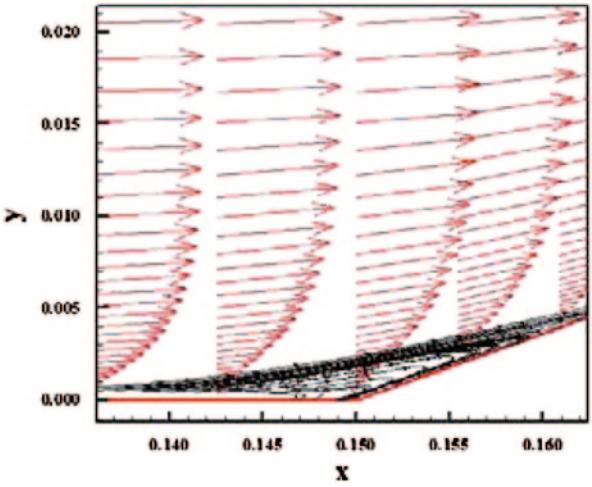


Fig. 16.29 Circulation bubble (H-CS)

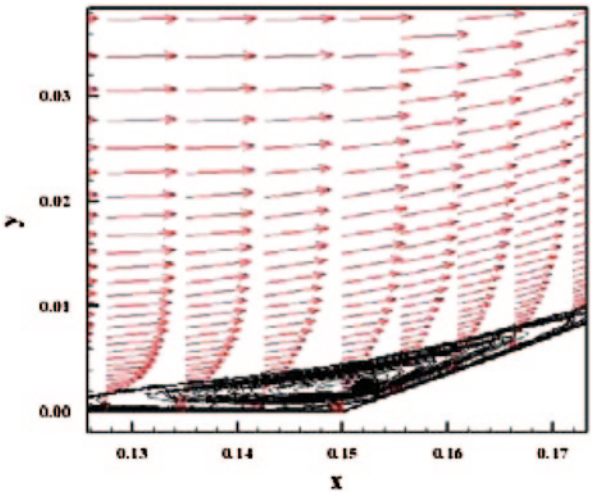


Fig. 16.30 Circulation bubble (FPP-CS)

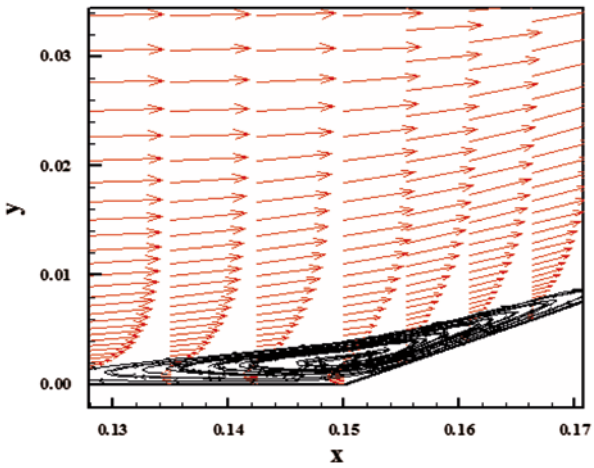


Fig. 16.31 Circulation bubble (LS-CS)

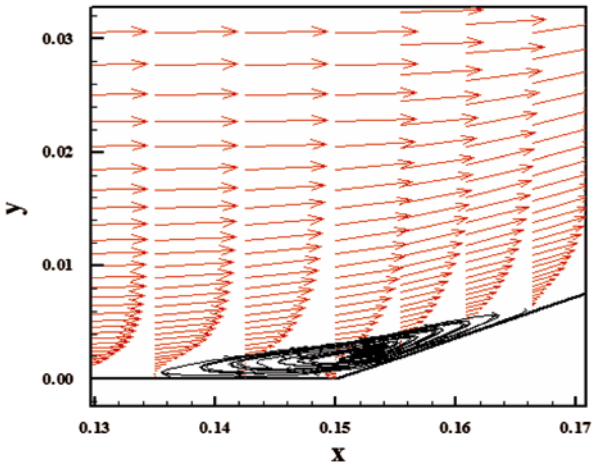


Fig. 16.32 Circulation bubble (RK-CS)

16.3 Conclusions

This work, first part, describes five numerical tools to perform perfect gas simulations of the laminar and turbulent viscous flow in two-dimensions. The [2–6] schemes, in its first- and second-order versions, are implemented to accomplish the numerical simulations. The Navier–Stokes equations, on a finite volume context and employing structured spatial discretization, are applied to solve the supersonic flow along a ramp in two-dimensions. Three turbulence models are applied to close the system, namely: [8–10]. On the one hand, the second-order version of the

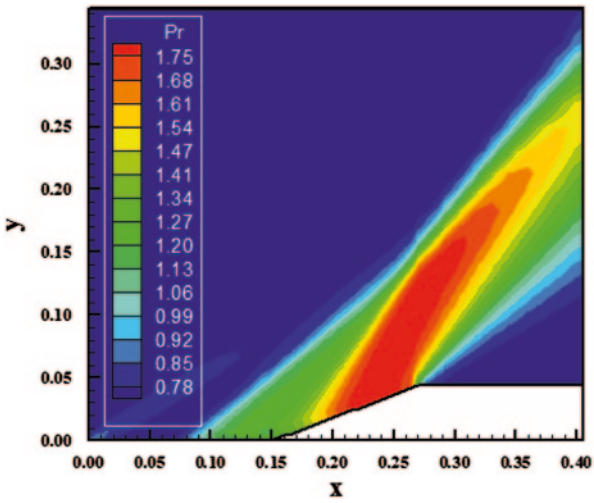


Fig. 16.33 Pressure contours (VL-BL)

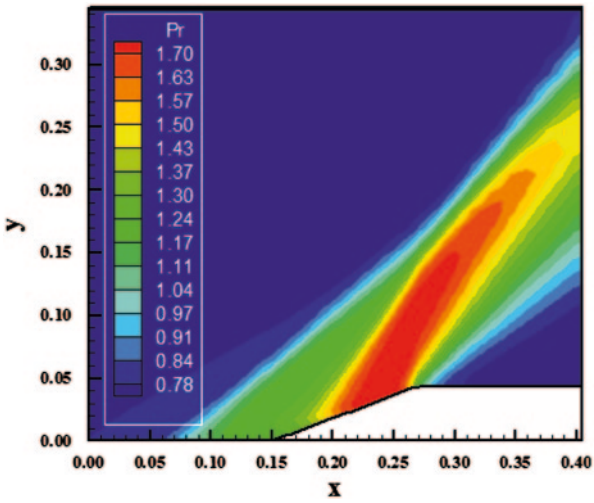


Fig. 16.34 Pressure contours (H-BL)

[2, 4–6] schemes are obtained from a "MUSCL" extrapolation procedure, whereas on the other hand, the modified flux function approach is applied in the [3] scheme for the same accuracy. The convergence process is accelerated to the steady state condition through a spatially variable time step procedure, which has proved effective gains in terms of computational acceleration (see [11, 12]). The results have shown that the [2, 4–6] schemes have yielded the best results in terms of the prediction of the shock angle at the ramp.

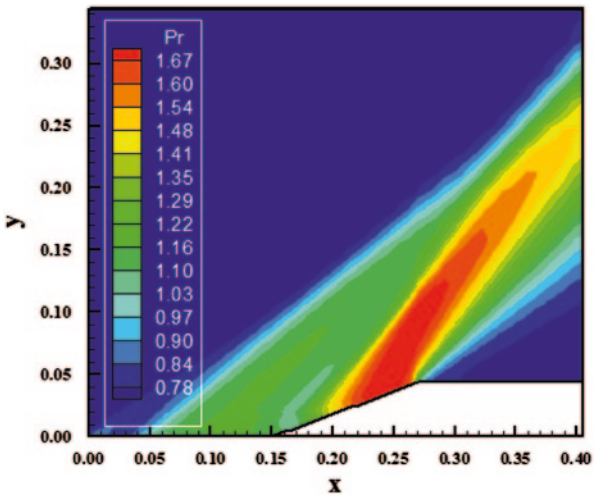


Fig. 16.35 Pressure contours (FPP-BL)

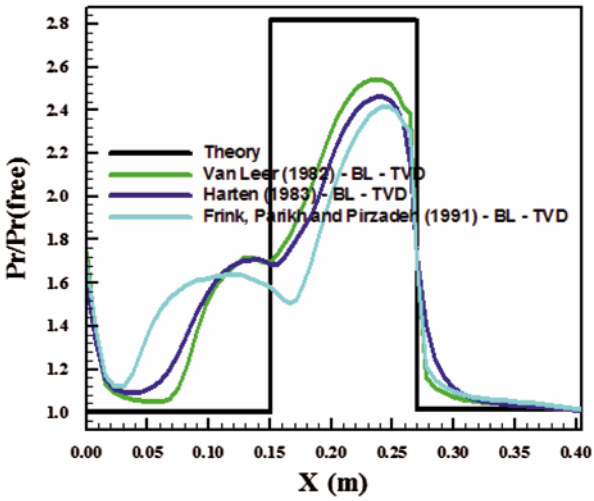


Fig. 16.36 Wall pressure distributions

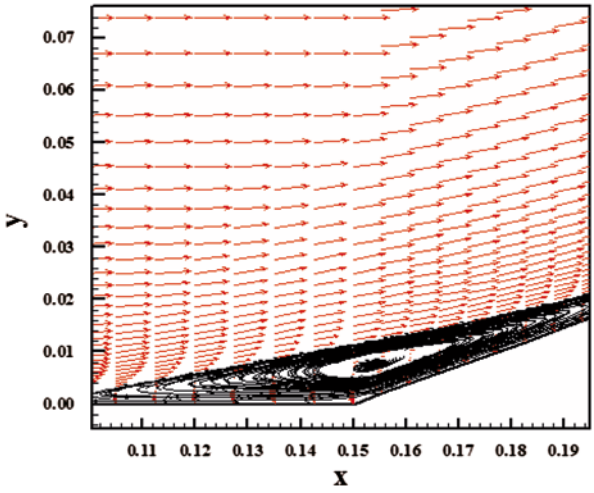


Fig. 16.37 Circulation bubble (VL-BL)

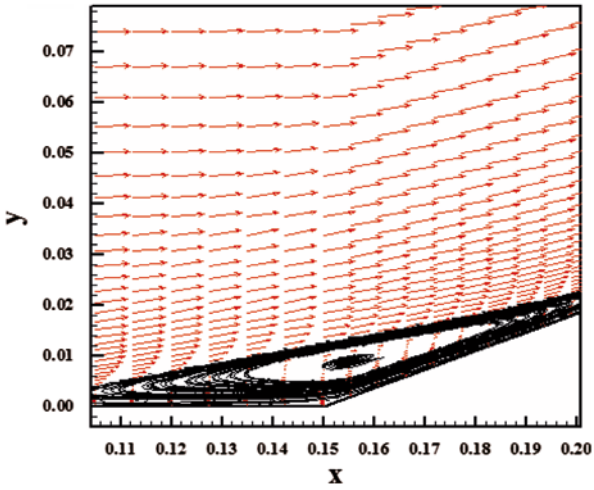


Fig. 16.38 Circulation bubble (H-BL)

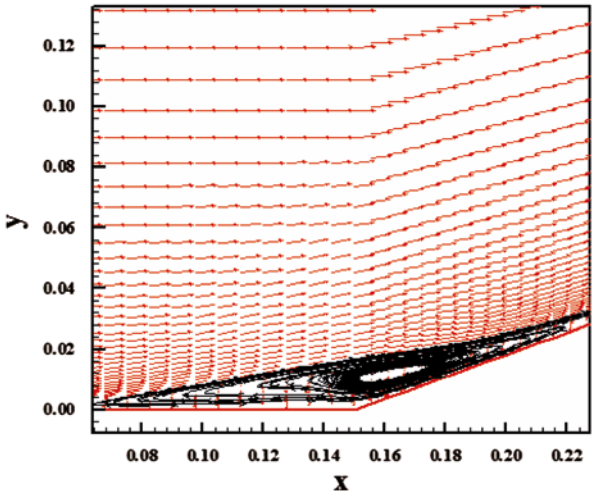


Fig. 16.39 Circulation bubble (FPP-BL)

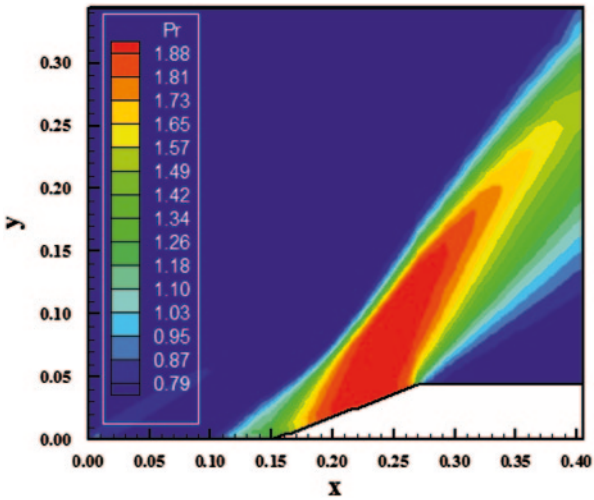


Fig. 16.40 Pressure contours (VL-SA)

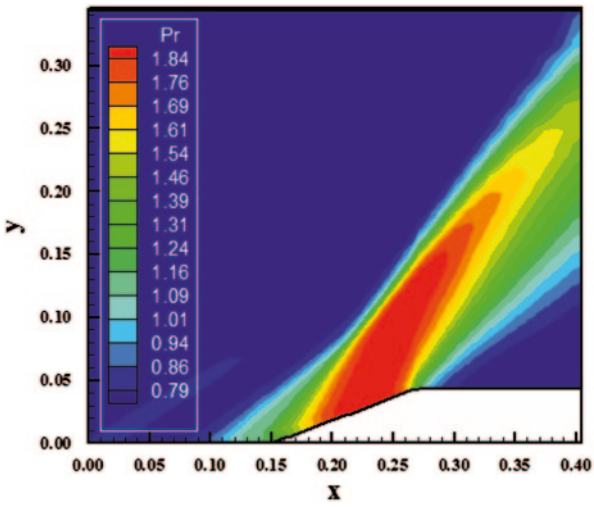


Fig. 16.41 Pressure contours (H-SA)

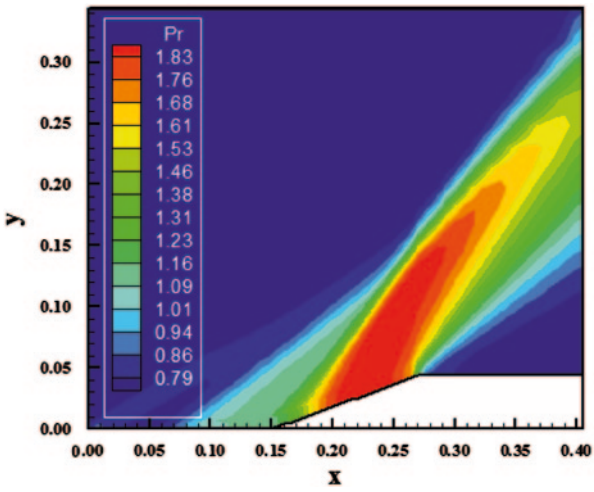


Fig. 16.42 Pressure contours (FPP-SA)

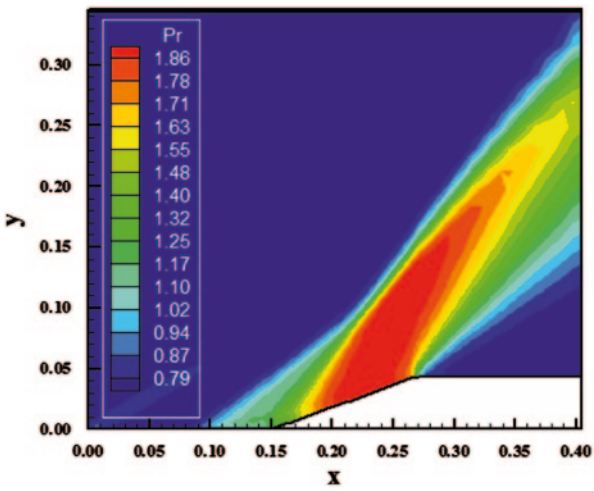


Fig. 16.43 Pressure contours (RK-SA)

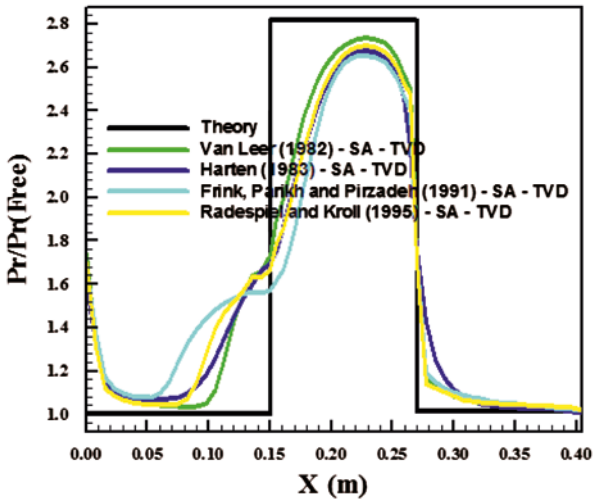


Fig. 16.44 Wall pressure distributions

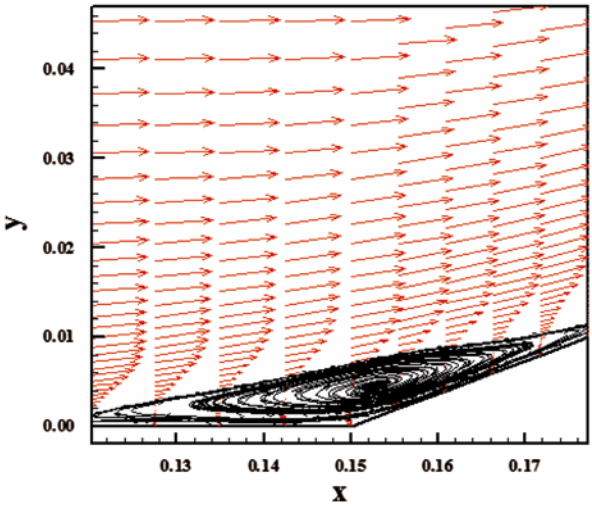


Fig. 16.45 Circulation bubble (VL-SA)

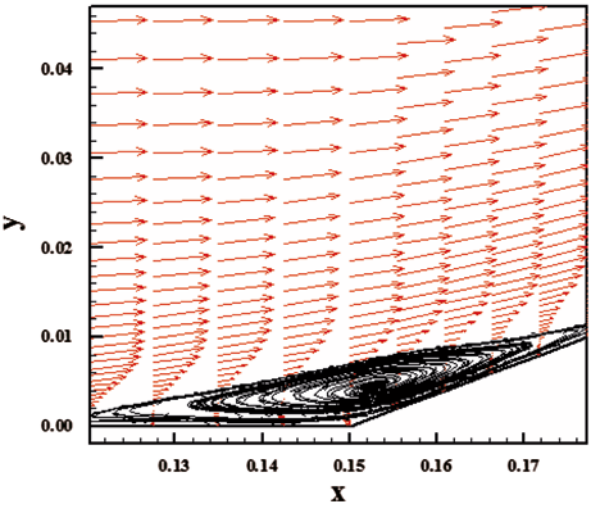


Fig. 16.46 Circulation bubble (H-SA)

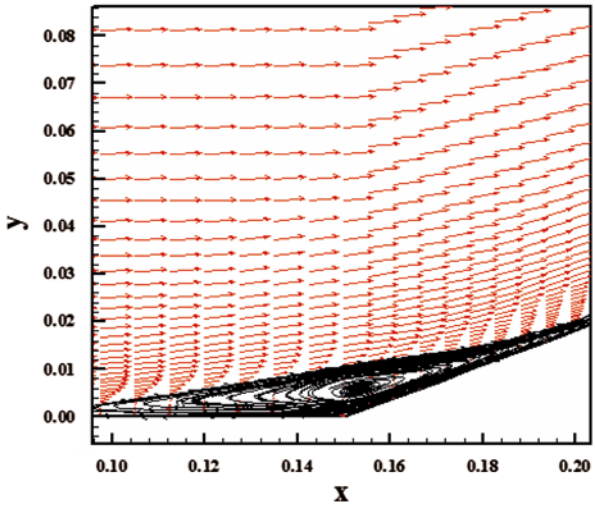


Fig. 16.47 Circulation bubble (FPP-SA)

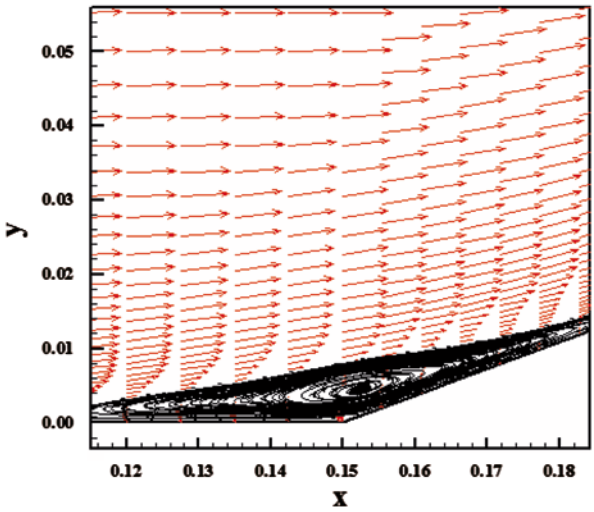


Fig. 16.48 Circulation bubble (RK-SA)

References

1. Kutler P (1975) Computation of Three-Dimensional, Inviscid Supersonic Flows, Lecture Notes in Physics, 41: 287–374.
2. Van Leer B (1982) Flux-Vector Splitting for the Euler Equations, Proceedings of the 8th International Conference on Numerical Methods in Fluid Dynamics, E. Krause, Editor, Lecture Notes in Physics, 170: 507–512, Springer-Verlag, Berlin.
3. Harten A (1983) High Resolution Schemes for Hyperbolic Conservation Laws, Journal of Computational Physics, 49: 357–393.
4. Frink NT, Parikh P, and Pirzadeh S (1991) Aerodynamic Analysis of Complex Configurations Using Unstructured Grids, AIAA 91-3292-CP.
5. Liou M, and Steffen Jr. CJ (1993) A New Flux Splitting Scheme, Journal of Computational Physics, 107: 23–39.
6. Radespiel R, and Kroll N (1995) Accurate Flux Vector Splitting for Shocks and Shear Layers, Journal of Computational Physics, 121: 66–78.
7. Maciel ESG (2014) Laminar and Turbulent Simulations of Several TVD Schemes in Two-Dimensions—Part I—Theory, Submitted to the *VIII National Congress of Mechanical Engineering (VIII CONEM)*, Brazil (under review).
8. Cebeci T, and Smith AMO (1970) A Finite-Difference Method for Calculating Compressible Laminar and Turbulent Boundary Layers, Journal of Basic Engineering, Trans. ASME, Series B, 92: 523–535.
9. Baldwin BD, and Lomax H (1978) Thin Layer Approximation and Algebraic Model for Separated Turbulent Flows, AIAA Paper 78–257.
10. Sparlat PR, and Allmaras SR (1992) A One-Equation Turbulence Model for Aerodynamic Flows, AIAA Paper 92-0439.
11. Maciel ESG (2005) Analysis of Convergence Acceleration Techniques Used in Unstructured Algorithms in the Solution of Aeronautical Problems—Part I, Proceedings of the XVIII International Congress of Mechanical Engineering (XVIII COBEM), Ouro Preto, MG, Brazil. [available in CD-ROM].
12. Maciel ESG (2008) Analysis of Convergence Acceleration Techniques Used in Unstructured Algorithms in the Solution of Aerospace Problems—Part II, Proceedings of the XII Brazilian Congress of Thermal Engineering and Sciences (XII ENCIT), Belo Horizonte, MG, Brazil. [available in CD-ROM].
13. Anderson Jr. JD (1984) Fundamentals of Aerodynamics, McGraw-Hill, Inc., EUA, 563 p, 1984.

Chapter 17
A Parametric Non-Mixture Cure Survival Model with Censored Data

Noor Akma Ibrahim, Fauzia Taweab and Jayanthi Arasan

Abstract In some medical studies, there is often an interest in the number of patients who are not susceptible to the event of interest (recurrence of disease) and expected to be cured. This article investigates the cure rate estimation based on non-mixture cure model in the presence of left, right and interval censored data. The model proposed based on log-normal distribution that incorporates the effects of covariates on the cure probability. The maximum likelihood estimation (MLE) approach is employed to estimate the model parameters and a simulation study is provided for assessing the efficiency of the proposed estimation procedure under various conditions.

Keywords Censored data · Cure fraction · Interval · Lognormal distribution · MLE method · Non-mixture cure model

17.1 Introduction

Cure fraction models are survival models that account for the probability of a subject being cured. Recently, these models are broadly used for analysing data from cancer clinical trials and from other diseases. In the literature there are two major approaches to model survival data with cure fraction. The first one is the cure rate model, which is a mixture of two separate regression models for the survival function of uncured individuals and the cure fraction of the cured subjects. This model proposed by [1] with subsequent extensive investigations in the literature [2–6], among others.

N. A. Ibrahim (✉) · J. Arasan
Department of Mathematics, Universiti Putra Malaysia,
43400 Serdang, Selangor Darul Ehsan, Malaysia
e-mail: nakma@upm.edu.my

F. Taweab
Institute for Mathematical Research, Universiti Putra Malaysia,
43400 Serdang, Selangor Darul Ehsan, Malaysia

F. Taweab
Department of Statistics, University of Tripoli, PO Box 3601381 Tripoli, Libya
e-mail: taweabf@yahoo.com.my

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_17,

The second approach for modeling the cure fraction is known by the non-mixture cure model. This model proposed by [7] as an alternative to the cure rate model to keep the proportional hazards structure for the whole population, while allowing for a straight-forward interpretation of the covariate effects on the probability of cure [8–10]. The non-mixture cure model was motivated by the underlying biological mechanism and developed based on assumption that the treatment leaves the subject with a number of tumor cells that may grow slowly over time and produce a detectable cancer [7, 11, 12]. Both mixture and non-mixture cure models have received considerable attention based on right-censored data. However, the literature on the cure models with interval-censored data is not many and only a few recent studies have investigated the cure models with this type of data.

Interval censoring is frequently encountered in medical and biological researches. With this type of survival data the failure time cannot be directly observed, and it is only known to lie within an interval obtained from a sequence of examinations times. Interval-data have been extensively studied but without involving the cure fraction. The interested readers can refer, to some reviews articles by authors such as [13–15]. On interval-censoring with a cured subject, the published studies include [10, 16, 17].

In this present article, a parametric non-mixture cure model in the presence of interval, right and left censored data is considered. The estimation method is based on the maximum likelihood approach in which lognormal distribution is used to model failure time for the uncured subjects. The paper is organized as follows. In Sect. 17.2, the non-mixture cure model is described, and its performance under right censoring is evaluated via simulation studies in Sect. 17.2.1. Section 17.3 provides the MLE estimation for the model in the presence of interval, right and left censored data. Simulation studies are reported in Sect. 17.4. We conclude with a brief discussion in Sect. 17.5.

17.2 Model Specification

Chen et al. [7] defined a non-mixture formulation as follows. Let N denote the number of carcinogenic cells that remain active and capable of developing a cancer for the i^{th} subject. Assume that N has a Poisson distribution with a mean of θ. Let Z_j, $j = 1, 2, ..., N$ express the random time for the j^{th} cancer cell which can produce a detectable cancer mass where Z_j is assumed to be independently and identically distributed with $F(.)$. Then, the recurrence of cancer can be defined by the random variable T such that $T = \min\left\{Z_j,\ 0 \le j \le N\right\}$. The survival function for the population is given by:

$$\begin{aligned} S(t) &= P(\text{no cancer by time } t) \\ &= P(N=0) + P(Z_1 > t, ..., Z_N > t, N \geq 1) \\ &= e^{-\theta} + \sum_{N=1}^{\infty} \frac{\theta^N e^{-\theta}}{N!} \left[1 - F(t)\right]^N \\ &= \sum_{N=0}^{\infty} \frac{\theta^N e^{-\theta}}{N!} \left[1 - F(t)\right]^N \\ &= e^{-\theta F(t)} = p^{F(t)} \end{aligned} \tag{17.1}$$

where p is the probability of cure which can be defined as

$$p = S(\infty) = \lim_{t \to \infty} e^{-\theta F(t)} = e^{-\theta} \tag{17.2}$$

For the i^{th} individual, with $i = 1, 2, ..., n$, consider $y_i = \min(T_i, C_i)$, where C_i is a right censored variable. Let δ_i represents the censoring indicator which equals 1 if y_i is an actual failure time (uncensored) and 0 if it is right censored. Considering that censoring times are independent and non-informative, [6, 7, 20] show that the contribution of the i^{th} subject for the likelihood is given by

$$L_i = \prod_{i=1}^{n} [-\log(p) f(t_i)]^{\delta_i} S(t_i) \tag{17.3}$$

The model can be further extended by incorporating covariates X into the cure probability p and the survival function for uncured subjects. Moreover, a parametric model can be specified for the failure time. In this work, we consider that the cure probability linked into covariates through θ by setting $\theta = e^{X'\beta}$, where β are the regression coefficients, and that a log-normal distribution for modeling the event time of the uncured individuals. The density and cumulative distribution function for this distribution are defined as

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln t - \mu)^2}{2\sigma^2}\right]$$

$$F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

where Φ is the standard normal distribution function.

Then, the log likelihood associated with n observed data can be written as

$$L = \sum_{i=1}^{n} \delta_i \left[\log(-\log p) + \log f(t_i) \right] + (\log p) F(t_i)$$
$$= \sum_{i=1}^{n} \delta_i \left[\log(\exp(x'\beta)) - \log(t_i \sigma \sqrt{2\pi}) - \frac{(\ln t_i - \mu)^2}{2\sigma^2} \right] - \exp(x'\beta) \Phi\left(\frac{\ln t_i - \mu}{\sigma} \right) \tag{17.4}$$

The MLE of the parameters can be obtained by usual optimization methods such as the Newton-Raphson method.

17.2.1 Simulation Studies

Simulations studies using 1,000 samples each with $n = 100$, $n = 300$ and $n = 500$ were conducted for the model for both censored and uncensored individuals with one covariate. The covariate values were generated from a uniform distribution within $(-1, 0)$. The values of 0, 0.1, 0.3 and 0.1 were chosen as the parameters of β_0, β_1, μ and σ. Random numbers u_i were simulated from uniform distribution within (0, 1) to determine whether someone is cured. If subject is cured $(u \le p)$, then $T = \infty$. If the subject is not cured the failure time T was set to the solution of $u = \exp(-\exp(\beta_0 + \beta_1 x) F(t))$. The censoring times were generated from lognormal distribution (μ, σ), where the values of μ and σ would be adjusted to get the desired approximate censoring rate in the data.

Table (17.1) presents the bias, SE, and MSE of the parameters estimates at two different levels of censoring. The simulation results show that the biases of the estimators are very small. The SE and MSE values increase with the increase in the rate of censoring and decrease in the sample size, which indicates that lower level of censoring and large sample size make estimates more efficient and rather accurate.

17.3 The Model with Interval Censored Data

Under interval censoring mechanisms, the failure time T cannot be observed exactly, but instead is known to have occurred within an interval $(L_i, R_i]$ where $T_i \in (L_i, R_i]$, and $L_i \le R_i$. Here, L_i is the latest examination time before the event and R_i is the earliest examination time after the event. The i^{th} subject is left-censored if she/he has met the event of interest at unknown time prior to R_i; $T_i \in (0, R_i]$. The subject is right censored if she/he has been event-free at the last known time, $T_i \in (L_i, \infty)$. Left and right censored can be considered as special cases of interval censored data [18, 19]. For convenience of notation, let us introduce the left, interval and right censoring indicators as

Table 17.1 Bias, SE, and MSE of the MLE estimators for two censoring rate

		True	Bias	SE	MSE
Moderate censoring (35–40)					
$n = 100$	β_0	0	0.004	0.263	0.069
	β_1	0.1	0.002	0.469	0.221
	μ	0.3	$-3e^{-04}$	0.014	$1.69e^{-04}$
	σ	0.1	0.001	0.009	$8.20e^{-05}$
$n = 300$					
	β_0	0	−0.005	0.149	0.022
	β_1	0.1	−0.001	0.255	0.065
	μ	0.3	$2e^{-04}$	0.008	$6.40e^{-05}$
	σ	0.1	$-4e^{-04}$	0.005	$2.52e^{-05}$
$n = 500$					
	β_0	0	−0.002	0.116	0.013
	β_1	0.1	0.014	0.204	0.042
	μ	0.3	$3e^{-04}$	0.006	$3.61e^{-05}$
	σ	0.1	$2e^{-04}$	0.004	$1.60e^{-05}$
Heavy censoring (60–65)					
$n = 100$	β_0	0	−0.008	0.327	0.107
	β_1	0.1	0.006	0.588	0.346
	μ	0.3	$3e^{-04}$	0.019	$3.60e^{-04}$
	σ	0.1	0.001	0.012	$1.48e^{-04}$
$n = 300$					
	β_0	0	−0.008	0.183	0.034
	β_1	0.1	−0.022	0.328	0.108
	μ	0.3	$-2e^{-04}$	0.011	$1.21e^{-04}$
	σ	0.1	0.001	0.007	$4.96e^{-05}$
$n = 500$					
	β_0	0	−0.003	0.157	0.025
	β_1	0.1	$-1e^{-04}$	0.263	0.069
	μ	0.3	$2e^{-04}$	0.008	$6.40e^{-05}$
	σ	0.1	0.001	0.005	$2.55e^{-05}$

$\delta_{Li} = 1$ If subject is left censored, 0 otherwise;
$\delta_{Ii} = 1$ If subject is interval censored, 0 otherwise;
$\delta_{Ri} = 1$ If subject is right censored, 0 otherwise,

Note that $\delta_{Ri} = 1 - (\delta_{Li} + \delta_{Ii})$. Then, the likelihood function for the n observed interval data will be

$$
\begin{aligned}
L_c &= \prod_{i=1}^{n} [1 - \exp(-e^{X'\beta} F(R_i^+))]^{\delta_{Li}} \times [\exp(-e^{X'\beta} F(L_i^-)) \\
&\quad - \exp(-e^{X'\beta} F(R_i^+))]^{\delta_{Ii}} \times [\exp(-e^{X'\beta} F(L_i^-))]^{\delta_{Ri}} \\
&= \prod_{i=1}^{n} \left[1 - \exp\left(-e^{X'\beta} \Phi\left(\frac{\ln R_i - \mu}{\sigma} \right) \right) \right]^{\delta_{Li}} \\
&\quad \times \left[\exp\left(-e^{X'\beta} \Phi\left(\frac{\ln L_i - \mu}{\sigma} \right) \right) - \exp\left(-e^{X'\beta} \Phi\left(\frac{\ln R_i - \mu}{\sigma} \right) \right) \right]^{\delta_{Ii}} \times \left[\exp\left(-e^{X'\beta} \Phi\left(\frac{\ln L_i - \mu}{\sigma} \right) \right) \right]^{\delta_{Ri}}
\end{aligned}
\tag{17.5}
$$

The maximum likelihood estimation of the parameters can be obtained by using the Newton-Raphson iterative procedure.

17.4 Simulation Studies

Simulation studies were conducted using 1,000 runs each with sample sizes of 100, 300, and 500 for this model. One covariate X was considered and generated from a uniform distribution in $(-1, 0)$. The random survival times T were generated under the cure model (1). Thus, a uniform (0,1) random variable u was generated and the subject is cured if $u \leq p$. Otherwise, the failure times T were set to the solution of $u = \exp(-e^{\beta_0+\beta_1 x} F(t))$. The true value of μ and σ were chosen to be 0.3 and 0.1 respectively. The value of β_0 was chosen to be 0 or 1, while β_1 as 0.1 and 0.5, which results in a cure rate of about 33 % and 0.06 %. On average, about 30 % observations were left censored, 40 % were interval censored and 30 % were right censored.

The visiting or examination times were simulated independent of X and T following Goulin [21], assuming that the number of visiting times is 10 visits for each subject, and that the time between two visits has a uniform distribution within $(0, c)$, where c is a constant controlling censoring rate.

In each simulation, we assessed the bias, standard error (SE), and mean square error (MSE) of the estimates and the results are collectively presented in Tables 17.2 and 17.3.

The simulation studies suggest that the proposed method has very small biases. The SE decreased with increasing sample sizes for all considered parameters. Given the consistency of the estimator and the increased precision with increasing samples size, the mean square errors (MSE) also decreased with increasing sample size.

17.5 Conclusions

In the analysis of lifetime data, usually we could have a fraction of the population not exposed to the event of interest, especially in medical fields. Non-mixture cure model is considered as a major approach to handle this kind of data. In this paper, a parametric non-mixture cure model is proposed for interval censored data in the presence of covariates. The maximum likelihood estimates (MLE) method is used to estimate the parameters. For various sample sizes, we implemented a simulation process to generate samples with cure fraction, and then under this setup the MLE's for the model were obtained. The values of the bias and the MSE that were obtained from simulation studies show that the proposed estimation method performs well in the situations considered.

Table 17.2 Bias, SE, and MSE of the MLE estimators

	True	Bias	SE	MSE
$n = 100$				
β_0	0	−0.013	0.356	0.127
β_1	0.1	0.064	0.314	0.103
μ	0.3	−0.005	0.203	0.041
σ	0.1	−0.032	0.178	0.033
$n = 300$				
β_0	0	−0.009	0.335	0.112
β_1	0.1	0.031	0.275	0.077
μ	0.3	0.019	0.163	0.027
σ	0.1	−0.008	0.087	0.008
$n = 500$				
β_0	0	−0.006	0.305	0.093
β_1	0.1	0.003	0.266	0.071
μ	0.3	0.002	0.143	0.020
σ	0.1	0.009	0.083	0.007

SE mean of standard errors, *MSE* mean square errors for MLE estimators

Table 17.3 Bias, SE, and MSE of the MLE estimators

	True	Bias	SE	MSE
$n = 100$				
β_0	1	−0.024	0.163	0.027
β_1	0.5	0.036	0.220	0.050
μ	0.3	0.016	0.374	0.140
σ	0.1	0.027	0.169	0.029
$n = 300$				
β_0	1	−0.008	0.105	0.011
β_1	0.5	0.012	0.143	0.021
μ	0.3	0.003	0.249	0.062
σ	0.1	−0.024	0.118	0.015
$n = 500$				
β_0	1	0.002	0.085	0.007
β_1	0.5	0.006	0.126	0.016
μ	0.3	0.005	0.175	0.031
σ	0.1	0.016	0.083	0.007

SE mean of standard errors, *MSE* mean square errors for MLE estimators

Acknowledgement The authors are much thankful and grateful to the Institute for Mathematical Research, Universiti Putra Malaysia (UPM), for their generous support of this study.

References

1. Berkson J, Gage R (1952) Survival curve for cancer patients following treatment. J Amer Statist Assoc 47:501–515
2. Taylor JMG (1995) Semi-parametric estimation in failure time mixture models. Biometrics 51:899–907
3. Maller RA, Zhou X (1996) Survival Analysis with Long-Term Survivors, Chichester. John Wiley and Sons
4. Sy JP, Taylor JMG(2000) Estimation in a cox proportional hazards cure model. Biometrics 56:227–236
5. Achcar A, Jorge, Coelho- Barros Emi'lio, A, Josmar Mazuchell (2012) T Cure fraction models using mixture and non-mixture models. Tatra Mt Math Publ 51:1–9
6. Mizoi MF, Bolfarine H, Lima ACP (2007) Cure rate models with measurement errors. Communications in Statistics—Simulation and Computation 36:185–196
7. Chen MH, Ibrahim JG, Sinha DA (1999) A new bayesian model for survival data with a surviving fraction. J Amer Statist Assoc 94:909–9198
8. Tsodikov AD, Ibrahim JG, Yakovlev AY (2003) Estimating cure rates from survival data: an alternative to two-component mixture models. J Amer Statist Assoc 98:1063–1078
9. Banerjee S, Carlin BP J (2004) The Parametric spatial cure rate models for interval-censored time-to-relapse data. Biometrics 60:268–275
10. Liu Hao, Shen (2009) A semi parametric regression cure model for interval censored data. J Amer Statist Assoc 487:1168–1178
11. Gutierrez RG (2002) Parametric frailty and shared frailty survival models. Stata 2:22–44
12. Zeng D, Yin G, Ibrahim JG (2006) Semi parametric transformation models for survival data with a cure fraction. J Amer Statist Assoc 101:670–684
13. Lindsey JC, Ryan LM (1998) Tutorial in biostatistics methods for interval censored data. Stat Med 17:219–238
14. Gomez G, Calle ML, Oller R (2004) Frequentist and bayesian approaches for interval-censored data. Statist Pap 45:139–173
15. Gomez G, Calle ML, Oller R, Langohr K (2009) Tutorial on methods for interval-censored data and their implementation in R. Statist Model 9:259–297
16. Lam KF, Xue H (2005) A semiparametric regression cure model with current status data. Biometrika 92:573–586
17. Kim Y, Jhun M (2008) Cure rate model with interval censored data. Stat Med 27:3–14
18. Sun J (2006) The statistical analysis of interval censored failure time data. Springer, New York
19. Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, Wiley, New York
20. Claire L Weston, John R Thompson (2010) Modeling survival in childhood cancer studies using two- stage non-mixture cure models. Journal of Applied Statistics 37:1523–1535
21. Zhao Guolin MA (2008) Nonparametric and Parametric Survival Analysis of Censored Data with Possible Violation of Method Assumptions. Ph.D. Diss, North Carolina University

Chapter 18
Information Technology Model for Supporting Open Utility Market

E. Grabovica, Dz. Borovina and S. Kovacevic

Abstract This article gives a brief description of a role that an Information technology has in one power system company that operates in an open utility market. When we consider open energy market, it is well known that Utility sector in European Union has already been liberalized and all the steps in joining that market are defined in legislatives. When it comes to IT role in such a company, IT plays a significant role in adopting to a new market. This work describes some of IT segments critical for achieving those goals, that have already been or will be implemented in public company for producing, distributing and supplying electrical energy in Bosnia and Herzegovina. As mentioned, some of IT projects have already been implemented in order to prepare the company for the upcoming open market. However, some information systems still need to be implemented, as a crucial for a competitive position of a company in a new market.

Keywords Customer information system · Billing · Customer relationship management · Disaster recovery

18.1 Introduction

All business areas in utility sectors, such as producing, transmitting, distributing and retail of electrical energy traditionally have been defined as monopolistic areas which have no competition. In that manner, all business processes were defined and managed by unified, centralized and vertically integrated subject which has a monopol on regional market of electrical energy. However, the only segment that actually should be protected from the competitors is network operator. All the other

E. Grabovica (✉) · D. Borovina · S. Kovacevic
Public Company for producing, distributing and retail of electrical energy,
Vilsonovo setaliste 15, Sarajevo, Bosnia and Herzegovina
e-mail: e.grabovica@elektroprivreda.ba

Dz. Borovina
e-mail: dz.borovina@elektroprivreda.ba

S. Kovacevic
e-mail: se.kovacevic@elektroprivreda.ba

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_18,

segments such as producing and supplying of electrical energy can and should be considered as potential concurrent services.

Inside the European Union, utility sector has been liberalized and defined in EU law regulations (Directions 96/92/EC and 2003/54/EC), in order to make the energy market open for the competitors. According to these directives, all states that are members of EU must establish full market competition, as well as the possibility for customers to choose their supplier. Deadline for these obligations were July 2007. In Bosnia and Herzegovina, all customers except households are in a position to choose supplier of electrical energy. Households will have the same opportunity from the January 2015. In order to establish successful competition in power supply process, it will be necessary to change the role and the position of Distributed System Operator. In EU states, Operators that have more than 100,000 customers using their network must implement all mechanisms of separations, while operators with less customers have to implement separation of accounting data and information [1].

18.2 Problem Formulation

Business and technical information systems must be fast, flexible, scalable and secure in order to provide all needed support for business in a deregulated market of electrical energy. Sluggishness and slow changes will not be tolerated as it used to be in a privileged position of a company. Information systems must be able to transform themselves and to adjust to changes in business and organizational aspect, as well as in terms of providing new services to customers.

On the customer service side, CIS should support multiple client interaction channels—such as call center, interactive voice response/voice response units (IVR/VRUs) and SMS—as well as customer self-service needs [2].

In a competitive market, a CIS also needs to enable data exchanges with other market participants (such as metering service companies, network companies, competitive retailers/suppliers and market operators).

The CIS product requirements defined by electric utilities tend to be more complex because of the intricate nature of the business. Issues such as the inability to store and manage commodity (electricity), more complex market structures (such as retail competition and unbundling), smart grids and the deployment of advanced metering infrastructure (AMI) tend to keep the electric utility at the forefront of business innovation in the CIS market, compared with other utility sectors [2].

Key elements of Customer Information system in supply company that operates on open market support are:

- Billing system
- Customer Relationship Management (CRM)
- Those systems will directly be in charge for customer services support. Besides, other information systems that will play important role in supporting open market changes are:

- Meter Data Management (MDM)
- Financial Management Information System (SAP)
- Distributed information system for new customers
- Database of electrical objects (DEEO).

18.3 Billing

When considering Billing system in terms of open market, it is necessary to point out that electrical energy supplier that works in open market must be competitive. It has to be fast when serving customers, at the same time providing new services and solving all problems customer has while using the service. Supplier must work proactively having in mind the competition. Billing application must provide the following functions when serving customer needs in open market [3]:

- Signing, analysis and monitoring contracts with qualified customers
- Signing, analysis and monitoring contracts with producers/dealers
- Collecting/archiving accounting data for qualified customers
- Accounting and bill delivery to qualified customers
- Charging, financial monitoring of debits and interests
- Customer data analysis, financial cards, reclamations, complaints.

Data exchange with other interested acters in the open market, in order to provide support for new processes, such as: change of supplier and customer moving. Figure 18.1 shows data exchange workflow between CRM, MDM and Billing modules, including new services.

New Customer Information system should support many advanced and sophisticated applications and modules that are related to open market, such as:

- Contracts
- Marketing strategy
- Accounting and billing
- Selling and services
- Advanced Metering Management
- Energy consumption metering
- Customer accounts and finance metering.

18.4 Customer Relationship Management

In terms of Customer Relationship management, communication with customers should be analyzed and implemented in a closely coordination with supply process and its requirements. In a Public company for supplying electrical energy, CRM solution that was recently implemented is Oracle Siebel Energy & Media.

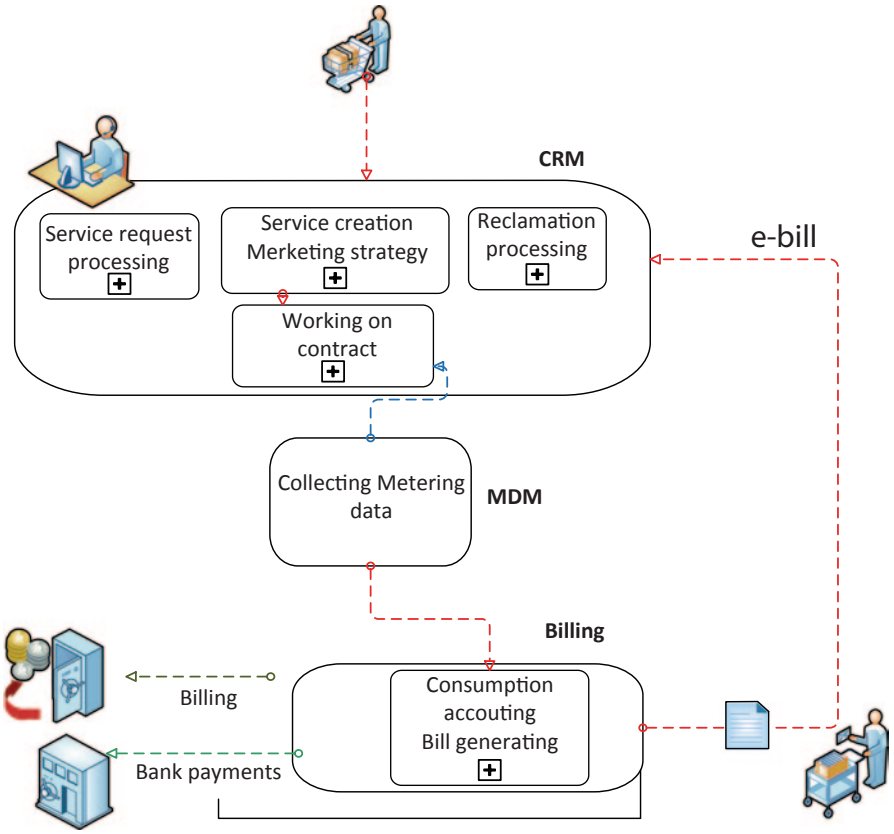


Fig. 18.1 Customer Information system diagram in a Utility company [3]

For companies that compete for customers in open market, CRM is extremely important segment of business. It should provide high available access to the service, reliable and accurate data, as well as professional and kind staff that serves customer needs.

According to Gartner researches, some of the strongest new areas of CRM focus, that should be implemented in supply company that operates on open market are [4]:

- Cross-channel CRM customer engagement applications, including customer-controlled communication
- Social networking systems that improve customer service through input taken directly from customers
- Video customer service and delivery systems, especially for the support of mobile consumers
- Mobile-based, location and context-sensitive technologies
- Customer service analytics, including big data

- BPM tools that enable the entire migration of increasingly complex customer service tasks and interactions for Web customer service
- Applications that optimize customer service agent interactions through advances in skills management, knowledgebase, search and real-time decision support
- Technologies that support rapid iterations and improvements in business processes
- Analytical tools that predict the most likely intent of customers' requests for service, as well as emerging needs for services and the optimization of each interaction to cross-sell or upsell products and services.

All those functions and tools should provide a competitive advantage for the company.

Some of new functions that a modern and innovative CRM system in our company will have to provide at the start point are [5]:

- Multi-channel outgoing campaigns (call, SMS, e-mail):
  - Planned shutdowns of energy objects notifications
  - Warning for customers who do not pay bills
- E-bill service: Paying bills using web services
- SMS services:
  - Notification of accounting data
  - Notification of planned shutdowns
  - Notification of planned reconstructions.

For any of those services, it is necessary to define marketing campaign for promoting services, in which customers are enforced to choose a specific service they are interested in, as well as to provide contact information necessary for using the service.

Methods for implementing previously defined services are:

- Notification of planned shutdowns by sms
  - CRM filters all the customers that will be covered by the planned shutdown
  - We define the type of the campaign (call/email/sms)
  - CRM filters customers that have contact data such as phone number, email address
  - CRM sends generic sms message such as "Your area will be out the electricity for a period…, thank you for the understanding…"
- Warning for customers who do not pay bills
  - Billing system send the list of customers who did not pay the bill
  - We define the type of the campaign (call/email/sms)
  - CRM filters that have contact data such as phone number, email address
  - CRM sends generic message such as "We kindly ask you to pay the bill for the electricity, otherwise your power supply will be shutdown…"
- Outgoing phone call campaigns for planned shutdowns
  - CRM will have one daily task that will be in charge for generating this list of customers that will be covered by planned shutdown in the period of 3 days

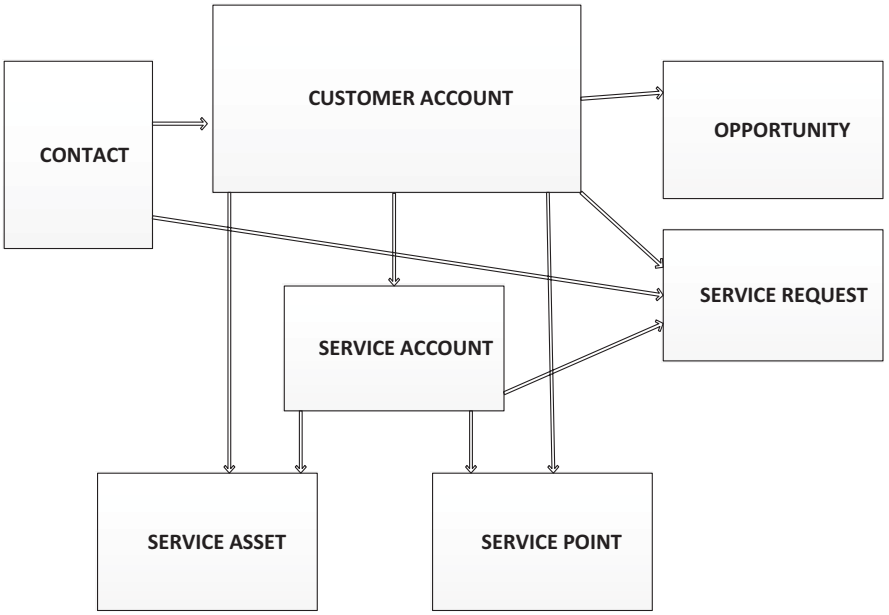


Fig. 18.2 CRM Data exchange model [5]

- Detailed information regarding planned shutdowns CRM will be receiving from the Dispathing application that is primarly responsible for managing those data.
- Generated list of customers that will be in a planned shutdown will be imported to Contact center and CRM, together with data such as: phone number, date and time of shutdown.
- Contact center will automatically make phone calls to the customers from the list and present predefined message "Your area will be out the electricity for a period…, thank you for the understanding…"

Figure 18.2 shows CRM data model that CRM uses for exchanging data with other information systems in a company.

Key entities of CRM data exchange model should be:

- Customer Account
- Service Account
- Service Point
- Service Request
- Contact.

CRM data model is based on existing relations between Customer Account and Service account data, as well as between Service account and Service point data. Customer account entity keeps basic information about customers before they start

Table 18.1 Availability parameters in CRM [5]

Service availability	☐ Monday–Friday		☒ Each day		
Availability hours during working week	☐ 7–17 h		☒ 0–24 h		
Availability hours during working weekends and holidays	☐ 7–17 h		☒ 0–24 h		
Regular maintenance	Every second Wednesday in a month				
Time of regular maintenance	7–7:45 h				
Required availability weekly	☐ 75%	☐ 90%	☐ 95%	☒ 99%	☐ 99.9%
Required availability monthly	☐ 75%	☐ 90%	☐ 95%	☒ 99%	☐ 99.9%
Reaction time	☐ 5 min	☒ 15 min		☐ 30 min	
Correction time	☐ 15 min	☐ 30 min	☒ 60 min		☐ 150 min

using electrical energy, while Service Account entity keeps information about customers who are already connected to electrical power network.

The following table gives availability parameters in a company (Table 18.1).

18.5 Conclusions

Full benefit of implementing new IT model will be completely achieved after the market deregulation. However, results that have been achieved until now are:

- Central database repository and management of customer data, as well as exchanging relevant customer data sets with corresponding databases.
- Full visibility and monitoring of database traffic in the system.
- Centralized management of security and event information in the system.

The process of market deregulation acompained with the market competition requires one completely new approach and crucial changes in organization and management at the same time. IT role in that process is very important for the following reasons:

- IT should provide technical, logistic and administrative support for improving existing and implementing new customer services at deregulated market (outgoing campaigns, e-services, m-services,…)

- IT should help in achieving better competitive position on the market
- From the position of support process IT should evolve to strategic position which will allow IT to participate in defining strategic goals and new IT services that will be provided to customers.

The main goal of IT is to balance user needs and business priorities while maintaining control.

Business process reengineering, adoption to legislative, as well as the internal reorganization of the company will not be possible without IT support. IT will eventually be involved in all aspects of planning and defining business goals, choosing appropriate IT solution, up to the final implementation.

References

1. Dr. Daniel Grote and Peter Fisher, Role of distributed system operator as a neutral promotor of open energy market, USAID, 2013.
2. Z. Sumic.(June 2013). Magic Quadrant for Utilities Customer Information Systems. [Online]. Available: www.gartner.com.
3. Study on implementing new Billing system, Public Company for producing and distributing electrical energy in Bosnia and Herzegovina, 2013, pp. 9–25.
4. M. Maoz. (July 2013). Hypo cycle for CRM Customer Service and support. [Online]. Available:www.gartner.com.
5. Technical documentation on implementing new centralized Contact center and CRM, Public Company for producing and distributing electrical energy in Bosnia and Herzegovina, 2013, pp. 52–55, 64, 73.

Chapter 19
Thermodynamic Properties of Engine Exhaust Gas for Different Kind of Fuels

H. K. Kayadelen and Y. Ust

Abstract Engine performance highly depends on the thermodynamic properties of the working fluid involved in different processes of engine cycles. However, as engines run with broad range of fuels and fuel air/ratios it is usually not possible to reach thermodynamic tables of products of combustion in the exhaust gas content for a specific fuel/air ratio and at a relevant temperature and pressure. This study addresses a MATLAB adaptable code to investigate engine performance according to the specifically defined thermodynamic properties of the exhaust mixture. Specific heat, enthalpy, entropy and isentropic exponent values are calculated precisely according to the chemical equilibrium approach, assuming there are 10 products of combustion in the exhaust gas content. Graphical illustrations are considered to be reference for future engine performance studies which are derived for different fuels and fuel air ratios under various temperatures and pressures.

Keywords Exhaust gas · Internal combustion engines · Thermodynamic properties · Engine performance

19.1 Introduction

According to Heywood [1]; Rashidi [2] and Rakopoulos et al. [3], it is a good approximation for performance estimates in engines to regard the burned gases produced by the combustion of fuel and air as in chemical equilibrium and therefore knowledge of the exact gas composition inside the combustion chamber is critical to the accurate calculation of the thermodynamic cycle models of internal combustion engines.

Numerous authors studied on prediction of emissions for particular engine parameters. Some of those concerning reciprocating internal combustion engines are

H. K. Kayadelen (✉)
Department of Marine Engineering Operations,
Yildiz Technical University, 34349 Istanbul, Turkey
e-mail: hasankayhakayadelen@hotmail.com

Y. Ust
Department of Naval Architecture and Marine Engineering,
Yildiz Technical University, 34349 Istanbul, Turkey
e-mail: yust@yildiz.edu.tr

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering*,
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_19,

given in [4–9]. Apart from those some research on emission prediction from gas turbine engines are given in [10–14].

As these studies usually concentrate on some particular emissions they are not able to predict the full equilibrium scheme. Additionally, existing engine combustion studies usually make an assumption of complete combustion of a $C_\alpha H_\beta$ fuel with excess air, so they treat the exhaust stream as a mixture of complete combustion products only comprising of CO_2, H_2O and N_2. This approach may lack in precision although there is sufficient oxygen which can completely oxidize all the fuel because of the dissociations of combustion products at high temperatures. For example, if the temperature of a mass of carbon dioxide gas in a vessel is increased sufficiently, some of the CO_2 molecules dissociate into CO and O_2 molecules. If the mixture of CO_2, CO and O_2 is in equilibrium, this means CO_2 molecules are dissociating into CO and O_2 at the same rate as CO and O_2 molecules are recombining in the proportions required to satisfy the equation $CO + ½\ O_2 = CO_2$. When hydrocarbon fuels are subjected to combustion at low temperatures, for the rich case the major product species present are N_2, H_2O, CO_2, CO and H_2 where for the lean case N_2, H_2O, CO_2, and O_2. But at higher temperatures (greater than about 2,200 K), these major species dissociate and react to form additional species in significant amounts [1]. So actual combustion reactions do not go to completion and it will be useful to develop an equilibrium product composition [15] by which individual species in the burned gases react together, produce and remove each species at equal rates but no net change in species composition results [1].

In this study the thermodynamic change of the working fluid will be calculated for 10 main products as suggested by Ferguson [16] to get more accurate values of gas properties which may lead considerable changes in performance results in comparison to those made by an assumption of complete combustion. Reaction can be easily remodelled if more reactant or product species is required and it is possible to estimate the new properties of the working fluid to be used in the performance estimation of any combustion engine.

In order to obtain the equilibrium compositions and thermodynamic properties, the chemical equilibrium routines of Olikara and Borman [17] presented by Ferguson [16] based on equilibrium constant approach are used in this present work.

19.2 Problem Formulation

In the analytical model presented for the adiabatic combustion in this section it is assumed that all gases are ideal gases and their enthalpies only change with temperature. Considering that for $\phi < 3$ and there are 10 constituents, the high temperature combustion model is given.

The air supplied for the combustion is assumed to be completely dry without any moisture and containing only 0.21 moles of O_2 and 0.79 moles of N_2. Standard reference temperature and pressure are 25 °C and 1 atm. respectively.

Table 19.1 Chemical relations of equilibrium

$1/2\,H_2 \rightleftarrows H$	$K_1 = y_7\sqrt{p}/\sqrt{y_6}$	$c_1 = K_1/\sqrt{p}$	$y_7 = c_1\sqrt{y_6}$
$1/2\,O_2 \rightleftarrows O$	$K_2 = y_8\sqrt{p}/\sqrt{y_4}$	$c_2 = K_2/\sqrt{p}$	$y_8 = c_2\sqrt{y_4}$
$1/2\,H_2 + 1/2\,O_2 \rightleftarrows OH$	$K_3 = y_9/\left(\sqrt{y_4}\sqrt{y_6}\right)$	$c_3 = K_3$	$y_9 = c_3\sqrt{y_4}\sqrt{y_6}$
$1/2\,O_2 + 1/2\,N_2 \rightleftarrows NO$	$K_4 = y_{10}/\left(\sqrt{y_4}\sqrt{y_3}\right)$	$c_4 = K_4$	$y_{10} = c_4\sqrt{y_4}\sqrt{y_3}$
$H_2 + 1/2\,O_2 \rightleftarrows H_2O$	$K_5 = y_2/\left(y_6\sqrt{y_4}\sqrt{p}\right)$	$c_5 = K_5\sqrt{p}$	$y_2 = c_5 y_6\sqrt{y_4}$
$CO + 1/2\,O_2 \rightleftarrows CO_2$	$K_6 = y_1/\left(y_5\sqrt{y_4}\sqrt{p}\right)$	$c_6 = K_6\sqrt{p}$	$y_1 = c_6 y_5\sqrt{y_4}$

The chemical equation for the combustion model is given below:

$$\varepsilon\phi C_\alpha H_\beta O_\gamma N_\delta + \left(0.21O_2 + 0.79N_2\right) + \rightarrow \nu_1 CO_2 + \nu_2 H_2O + \nu_3 N_2 + \nu_4 O_2 + \nu_5 CO + \nu_6 H_2 + \nu_7 H + \nu_8 O + \nu_9 OH + \nu_{10} NO \tag{19.1}$$

Here ν_1 to ν_{10} represents the number of moles for each species, α, β, γ, δ are the numbers of carbon, hydrogen, oxygen and nitrogen atoms present in the fuel. ϕ is equivalence ratio and ε is the molar air-fuel ratio obtained from the stoichiometric combustion of the fuel which are calculated as below:

$$\phi = \frac{FA}{FA_s} \tag{19.2}$$

$$\varepsilon = \frac{0.21}{\alpha + \frac{\beta}{4} - \frac{\gamma}{2}} \tag{19.3}$$

In order to solve for the 10 unknown mole numbers of Eq. (19.3), 10 equations are needed which six of them can be provided by the criteria of equilibrium among the products expressed by the following chemical relations in Table 19.1 [16]:

where the unit pressure p is in atmospheres and K_1–K_6 are the equilibrium constants (based on partial pressures) of each reactions.

There are four more equations which come from the combustion model atom balancing:

$$\mathbf{C} \quad \varepsilon\phi\alpha = \left(y_1 + y_5\right)N \tag{19.4}$$

$$\mathbf{H} \quad \varepsilon\phi\beta = \left(2y_2 + 2y_6 + y_7 + y_9\right)N \tag{19.5}$$

$$\mathbf{O} \quad \varepsilon\phi\gamma + 2\cdot 0.21 = \left(2y_1 + y_2 + 2y_4 + y_5 + y_8 + y_9 + y_{10}\right)N \tag{19.6}$$

$$\mathbf{N} \quad \varepsilon\phi\delta + 2 \cdot 0.79 = \left(2y_3 + y_{10}\right)N \tag{19.7}$$

where y_i stands for the mole fractions of 10 species and N is the total number of moles of the species:

$$y_i = \frac{\upsilon_i}{\sum_{i=1}^{10} \upsilon_i} \tag{19.8}$$

$$N = \sum_{i=1}^{10} \upsilon_i \tag{19.9}$$

With the added N, total number of unknowns are now 11 and but we have 10 equations so far. From the definition of mole fraction one can write the following equation which makes the total number of unknowns and total number of equations equal.

$$\sum_{i=1}^{10} y_i - 1 = 0 \tag{19.10}$$

We obtain a series of equations non-linear equations which can be solved by Newton–Raphson iteration to obtain the mole fractions at the equilibrium [16] and can be written as follows: Details of the solution procedure can be found in [17, 18].

$$f_i = (y_i) = 0 \tag{19.11}$$

At constant pressure, temperature variation is effective on specific heat because of the dissociations of species at high temperatures. The effect of temperature on mole fractions should be accounted for the equilibrium specific heat calculation differentiating Eq. (19.11) with respect to temperature which can be written as:

$$\frac{\partial f_j}{\partial T} + \frac{\partial f_j}{\partial y_i}\frac{\partial y_i}{\partial T} = 0 \tag{19.12}$$

The solution matrix from the above equation is used in finding the specific heat of the combustion products in Eq. (19.18).

Molar specific heat, enthalpy and entropy values of each species can be obtained from following expressions by using curve fit coefficients ($a_1 \ldots a_n$) for thermodynamic properties of (C–H–O–N) systems [19]:

$$\frac{\overline{h}_i^{o}}{R_u T} = a_{1,i} + \frac{a_{2,i}}{2}T + \frac{a_{3,i}}{3}T^2 + \frac{a_{4,i}}{4}T^3 + \frac{a_{5,i}}{5}T^4 + \frac{a_{6,i}}{T} \tag{19.13}$$

$$\frac{\overline{c}_{p,i}}{R_u} = a_{1,i} + a_{2,i}T + a_{3,i}T^2 + a_{4,i}T^3 + a_{5,i}T^4 \tag{19.14}$$

$$\frac{\overline{s}_i^{o}}{R_u} = a_{1,i}\ln T + a_{2,i}T + \frac{a_{3,i}}{2}T^2 + \frac{a_{4,i}}{3}T^3 + \frac{a_{5,i}}{4}T^4 + a_{7,i} \tag{19.15}$$

At constant pressure, enthalpy of the mixture change due to the dissociations as the mole fractions of the mixture change with temperature. This will change the ultimate specific heat of the gas mixture defined as follows:

$$h = \sum_{i=1}^{10} y_i \overline{h}_i^{o} \text{ [kJ/kmol]}. \tag{19.16}$$

$$h = \frac{1}{M}\sum_{i=1}^{10} y_i \overline{h}_i^{o} \text{ [kJ/kg]}. \tag{19.17}$$

$$\left(\frac{\partial h}{\partial T}\right)_P = c_{p_g} = \sum_{i=1}^{10} \frac{y_i}{M}\frac{\partial \overline{h}_i^{o}}{\partial T} + \frac{\overline{h}_i^{o}}{M}\frac{\partial y_i}{\partial T} - \frac{y_i \overline{h}_i^{o}}{M^2}\frac{\partial M}{\partial T} \tag{19.18}$$

Using Eq. (19.13–19.17), rearranging Eq. (19.18) gives:

$$\left(\frac{\partial h}{\partial T}\right)_P = c_{p_g} = \frac{1}{M}\left[\sum_{i=1}^{10} y_i \overline{c}_{p_i} + \frac{\partial y_i}{\partial T}\overline{h}_i^{o} - \frac{M_T}{M} y_i \overline{h}_i^{o}\right] \text{[kJ/kg K]}. \tag{19.19}$$

Where

$$M_T = \frac{\partial M}{\partial T} = \sum_{i=1}^{10} M_i \frac{\partial Y_i}{\partial T} \tag{19.20}$$

T is the combustion temperature in Kelvin at which the mole fractions of each equilibrium species, y_i are produced, M_i is the molecular weight of species i, and M is the the molecular weight of the mixture as follows:

$$M = \sum_{i=1}^{10} m_i = \sum_{i=1}^{10} y_i M_i \tag{19.21}$$

From the law of conservation of mass, the mass of the products is equal to the mass of reactants (m_R). A definition can be made as follows:

$$m_R = m_a + m_f \tag{19.22}$$

The total number of moles of the products can be found by dividing the mass of reactants into the molecular weight of the combustion products as follows:

$$N = \frac{m_R}{M} \tag{19.23}$$

Lastly, the number of moles $\nu_1, \nu_2 \ldots \nu_{10}$ are obtained from:

$$\upsilon_i = y_i \, N \tag{19.24}$$

19.3 Results and Discussion

Figures 19.1, 19.2, 19.3, 19.4 and 19.5 are for the thermodynamic properties for constant unburned gas temperature and Figs. 19.5, 19.6, 19.7, 19.8, 19.9 and 19.10 are the thermodynamic properties for constant pressure. For the constant unburned mixture temperature case the unburned temperature is assumed to be 300 K before combustion. For the constant pressure case the pressure is assumed to be 1 bar during combustion. The dedicated temperatures for different equivalence ratios are the adiabatic flame temperatures for each case which are presented in Figs. 19.5 and 19.10.

Figure 19.1 shows the change of constant pressure specific heat vs. pressure for different fuels and equivalence ratios. It can be seen that cp changes with pressure significantly for stoichiometric mixture. For other equivalence ratios the change is minor especially after pressure of 10 bar. cp increases with equivalence ratio for the lean case. Increasing equivalence ratio after its stoichiometric value decreases cp.

Figure 19.2 shows the change of specific entropy vs. pressure for different fuels and equivalence ratios. It can be seen that s decrease with pressure. Highest s is for methane for the equivalence ratio 1.2. In general s increase with equivalence ratio for all fuels. Highest entropy values are for fuel methane and lowest are for diesel.

Figure 19.3 shows the change of specific enthalpy vs. pressure for different fuels and equivalence ratios. No net change in h with pressure is observed. Lowest enthalpy values are for fuel methane and highest are for diesel.

Figure 19.4 shows the change of ratio of specific heats vs. pressure for different fuels and equivalence ratios. I It can be seen that k changes with pressure significantly for stoichiometric mixture. For other equivalence ratios the change is minor especially after pressure of 10 bar. k decreases with equivalence ratio for the lean case. Decreasing equivalence ratio after its stoichiometric value increases k for each fuel. Highest k values are for fuel methane and lowest are for diesel.

Figure 19.5 shows the change of burned gas temperature vs. pressure for different fuels and equivalence ratios. The temperature values are the temperature values at which Figs. 19.1–19.4 are obtained. The change of burned gas temperature with pressure is only significant at stoichiometric combustion at pressure less than 10 bar.

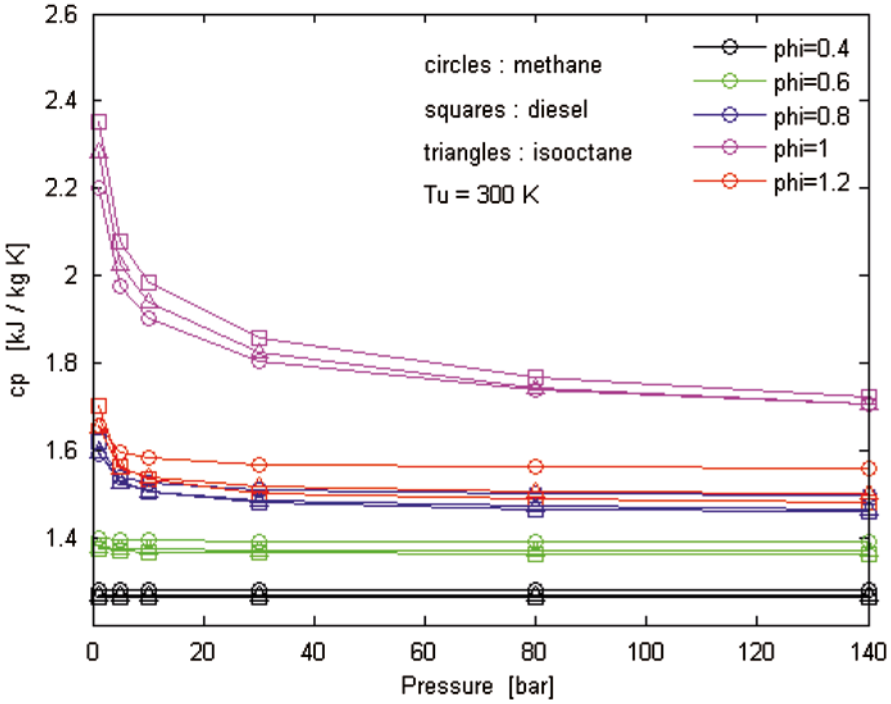


Fig. 19.1 Change of constant pressure specific heat vs. pressure

Figure 19.6 shows the change of constant pressure specific heat vs. unburned temperature for different fuels and equivalence ratios. Increasing unburned temperature increases the dedicated adiabatic flame temperature and accordingly the constant pressure specific heat. cp increases with equivalence ratio for the lean case. Increasing equivalence ratio after its stoichiometric value decreases cp. For this case, highest cp values are for fuel diesel and lowest are for methane.

Figure 19.7 shows the change of specific entropy vs. unburned temperature for different fuels and equivalence ratios. It can be seen that s increase with unburned temperature. Highest s is for methane for the equivalence ratio 1.2. In general s increase with equivalence ratio for all fuels. Highest entropy values are for fuel methane and lowest are for diesel.

Figure 19.8 shows the change of specific enthalpy vs. unburned temperature for different fuels and equivalence ratios. It can be seen that h increase with unburned temperature. Lowest enthalpy values are for fuel methane and highest are for diesel.

Figure 19.9 shows the change of ratio of specific heats vs. pressure for different fuels and equivalence ratios. It can be seen that k decrease with increasing unburned mixture. Highest k values are for fuel methane and lowest are for diesel.

Figure 19.10 shows the change of burned gas temperature vs. unburned temperature for different fuels and equivalence ratios. The burned gas temperature always

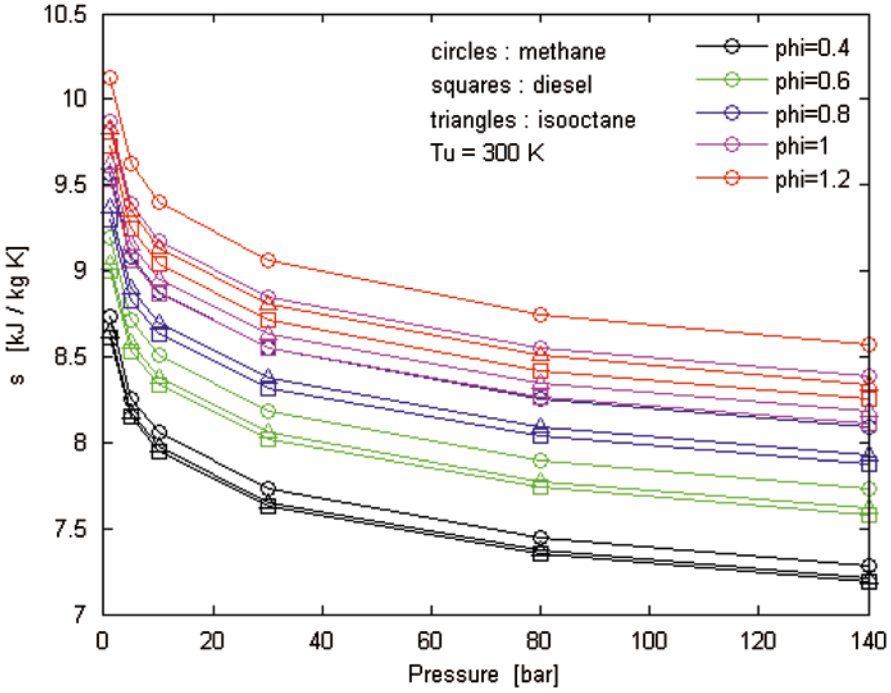


Fig. 19.2 Change of specific entropy vs. pressure

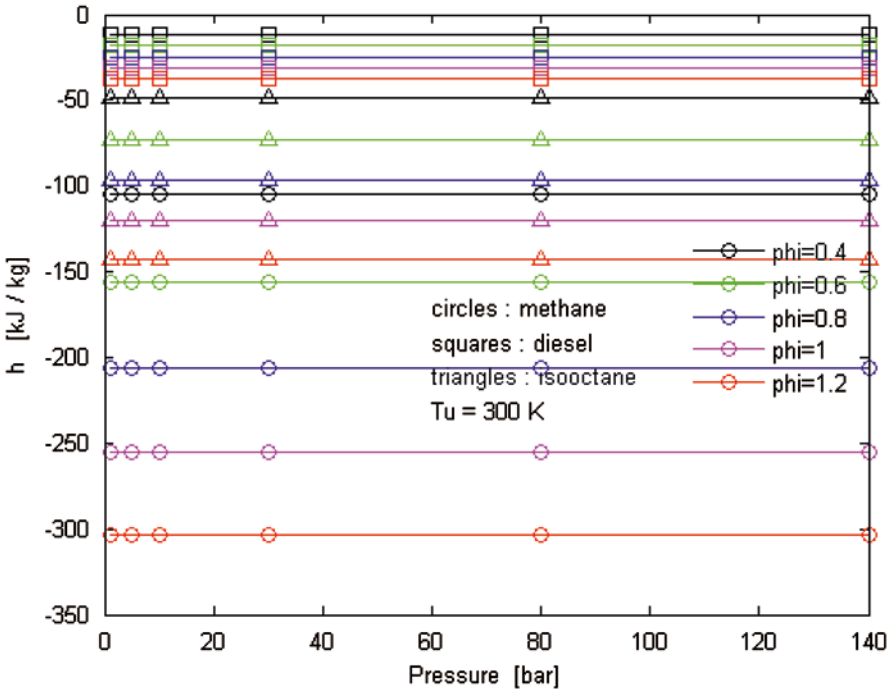


Fig. 19.3 Change of specific enthalpy vs. pressue

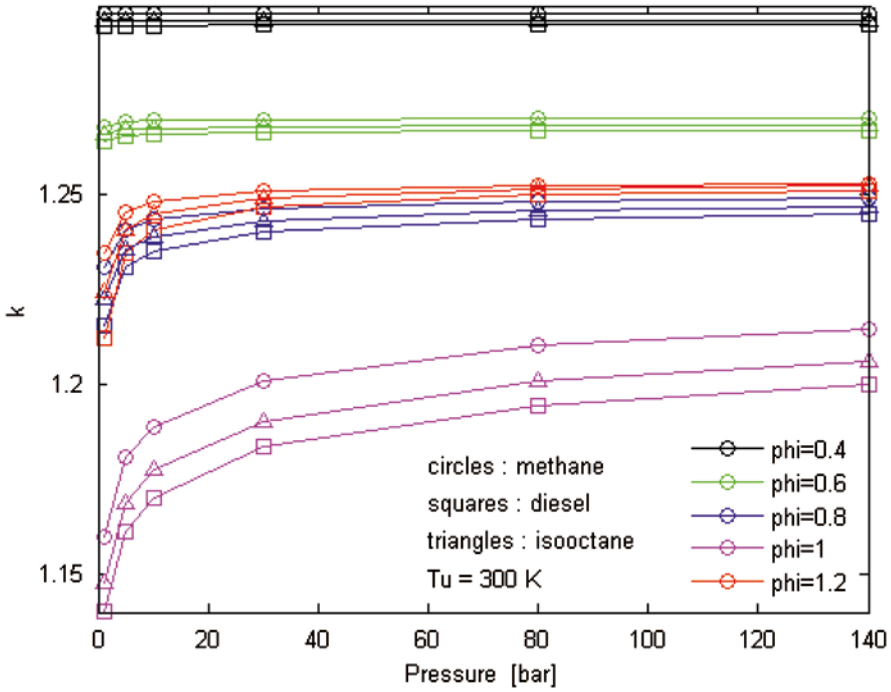


Fig. 19.4 Change of isentropic exponent vs. pressure

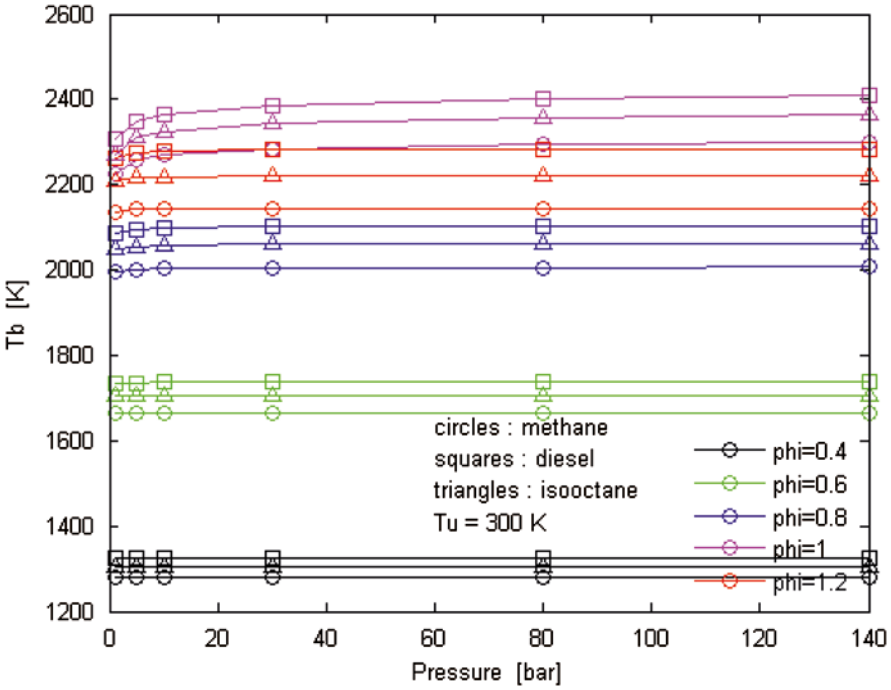


Fig. 19.5 Change of burned gas temperature vs. pressure

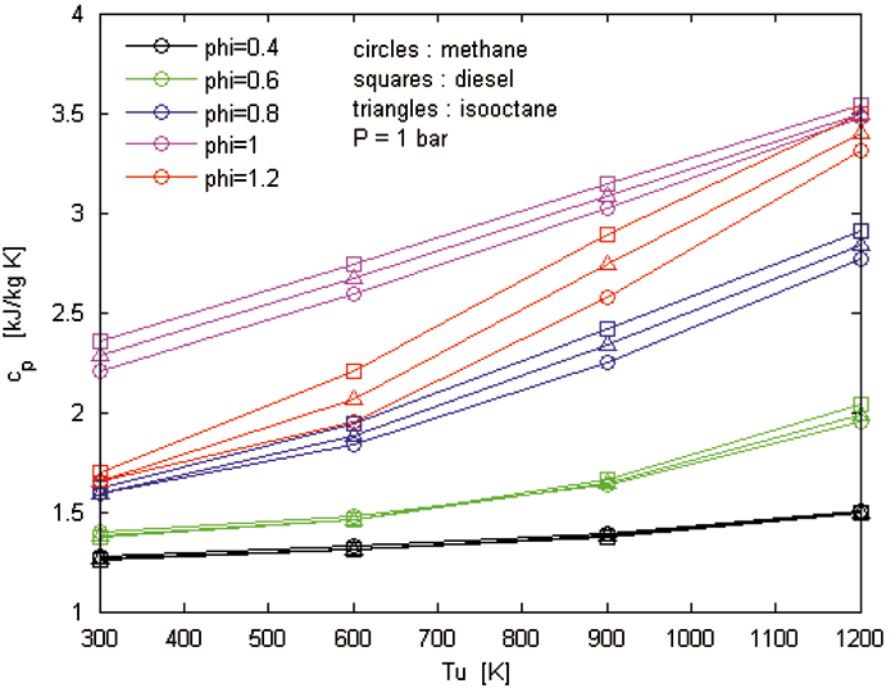


Fig. 19.6 Change of constant pressure specific heat vs. unburned temperature

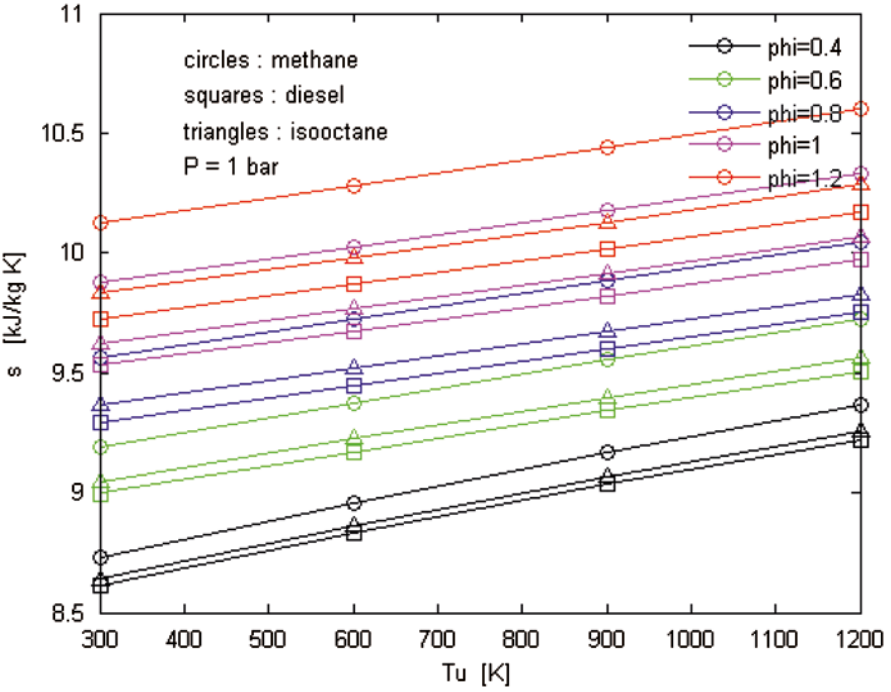


Fig. 19.7 Change of specific entropy vs. unburned temperature

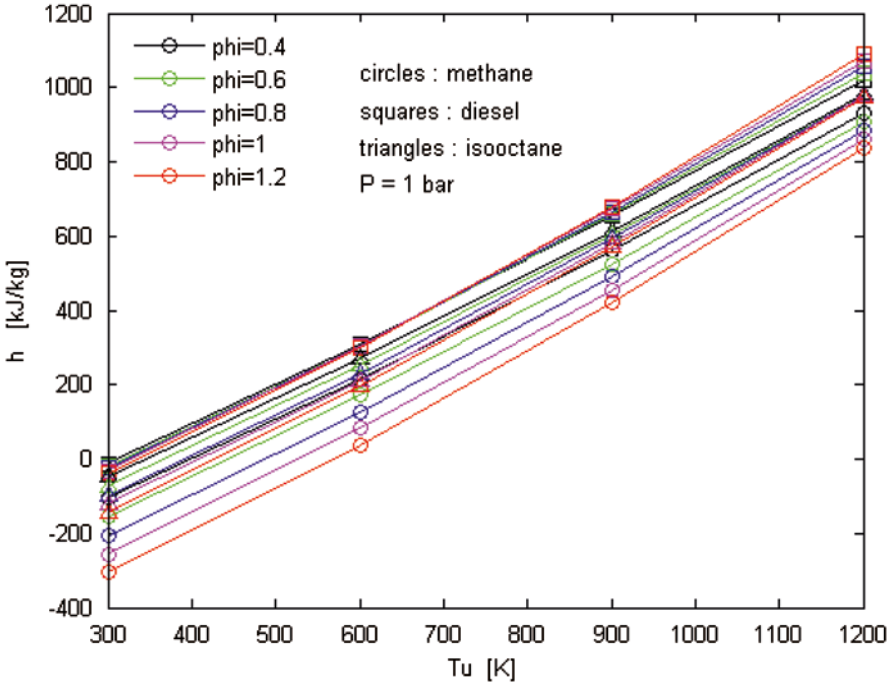


Fig. 19.8 Change of specific enthalpy vs. unburned temperature

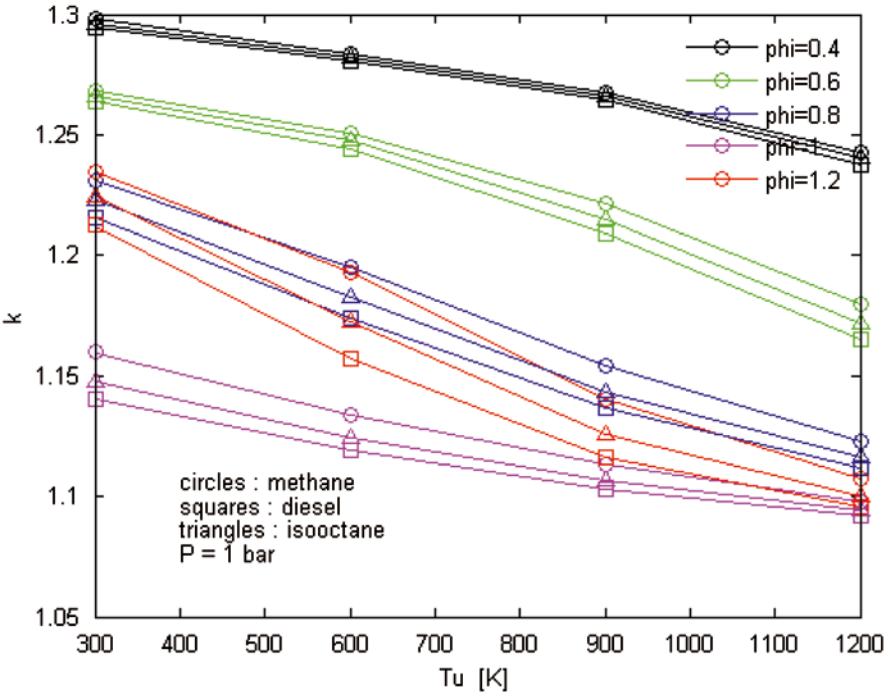


Fig. 19.9 Change of isentropic exponent vs. unburned temperature

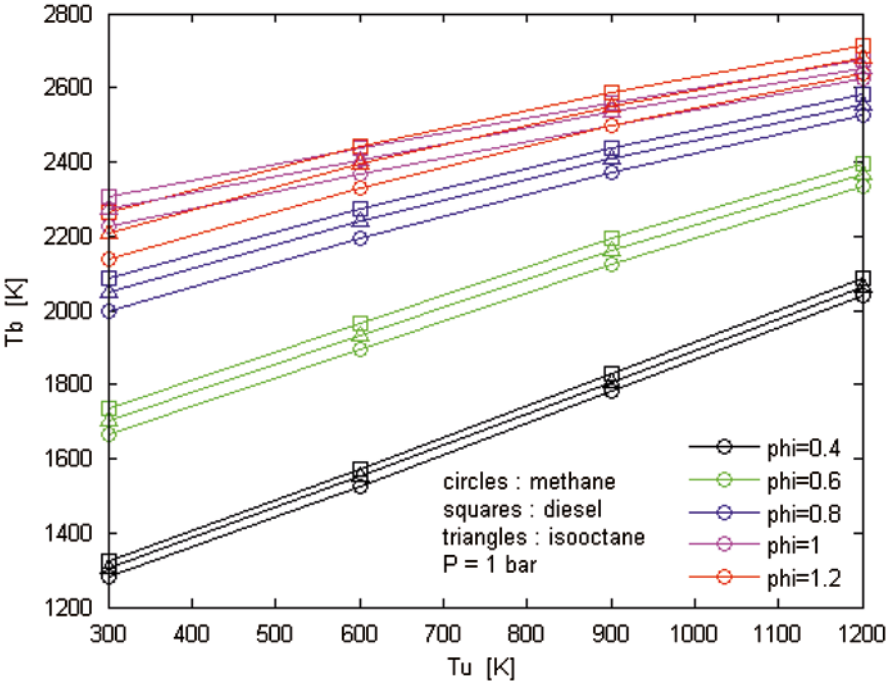


Fig. 19.10 Change of burned gas temperature vs. unburned temperature

increases with unburned mixture temperature. The temperature values are the temperature values at which Figs. 19.6–19.9 are obtained.

References

1. Heywood J B (1988) Internal combustion engine fundamentals. McGraw-Hill, New York
2. Rashidi M (1998) Calculation of equilibrium composition in combustion products. Appl. Therm. Eng., 18(3–4): 103–9
3. Rakopoulos C D, et al. (1994) A fast algorithm for calculating the composition of diesel combustion products using 11 species chemical-equilibrium scheme. Adv. Eng. Software, 19(2): 109–19
4. Lughofer E, et al. (2011) Identifying static and dynamic prediction models for NOx emissions with evolving fuzzy systems. Appl. Soft Comput., 11(2): 2487–500
5. Canakci M, et al. (2009) Prediction of performance and exhaust emissions of a diesel engine fueled with biodiesel produced from waste frying palm oil. Expert Syst. Appl., 36: 9268–80
6. Sahin Z and O Durgun (2009) Prediction of the effects of ethanol-diesel fuel blends on diesel engine performance characteristics, combustion, exhaust emissions, and cost. Energy & Fuels, 23: 1707–17
7. Bebar L, et al. (2002) Low NOx burners – prediction of emissions concentration based on design, measurements and modelling. Waste Manage. (Oxford), 22(4): 443–51
8. Ouenou-Gamo S, M Ouladsine, and A Rachid (1998) Measurement and prediction of diesel engine exhaust emissions. ISA Trans., 37(3): 135–40

9. Arsie I, et al. (1998) Models for the prediction of performance and emissions in a spark ignition engine: A sequentially structured approach. Society of Automotive Engineers, Warrendale, Pa
10. Foster T J, et al. (1998) Measurement and prediction of NO and no2 emissions from aero engines in RTO AVT Symposium on Gas turbine engine combustion, emissions and alternative fuels Lisbon
11. Tsague L, et al. (2007) Prediction of emissions in turbojet engines exhausts: Relationship between nitrogen oxides emission index (einox) and the operational parameters. Aerosp. Sci. Technol., 11(6): 459–63
12. Wilson C W, et al. (2004) Measurement and prediction of emissions of aerosols and gaseous precursors from gas turbine engines (partemis): An overview. Aerosp. Sci. Technol., 8(2): 131–43
13. Katsuki M and Y Mizutani (1977) Simplified reactive flow model of gas-turbine combustors for predicting nitric-oxide emission. Combust. Sci. Technol., 17(1–2): 19–28
14. Mohamed H, H Ben Ticha, and S Mohamed (2004) Simulation of pollutant emissions from a gas-turbine combustor. Combust. Sci. Technol., 176(5–6): 819–34
15. Keating E L (2007) Applied combustion. 2nd ed.: CRC Press/Taylor & Francis, Boca Raton
16. Ferguson C R (1986) Internal combustion engines: Applied thermosciences. John Wiley, New York
17. Olikara C and G L Borman (1975) A computer program for calculating properties of equilibrium combustion products with some applications to I.C. Engines. Society of Automotive Engineers, Warrendale, Pa
18. Kayadelen H K and Y Üst (2013) Prediction of equilibrium products and thermodynamic properties in H2O injected combustion for CHON type fuels. Fuel, 113: 389–401
19. Turns S R (2000) An introduction to combustion: Concepts and applications. 2nd ed.: WCB/McGraw-Hill, Boston

Chapter 20
Application of the Artificial Neural Networks and Fuzzy Logic for the Prediction of Reactivity of Molecules in Radical Reactions

V. E. Tumanov

Abstract This paper discusses the use of feed-forward artificial neural network to predict the reactivity of organic molecules in the bimolecular radical reactions in the liquid phase and the use of the fuzzy knowledge base to identify the empirical dependence of the activation energy of reactions phenyl radical ($C_6H^{\circ}_5$, 4-CH_3–$C_6H^{\circ}_5$, 4-Br–$C_6H^{\circ}_5$, 4-Cl–$C_6H^{\circ}_5$ etc.) with hydrocarbons in the liquid phase from thermochemical data. Also artificial neural network was used to predict the values of C–H bonds dissociation energies of hydrocarbons on experimental data of radical reactions R$^{\circ}$ + RH.

Keywords Feed-forward artificial neural network · Subject-oriented science intelligence system · Reactivity of organic molecules · Radical reaction · Fuzzy knowledge base · Rate constant · Activation energy · Bond dissociation energy

20.1 Introduction

Currently artificial neural network (ANN) is widely used in solving applied problems of automated processing of scientific data. The main fields of ANN application in chemical and biochemical studies are given in a review [1]. Most works in this area are devoted to the correlation between the structure of chemical compounds and the physicochemical properties or biological activity they showed. In the physical chemistry, the main directions of ANN application are the simulation of chemical processes and the simulation of the dynamic properties of the molecules and the systems.

On the one hand, the physical chemistry of radical reactions accumulated large amount of experimental data on the reactivity (specific reaction rate or activation energies) of molecules in radical reactions [2, 3]. On the other hand, the experiments

V. E. Tumanov (✉)
Laboratory of Information Support for Research, Institute of Problems of Chemical Physics RAS, Semenov ave, 1, 142432, Chernogolovka, Russian Federation
e-mail: tve@icp.ac.ru

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,* Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_20,

to quantify the reactivity of molecules in radical reactions are an expensive and time-consuming task. Carrying out the quantum chemical calculations is time consuming, and the resulting data for these calculations are not sufficiently reliable. Therefore, the development of ANN based on existing experimental data to predict the reactivity of organic molecules in radical reactions is the vital task.

Knowledge of the reactivity of organic molecules in the radical reactions is necessary for the development of new organic materials, the design of new drugs, design of technological processes, planning and conducting a scientific experiment, the training of students and graduate students.

This paper discusses the use of feed-forward artificial neural network to predict the reactivity of organic molecules in the bimolecular radical reactions in the liquid phase and the use of the fuzzy knowledge base to identify the empirical dependence of the activation energy of reactions phenyl radical ($C_6H^{\circ}_5$, 4-CH_3–$C_6H^{\circ}_5$, 4-Br–$C_6H^{\circ}_5$, 4-Cl–$C_6H^{\circ}_5$ etc.) with hydrocarbons in the liquid phase from thermochemical data.

20.2 Problem Formulation

Experimentally, the activation energy (E) or a classical potential barrier (E_e) determines the reactivity of organic molecules in a radical reaction:

$$E_e = E - 0.5(hL\nu_i - RT) \tag{20.1}$$

ν_i is a frequency of the stretching vibrations for the bond being broken, R is the gas constant, h is the Planck constant, L is the Avogadro number, and T is the reaction temperature (K).

Specific rate constant (k) of chemical reaction is calculated by the formula:

$$k = nA_0 \exp(-E / RT) \tag{20.2}$$

where: A_0 is collision frequency per one equireactive bond, n is the number of equireactive bonds in a molecule.

When designing the information space for ANN predictions of the reactivity the functional relationship between the reactivity of the chemical reaction and the thermochemical properties (enthalpy of reaction—ΔH) is used.

N.N. Semenov was the first to pay attention to the functional relationship between the reactivity and reaction enthalpy (known as Polanyi—Semenov's ratio [4]):

$$E = B - \gamma\Delta H \tag{20.3}$$

where B and γ—empirical coefficients.

The works [5, 6] proposed the empirical models of elementary radical reaction, which allowed constructing non-linear correlation dependences between the classical potential barrier of the radical reaction and its thermochemical properties:

- approximation of the above mentioned dependence in the work [5] by the parabola:

$$br_e = \alpha\sqrt{E_e - \Delta H_e} - \sqrt{E_e} \tag{20.4}$$

- approximation of the above mentioned dependence in the work [6] in the form of the tacitly set curve:

$$br_e = D_{ei}^{1/2} \ln\left(\frac{D_{ei}^{1/2}}{D_{ei}^{1/2} - E_e^{1/2}}\right) + \alpha D_{ef}^{1/2} \ln\left(\frac{D_{ef}^{1/2}}{D_{ef}^{1/2} - (E_e - \Delta H_t)^{1/2}}\right) \tag{20.5}$$

Under the proposed empirical models assuming the harmonic stretching vibrations, the reaction of the radical abstraction $R^{o} + R_1H \rightarrow RH + R^{o}_1$ (where R^{o}·and R^{o}_1—alkyl radicals, and RH and R_1H—hydrocarbon molecules) has the following parameters [5]:

1. Enthalpy $\Delta H_e = D_i - D_f + 0.5\,(hLv_i - hLv_f)$ including the energy difference of zero-point vibrations of broken and formed bonds (it represents a change in the potential energy of the system). Here v_i—frequency of vibration of the molecule along the broken bond, v_f—frequency of vibration of the molecule along the formed bond, D_i—bond dissociation energy of the broken bond, $D_{ei} = D_i + 0.5hLv_i$, D_f—bond dissociation energy of the formed bond, $D_{ef} = D_f + 0.5hLv_f$.
2. The classical potential barrier of the activation E_e (1), which includes the zero-point energy of the broken bond.
3. The parameters $b = \pi(2\mu_i)^{1/2}v_i$ and $b_f = \pi(2\mu_f)^{1/2}v_f$, that describe the potential energy dependence of the atoms vibration amplitude along the breaking (i) and forming (f) valence linkage. $2b^2$—the force constant of the linkage, μ_i—the reduced mass of the atoms due to breaking bond, μ_f—the reduced mass of the atoms due to forming bond.
4. The parameter r_e, which is the integrated stretching of breaking and forming bonds in the transition state.
5. Pre-exponential factor A_0 per equireactive bond in the molecule.
6. According to statistically determined value of br_e, based on formula (4), it is possible to estimate the value of the classical potential barrier by the formula:

$$\sqrt{E_e} = \frac{br_e}{1-\alpha^2}\left(1 - \alpha\sqrt{\left(1 - \left(1-\alpha^2\right)\frac{\Delta H_e}{\left(br_e\right)^2}\right)}\right) \tag{20.6}$$

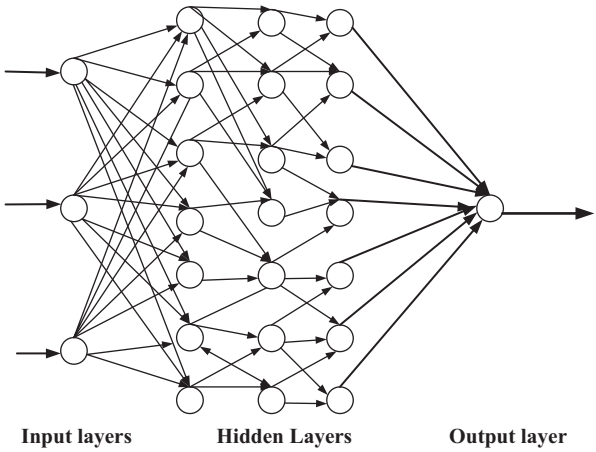


Fig. 20.1 Typical architecture of feed-forward artificial neural network

Thus, we can assume that the dependence of the classical potential barrier E_e of the thermochemical characteristics of the reagents and the kinetic characteristics of the radical reactions can be represented as the functional relation:

$$E_e = F(\Delta H_e, T, nA_0, \alpha) \tag{20.7}$$

Then the task of ANN works in predicting the values of the classical potential barrier E_e as a functional relation of the thermochemical and kinetic characteristics of the reagents with subsequent calculation of the activation energies and specific reaction rate by the formulae (1.1) and (1.2) reduce to the approximation of unknown functional relation (1.7).

20.3 Problem Solution

20.3.1 Artificial Neural Networks for the Prediction of Reactivity of Molecules

To approximate the dependence (1.7) we used feed-forward artificial neural network [7] with a typical architecture shown in Fig. 20.1. We used the ANN having four inputs, three inner layers, each of seven neurons and one output.

ANN work is set by the formulae:

$$\begin{aligned} NET_{jl} &= \sum_i w_{ijl} x_{ijl}, \\ OUT_{il} &= \Phi\left(NET_{jl} - \theta_{jl}\right), \\ x_{ij(l+1)} &= OUT_{il}, \\ \delta &= 0.5 \sum_j \sum_k \left(y_j^k - d_j^k\right)^2. \end{aligned} \tag{20.8}$$

Table 20.1 Training results of ANN

Reaction	E	E_e	E_{ANN}
$C^oH_3+CH_2ClBr$	27.17	36.57	30.97
$C_6H^o_5+(CH_3)_4C$	23.82	43.17	19.00
$C^oCl_3+CH_3(CH_2)_5CH_3$	46.88	67.35	52.35
$C^oH_3+cyclo$-$[(CH_2)_6]$	44.17	46.80	42.13
$C_6H^o_5+cyclo$ -$[(CH_2)_5]$	27.99	31.04	27.48
$C^oH_3+cyclo$-$[CH(CH_3)(CH_2)_4]$	30.23	39.67	33.18
$C^oCl_3+C_6H_5CH_3$	44.30	52.62	40.88
$C_6H^o_5+C_6H_5CH_3$	20.67	24.45	16.01

where the index i will always denote the input number, j—number of neurons in the layer, l—number of the layer; x_{ijl}—i-th input of j-th neuron in the layer l; w_{ijl}—weighting factor of the i-th input neuron number j in layer l; NET_{jl}—signal NET j-th neuron in layer l; OUT_{jl}—the output signal of the neuron; θ_{jl}—the threshold of neuron j in the layer l; x_{jl}—the input column vector of the layer l.

ANN input vector is set as the vector $x_0=\{T, D_{ei}, D_{ef}, nA_0, \alpha\}$, output data is equal to E_e.

The method of back propagation of the error [7] was used as training procedure. Activation function is a sigmoid function and is set by the following formula:

$$f(x)=\frac{1}{1+e^{-\beta x}} \tag{20.9}$$

The parameter $\beta>0$ was chosen experimentally.

For ANN training 3000 iterations were required on training set of 295 samples. Training set was constructed from the elemental radical reactions R'+RH in the liquid phase, where R'—a radical and RH—a hydrocarbon molecule.

Table 20.1 shows the comparison of the predictions of the values of the classical potential barrier of the reaction using ANN (E_{ANN}), the experimental values of activation energy (E) and the values calculated by the formula (6) of the classical potential barrier (E_e).

The error of the values prediction of the classical potential barrier of the radical reaction using ANN in the control sample (of 20 samples) was 3.34±2.0 kJ/mol, which is within the experimental error (±4 kJ/mol). The error of values prediction of the classical potential barrier for the radical reaction (1.6) on the same control sample was 9.5±7.0 kJ/mol. ANN predicts better than the calculation by formula (1.6). This is due to the size of the statistical error br_e, which defines the class of radical reactions. Thus, the ANN better approximates the functional dependence (1.7) by calculating the weight matrix relations.

20.3.2 Fuzzy Knowledge Base for Predicting the Reactivity of Phenyl Radical Reactions with Hydrocarbons

The artificial neural network (1.8) doesn't consider influence of solvent and the reactionary center on a prediction of value of a classical potential barrier. To consider such influence an attempt to use the fuzzy knowledge base was made.

In this paper to determine the parameter space (input) for the identification of the activation energy of radical reaction $E_{экс}$ the model intersecting terms Morse is used, and it is determined by the correlation ratio [6]:

It is assumed that classical potential barrier activation is given with nonlinear object (parameter α is constant for the entire sample):

$$E_e = F\left(D_{ef}, D_{ei}, S_1, S_2\right) \tag{20.10}$$

For modeling relation two qualitative parameters are used, where S_1—characteristic hydrocarbon reaction center (defined class of compounds considered 40 classes) and S_2—characteristic of the solvent (non-polar, polar, polar with complexing, non-polar with complexing) [2]. The sample (647 reactions) is obtained from the database of rate constants of radical liquid-phase reactions object-oriented system of scientific knowledge in physical chemistry of radical reactions [6]. The characteristic of the sample is the combination of quantitative and qualitative parameters in a model of identification.

In the process of building the fuzzy knowledge base, input and output variables in this ration (1.10) are considered as linguistic variables defined on the corresponding universal sets. As a member of the membership function for indecipherable terms G is selected:

$$\mu^G(x) = \frac{1}{1+\left(\dfrac{x-b}{c}\right)^2} \tag{20.11}$$

where initial values of the parameters b and c were chosen by expert manner. The fuzzy knowledge base by a system of logical statements in the form:

$$\mu^{d_j}\left(x_1, x_2, x_3, x_4\right) = \bigcup_{p=1}^{k_i}\left\{w_{jp}\left[\bigcap_{i=1}^{4}\mu^{a_i^{jp}}(x_i)\right]\right\} \tag{20.12}$$

where $x_1 \div x_4$—linguistic evaluation of input variables D_{ei}, D_{ef}, S_1, S_2; d_j—linguistic evaluation of the output variable E_e, w_{jp}—weight matrix rules.

For experimental samples for the reactions of phenyl radicals with hydrocarbons the fuzzy knowledge base has been set up with the help of genetic algorithm [7]. At this time it includes 634 fuzzy rules. The comparison of classical potential barrier of activation for the reactions of phenyl radicals with hydrocarbons *Ee* (in Table 20.2),

Table 20.2 Comparison of classical potential barrier of activation for the reactions of phenyl radicals with hydrocarbons

Radical	Hydrocarbon	Solvent	Ee (kJ/mole)	*Ekl_eks* (kJ/mole)
$C_6H_5^o$	*cyclo*-[CH=CHCH=CH(CH$_2$)$_2$]	Non-polar with complexing	12.01	11.96
$C_6H_5^o$	$C_6H_5CH(CH_3)_2$	Polar	21.10	21.06
$C_6H_5^o$	$(CH_3)_2CHOH$	Polar	21.40	21.37
$C_6H_5^o$	$(CH_3)_2CHOH$	Polar	21.46	21.44
$C_6H_5^o$	*cyclo*-[CH=CH(CH$_2$)$_4$]	Polar	21.78	21.91
$C_6H_5^o$	$C_6H_5CH_2OH$	Polar with complexing	21.90	21.86
4-NO_2-$C_6H_5^o$	$C_6H_5CH(CH_3)_2$	Non-polar with complexing	24.00	23.98
$C_6H_5^o$	$C_6H_5CH(CH_3)_2$	Non-polar with complexing	24.00	23.98
4-Br-$C_6H_5^o$	$C_6H_5CH(CH_3)_2$	Non-polar with complexing	24.41	24.38
$C_6H_5^o$	$C_6H_5CH(CH_3)_2$	Polar	24.43	24.42
$C_6H_5^o$	CH_3CH_2OH	Polar with complexing	24.50	25.51
$C_6H_5^o$	$C_6H_5C(CH_3)_2CH(O)$	Non-polar with complexing	25.58	25.55

is obtained using the fuzzy knowledge base with the calculated from the experimental values of the activation energy *Ekl_eks* corresponding reactions.

As can be seen from Table 20.2, there is good agreement between experimental values of classical potential barrier and values obtained by the fuzzy inference.

20.3.3 Estimate of Dissociation Energy of C–H Bonds of Hydrocarbons by Artificial Neural Network

The bond dissociation energy of organic molecules is one of its most important thermochemical characteristics. Determination of dissociation energies of organic molecules is a complex and time-consuming experimental task; as a rule, the results of quantum chemical calculations don't have sufficient reliability. Presently this characteristic is known not for many compounds [8]. Therefore, estimate of bond dissociation energies of complex organic molecules on the basis of the use of the ANN is an actual scientific problem.

The purpose of this section is to develop the ANN to estimate the dissociation energy C-H bonds in hydrocarbons on kinetic and thermochemical data radical reactions, assuming that the training sample feature space is constructed on the empirical model intersecting terms Morse for bimolecular radical reactions [6] is given by the nonlinear correlation relationship (1.5).

It can be assumed that the dependence of the dissociation energy bonds on parameters on radical reaction can be represented as a function of four variables:

$$D_{ei} = F\left(D_{ef}, E_e, \nu_i, \nu_f\right) \tag{20.13}$$

Table 20.3 Training results of ANN

Molecule	D_{ann}	*D* [8]	Δ*D*
cyclo-[($\underline{CH_2}$)$_5$]	416.3	425.8	−9.5
trans-$CH_3CH{=}CH\underline{CH_3}$	374.7	374.2	0.5
cis-$CH_3CH{=}CH\underline{CH_3}$	374.8	373.2	1.6
$C_6H_5\underline{CH_3}$	392.8	392.4	0.4
$C_6H_5\underline{CH_2}CH_3$	383.2	381.5	1.7
$C_6H_5\underline{CH}(CH_3)_2$	368.5	369.0	−0.5
$CH_3\underline{CH}(O)$	396.0	391.2	3.5
$(CH_3)_3C\underline{CH}(O)$	392.5	392.5	0.0
$CH_3(O)\underline{CH_2}CH_3$	402.6	410.0	−7.4
cyclo-[$C(O)(\underline{CH_2})_5$]	407.4	411.5	−4.1
cyclo-[$O(CH_2)_4$]	405.6	409.0	−3.4
$C_6H_5O\underline{CH_3}$	400.8	402.4	−1.6
$CH_3C(O)O\underline{CH_3}$	409.8	410.3	−0.5
$(CH_3)_2\underline{CH}OH$	403.8	407.9	−4.1
$(CH_3)_2\underline{CH}C(O)OH$	401.6	405.7	−4.1
cyclo-[$(CH_2)_5\underline{CH}(OH)$]	402.2	405.8	3.6
$CH_3\underline{CH_2}CN$	405.5	404.6	0.9
$C_6H_5S\underline{CH_3}$	407.0	406.5	0.5

Then the problem of the ANN to assess the value of bond dissociation energy of the organic molecule is reduced to the approximation of the unknown function (1.12).

For approximation of the dependence (1.2) the multilayer ANN back propagation Hopfield was used [7]. This ANN has three layers: the first layer has four neurons (equals the number of input parameters), the second layer has 10 neurons, the third layer has one neuron (result *Dei*), sigmoidal parameter of normalization 0.5. Training continues as long as the error parameter is more than 0.00001. Functioning of ANN is given by (1.7).

For ANN training 900023 iterations took place on the training set of 667 samples. The training sample was constructed from the elementary radical reactions R'+RH in the liquid phase, where R'—radical and RH—hydrocarbon molecule.

Compares the estimates of C-H bonds dissociation energy of hydrocarbons by using the ANN (*D*ann) and the experimental values of bond dissociation energy (*D*) taking into account the zero-point energy connection (for C–H bond is 17.4 kJ/mol) is given in Table 20.3.

Absolute error of predicted values of C–H bond dissociation via ANN on kinetic and thermochemical data of radical reactions doesn't exceed 10.0 kJ/mol (according to catalog's data [8] the error in determination of bond dissociation energy can reach up to 12.5 kJ/mol). Standard deviation is 4.3 kJ/mol, which is within experimental error (±4 kJ/mol). In this way, the developed ANN estimates bond dissociation energies with acceptable accuracy, which means a good approximation of the functional dependence (1.12).

20.4 Conclusions

It was the first time when feed-forward ANN was used for the approximation on the experimental data of the functional dependence of the classical potential barrier of the chemical reaction from the thermochemical characteristics of the reagents and reaction kinetic parameters.

The results of the prediction of the reactivity of liquid-phase reactions of the hydrocarbons with hydrocarbon radicals are within the limits of the experimental error.

For the first time (on the example the reactions of phenyl radicals with hydrocarbons) the attempt to identify the dependence of the activation of classical potential barrier of the reactions of phenyl radicals with hydrocarbons fuzzy knowledge base built on the basis of quantitative and qualitative parameters was done.

For the first time ANN was used to predict the values of C-H bonds dissociation energies of hydrocarbons on experimental data of radical reactions $R^{\circ}+RH$. The predict results of C-H bonds dissociation energy of hydrocarbons on kinetic and thermochemical data of radical reactions are within the experimental error.

References

1. Gasteiger J, Zupan J (1993) Neural networks in chemistry," *Angev. Chem. Int. Ed. Engl*. 32:503–527
2. Tumanov V, Gaifullin G (2012) Subject-oriented science intelligent system on physical chemistry of radical reactions", *Modern Advances in Intelligent Systems and Tools* 431:121–126
3. Mallard WG, Westley F, Herron JT, Hampson RF (1994) NIST Chemical Kinetics Database – Ver. 6.0. NIST Standard Reference Data, Gaithersburg, MD.
4. Semenov NN (1935) Chemical Kinetics and Chain Reactions. London, Oxford Univ. press
5. Denisov ET (1997) New empirical models of free radical abstraction reactions. Uspekhi Khimii 66:953–971
6. Denisov ET, Tumanov VE (1994) Transition-State Model as the Result of 2 Morse Terms Crossing Applied to Atomic-Hydrogen Reactions. Zhurnal Fizicheskoi Khimii 68:719–725
7. Lill JH (2011) Fuzzy Control and Identification. John Wiley & Sons, Incorporated
8. Luo YR (2003) HandOther of Bond Dissociation Energies in Organic Compounds. CRC Press, Boca Raton, FL

Chapter 21
Plasma-Fuel Systems for Environment Enhancement and Processing Efficiency Increasing

V. E. Messerle and A. B. Ustimenko

Abstract Plasma-fuel systems for thermochemical treatment for combustion, gasification, pyrolysis, hydrogenation, radiation-plasma, and complex conversion of solid fuels, including uranium-containing slate coal, and cracking of hydrocarbon gases, are presented. The use of these plasma technologies for obtaining target products (hydrogen, hydrocarbon black, hydrocarbon gases, synthesis gas, and valuable components of the coal mineral mass) meet the modern environment and economic requirements. Plasma coal conversion technologies are characterized by a small time of reagents retention in the plasma reactor and a high rate of the original substances conversion to the target products without catalysts. Thermochemical treatment of fuel for combustion is performed in a plasma-fuel system, representing a reaction chamber with a plasma generator, while other plasma fuel conversion technologies are performed in a combined plasma reactor of 100 kW nominal power, in which the area of heat release from the electric arc is combined with the area of chemical reactions.

Keywords Fuel · Processing · Plasma generator · Conversion efficiency · Environment

21.1 Introduction

The global energy sector is oriented to use—currently and in the foreseeable future (till 2100)—organic fuels, basically, low-grade coal, the share of which is 40.6 % in electricity engineering and 24 % in heat engineering. Therefore, the development of technologies for efficient and environmentally clean use of such coal is a priority

A. B. Ustimenko (✉)
Research Institute of Experimental and Theoretical Physics, Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: ust@physics.kz

V. E. Messerle
Combustion Problems Institute, Almaty, Kazakhstan

Institute of Thermophysics of Russian Academy of Science, Novosibirsk, Russia

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,* Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_21,

problem of today. The analyzed plasma fuel conversion technologies meet these requirements. The plasma fuel conversion technologies have become very urgent recently due to depletion of oil and gas deposits, reduced quality of solid fuels, and increasing NPP capacities.

This work presents the results of long-term studies of plasma technologies of pyrolysis, hydrogenation, thermochemical treatment for combustion, gasification, hybrid (radiation-plasma) and complex conversion of solid fuels, as well as cracking of liquefied petroleum gas [1–9]. The use of these technologies for production of target products (hydrogen, hydrocarbon black, hydrocarbon gas, synthetic gas, valuable components of coal mineral mass, including rare earth elements) corresponds to contemporary environmental and economic requirements to the main industrial sectors. Plasma solid fuel conversion technologies differ, primarily, in concentrations of the reducing gas (air, water vapor, carbon dioxide, and oxygen), conditioned by different values of the excess oxidant coefficient α. The value $\alpha = 0$ corresponds to coal pyrolysis, while the value $\alpha = 1$ corresponds to complete coal gasification. It should be noted that the theoretical quantity of air required for combustion of 1000 kg of such coal makes 5250 kg, which is almost 2.5 times higher than the quantity required for its complete gasification.

21.2 Plasma Technologies Discussion

During plasmachemical gasification of a low-grade coal with the ash content 40 % and the combustion heat 16,632 kJ/kg at $\alpha = 0.5$ the gaseous phase is basically represented by the synthetic gas ($CO + H_2$). With increasing temperature (1800–2600 K), all mineral components go to the gaseous phase in the form of gaseous substances, such as Al, Si, SiS, Fe, Al_2O, SiC_2, and others.

The plasmachemical cracking technology includes the heating of hydrocarbon gases in an electric-arc combined reactor to the temperatures of their pyrolysis (1900–2300 K), generating a highly dispersed hydrocarbon black and hydrogen in a single technological process. In the temperature range 2500–5000 K, the gaseous phase includes a number of hydrocarbons (C_3H, C_2H_2, C_4H_2, etc.) which, with increase in temperature, dissociate into their components, hydrogen and carbon. All condensed carbon goes to the gaseous phase at temperatures exceeding 3200 K.

Plasmachemical hydrogenation of solid fuels, representing coal pyrolysis in the hydrogen medium, makes it possible to produce acetylene and other unsaturated hydrocarbons (ethylene C_2H_4, propylene C_3H_6, ethane C_2H_6, etc.) from cheap low-grade coals by way of hydrogen plasma treatment [4]. Plasmachemical hydrogenation of coal is a new little-studied process of direct production of acetylene and alkenes in the gaseous phase, in contrast to traditional processes of coal hydrogenation (liquefying).

The experiments on hydrogenation of low-grade coal in a plasma reactor (Fig. 21.1), with the power of 50 kW and the consumption of coal 3 kg/h and of the propane–butane mixture 150 l/h, allowed the production of the following gas composition, wt.%: $C_2H_6 = 50$, $C_2H_2 = 30$, $C_2H_4 = 10$.

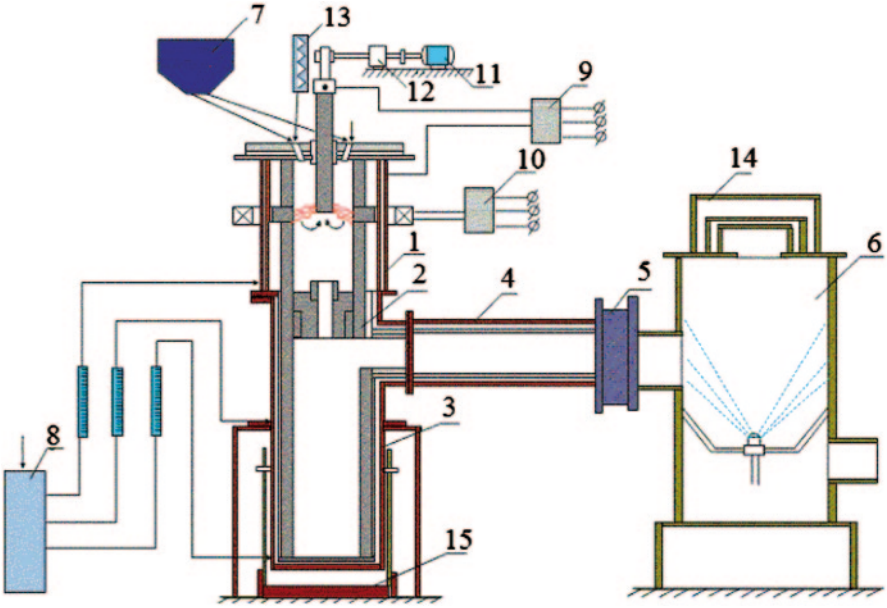


Fig. 21.1 The schematic diagram of the setup for plasmachemical fuel conversion: *1* plasmachemical reactor *2* diaphragm and chamber for gas and slag separation, *3* slag collector, *4* oxidation chamber, *5* diaphragm, *6* water scrubber, *7* solid fuel feeding, *8* water cooling system, *9*, *10* power supply system, *11*, *12* central electrode feeding system, *13* steam generator, *14* safety valve, *15* slag collector lift

Plasma ignition of coal is based on plasmachemical fuel conversion for combustion, resulting in the production from a low-grade coal of a two-component fuel (combustible gas and carbon residue). This high-reactivity two-component fuel is generated already at T = 900–1200 K. Thus, this process can be performed at a comparatively low specific power consumption (0.05–0.4 kW h/kg of coal) and can be used efficiently by thermal power plants for no-oil start-up of boilers and stabilized combustion of the pulverized coal flame [3, 7–9].

Plasma gasification, radiation-plasma, and complex coal conversion for the production of synthetic gas and valuable components from mineral coal were investigated using a versatile experimental plant (Fig. 21.1). In terms of environmental protection, these technologies are the most promising. The essence of these technologies is to heat the coal dust by electric-arc plasma, the oxidizing agent, to the complete gasification temperature, when the coal organic mass is converted to experimentally clean fuel, i.e., a synthetic gas free from ash particles as well as nitrogen and sulfur oxides.

Complex coal conversion includes, parallel to organic mass gasification, the reduction of mineral coal oxides in the same reduction volume by the carbon in the carbon residue and the generation of valuable components, such as hydrocarbon black, aluminum and carbon, as well as rare earth microelements: uranium, molybdenum, vanadium, etc.

Table 21.1 The integrated characteristics of plasma gasification of a low-grade brown coal

T (K)	Q_{sp} (kW·h/kg)	CO	H_2	X_C (%)	X_S (%)
		Volume (%)			
3100	5.36	45.8	49.4	92.3	95.2

Q_{sp} specific power consumption

Table 21.2 The reduction (Θ) of the coal mineral mass

Sampling places	T (K)	Θ (%)
Slag from the melt bathtub	2600–2800	8.5–44.0
Slag from the arc chamber wall	2600–2900	16.5–47.3
Material from the slag collector	2000–2200	6.7–8.3

The material and thermal balances helped to find the integral indicators for the process. Table 21.1 presents typical results of plasma-steam gasification of low-grade brown coal with the ash content 28 % and the calorific value 13,180 kJ/kg. The synthetic gas yield was 95.2 %, the carbon gasification (X_C) was 92.3 %, and coal desulfurization (X_S) was 95.2 %.

The reduction of solid residue samples from various units of the plant for plasmachemical fuel conversion and the special melt bathtub near graphite diaphragm 2 (Fig. 21.1) is shown in Table 21.2. As can be seen from the table, the reduced material was found in the slag in the form of ferrosilicon as well as silicon and iron carbides. The maximum reduction of the coal mineral oxides was observed in the slag from the walls of the reactor electric-arc chamber in the areas with maximum temperatures, reaching 47 %.

In the case of radiation-plasma conversion, the coal dust was pre-activated by an electron beam and then processed in plasmachemical reactor 1 (Fig. 21.1). The experiments were performed in a plasma gas generator with the rated power 100 kW. Measurements of the process material and heat balances gave the following integrated indicators: the mass-average temperature 2200–2300 K and the carbon gasification rate 82.4–83.2 %. It was found that the preliminary electronic activation of the coal dust fuel had a noticeable positive effect on the yield of the synthetic gas during its treatment. The yield of the synthetic gas during thermochemical treatment of the untreated coal dust before combustion was 24.5 %, and after electronic activation of coal the yield of the synthetic gas reached 36.4 %, i.e., a 48 % increase.

The essence of plasma technologies for the production of uranium, molybdenum, and vanadium oxides from solid fuel is the processing of its mixture by water steam in plasmachemical reactor 1 (Fig. 21.1) [1, 4–6]. The process of extraction of uranium, molybdenum, and vanadium from coal (slate coal) using plasma heating is as follows. The coal dust from hopper and water steam from steam boiler, with the coal-to-water steam weight ratio 8:12, is fed to plasmachemical reactor. In the reactor, the water steam plasma heats the coal dust to 2500–2900 K. When the coal is heated, the organic mass of the raw material is gasified and the uranium, molybdenum, and vanadium compounds in the mineral part are volatilized to the gaseous phase, containing synthetic gas, basically. Then the two-phase plasma flow (the gaseous phase + melted slag) is fed to chamber for separation of gas and slag,

Table 21.3 Integrated indicators of plasma processing of uranium-containing slate coal

No	G_f (kg/h)	G_{steam} (kg/h)	G_{steam}/G_f	T_{av} (K)	Q_{sp} (kW h/kg)	X_U (%)	X_{Mo} (%)	X_V (%)	X_C (%)
1	5.82	0	0	2900	2.84	48.0	54.5	58.6	56.2
2	8.40	0	0	2500	1.93	25.7	34.5	41.7	54.6
3	6.60	0.60	0.09	2700	2.20	78.6	79.0	81.3	66.4
4	4.33	0.40	0.09	3150	3.04	23.6	24.3	29.0	70.4

G_f consumption of fuel, G_{steam} steam rate, T_{av} averaged temperature, X_U uranium extraction, X_{Mo} molybdenum extraction, X_V vanadium extraction

wherefrom the slag is fed to slag collector, while the gaseous phase is sent to the series of heat exchangers for a two-step cooling and separate condensation of the target products. Table 21.3 presents the experimental results of plasma treatment of the uranium-containing slate coal with 0.02 % of uranium.

The experiments on plasma pyrolysis (cracking) of the propane–butane gas mixture were performed in a plasmachemical reactor with the rated power 100 kW (Fig. 21.1).

In these experiments, the consumption of the propane–butane mixture was 300 l/min and the electrical power of the plasmachemical reactor was 60 kW [4]. During the experiments, hydrogen and soot were separated in the water-cooled chamber for separation of the gaseous and condensed phases 2. Hydrogen was removed to oxidation chamber 4, while hydrocarbon black was precipitated on the reactor walls, water-cooled spiral copper collectors under the lid and the reactor output diaphragm as well as in soot collector 3. After the experiments, samples were taken from the above units of the reactor. Physical and chemical analysis of the hydrocarbon black samples was made by means of a transmission electron microscope, which showed that the products of plasma pyrolysis of the propane–butane gas mixture, condensed on the graphite electrodes of the plasma reactor, represented different nano-carbon structures, mostly, in the form of "huge" nanotubes (Fig. 21.2), having high electric conductivity and mechanical strength, 30 times higher than that of Kevlar fabric [4]. As shown on negative 9091, the sample mainly included large "wooly" carbon nanotubes about 100 nm in diameter and more than 5 μm in length. Negative 9094 shows huge carbon nanotubes with a drop-shaped inclusion in the metal phase. Their diameter reaches 300 nm. Negative 9104 shows a "stepped" carbon nanotube with the diameter 200 nm or more and an inner partition. Huge nanotubes may represent structures in the form of an octopus (negative 9110). The diameter of such octopus at the place of branching is about 400 nm. It is typical that the wall thickness of the hug nanotubes can vary from 30 nm (negative 9104) to 100 nm (negatives 9094 and 9110).

The experimental results confirmed that it is possible to produce hydrogen and condensed carbon containing nanostructures in the form of huge carbon nanotubes. These results were used to find a technical solution to create a pilot plant rated 1 MW with the capacity of the original natural gas 330 nm^3/h in order to perform plasmachemical cracking of hydrocarbon gases. The expected yield of the target

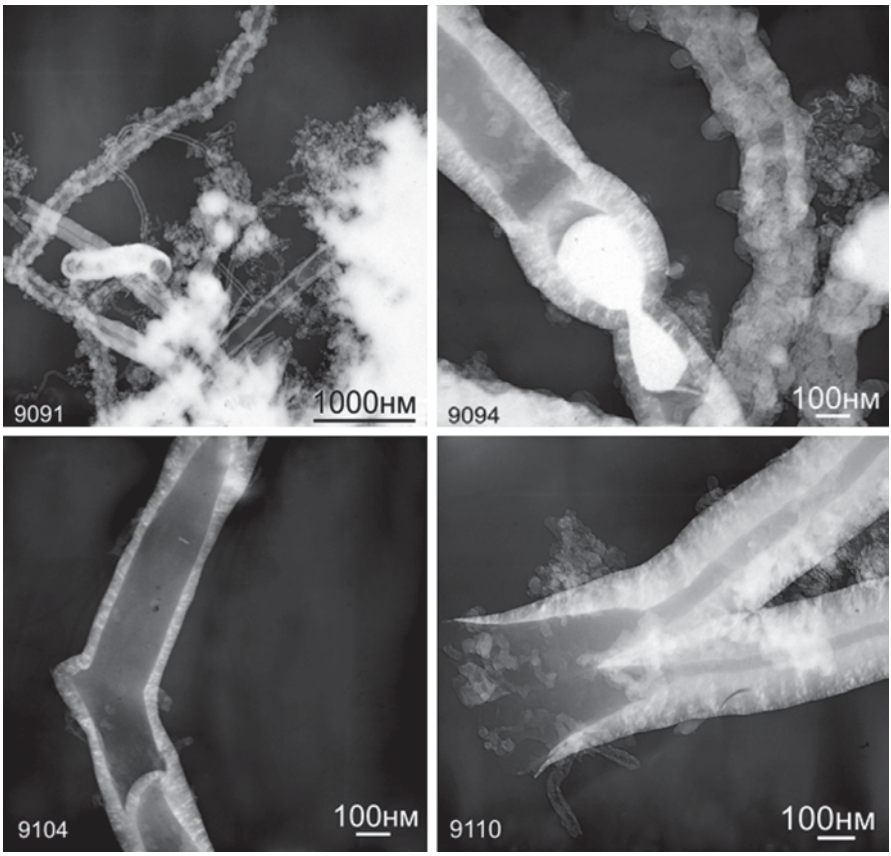


Fig. 21.2 Photos of carbon nanotubes produced by means of a transmission electron microscope

products will make 74% of hydrocarbon black (171 kg/h) and 25% of hydrogen (58 kg/h).

Plasmachemical conversion for combustion of coal from the original low-grade coal produces a high-reactivity two-component fuel that actively ignites when mixed with secondary air in the boiler furnace and burns stably without combustion of additional high-reactivity fuel, crude oil or gas, traditionally used for boiler firing and lighting of the dust-coal flame at thermal power plants (Table 21.4).

During complex coal conversion, the conversion of its mineral part requires high temperatures (2200–3100 K), increasing the specific power consumption to 2–4 kW h/kg. It gives a high degree of coal conversion (90–100%).

Plasma-steam gasification ensures the transfer to the gaseous phase of the organic coal mass, basically, which does not require very high temperatures as during complex treatment, thus allowing the process with comparatively low specific power consumption rates (0.5–1.5 kW h/kg) and high conversion rates (90–100%).

Radiation-plasma coal conversion increases the original fuel conversion rate by 48%.

Table 21.4 Optimal ranges of the recommended technological parameters for plasmachemical fuel conversion

Fuel/Plasma gas	T (K)	Specific power consumption (kW h/kg of fuel)	Fuel conversion rate (%)	Concentration (mg/Nm$^{3)}$	
				NO_x	SO_x
1. Plasmochemical preparation of coal for combustion (*air*)					
1.5–2.5	800–1200	0.05–0.40	15–30	1–10	1–2
2. Complex processing of coal (*water steam*)					
1.3–2.75	2200–3100	2–4	90–100	1–2	1
3. Plasma gasification of coal (*water steam*)					
2.0–2.5	1600–2000	0.5–1.5	90–100	10–20	1–10
4. Radiant-plasma processing of coal (*air*)					
1.5–2.5	800–1200	0.1–0.45	22–45	1–10	1–2
5. Plasma processing of uranium-bearing solid fuels (*water steam*)					
8–12	2500–3150	2–4	55–70	1–3	1–2
6. Plasmochemical hydrogenation of coal (*hydrogen*)					
10	2800–3200	6.5–8	70–100	0	0
7. Plasmochemical cracking of a propane-butane mixture					
18 м^3/ч	1500–2500	2.2–3.8	98–100	0	0

Plasma treatment of uranium-containing slate coal reached the following indicators: at temperatures 2700–2900 K the slate coal gasification rate was 56.2–66.4%, the conversion of microelements to the gaseous phase reached 48.0–78.6% for uranium, 54.5–79.0% for molybdenum, and 58.6–81.3% for vanadium, which agreed with the TERRA calculations in terms of quality [1].

Plasma hydrogenation of coal requires high temperatures (2800–3200 K), which results in high power consumption for this process (6.5–8 kW h/kg), thereby allowing high conversion rates (70–100%) for direct (one stage) production of acetylene and alkenes in the gaseous phase.

In order to ensure high conversion rates (98–100%) of the hydrocarbon gas in a combined plasma reactor, such high temperatures are not needed, which allows the process with relatively low specific power consumption (2.2–3.8 kW h/kg).

21.3 Conclusions

Plasmachemical conversion for combustion of coal from the original low-grade coal produces a high-reactivity two-component fuel that actively ignites when mixed with secondary air in the boiler furnace and burns stably without combustion of additional high-reactivity fuel, crude oil or gas, traditionally used for boiler firing and lighting of the dust-coal flame at thermal power plants.

During complex coal conversion organic mass of coal is transformed to synthesis gas and its mineral mass—to the set of valuable components including uranium Molybdenum, and vanadium ones.

Plasma-steam and plasma-air gasification ensures production of high quality synthesis gas, which can be used to synthesize methanol, and as a high potential reducing gas instead of metallurgical coke.

Radiation-plasma coal conversion can increases the original fuel conversion rate by 48 %.

Plasma hydrogenation of coal is resource-saving technology for direct production of acetylene and alkenes from solid fuel.

Plasma cracking allows getting hydrogen and black carbon from hydrocarbon gas.

References

1. Gorokhovski M, Karpenko EI, Lockwood FC, Messerle VE, Trusov BG, Ustimenko AB (2005) Plasma technologies for solid fuels: experiment and theory. Journal of the Energy Institute 78 (4): 157–171
2. Zhukov MF, Kalinenko RA, Levitski AA, Polak LS (1990) Plasmochemical processing of coal (in Russian). Moscow: Science
3. Messerle VE, Ustimenko AB (2012) Plasma ignition and combustion of solid fuel. (Scientific-and-technological basics) (in Russian). Saarbrucken, Germany: Palmarium Academic Publishing
4. Messerle VE, Ustimenko AB (2012) Plasma technologies for fuel Conversion. High Temperature Material Processes 16 (2): 97–107
5. Galvita V, Messerle VE, Ustimenko AB (2007) Hydrogen production by coal plasma gasification for fuel cell technology. International Journal of Hydrogen Energy 32 (16): 3899–3906
6. Messerle VE, Ustimenko AB (2007) Solid Fuel Plasma Gasification. In: Syred N, Khalatov A (eds.) Advanced Combustion and Aerothermal Technologies. Springer, pp 141–156
7. Gorokhovski MA, Jankoski Z, Lockwood FC, Karpenko EI, Messerle VE, Ustimenko AB (2007) Enhancement of Pulverized Coal Combustion by Plasma Technology. Combustion Science and Technology 179 (10): 2065–2090
8. Messerle VE, Karpenko EI, Ustimenko AB, Lavrichshev OA (2013) Plasma preparation of coal to combustion in power boilers. Fuel Processing Technology 107: 93–98
9. Karpenko EI, Messerle VE, Ustimenko AB (2007) Plasma-Aided Solid Fuel Combustion. Proceedings of the Combustion Institute 31: 3353–3360

Chapter 22
Computer Modeling of Optimal Technology in Material Engineering

V. A. Rusanov, S. V. Agafonov, A. V. Daneev and S. V. Lyamin

Abstract A technique of nonlinear mathematical programming good for grounding an optimal technological process of nitrogenization in a distributed environment of electrostatic field is proposed. The technique is based on the quadratic approximation for deviations of the vector argument of deviations of the vector argument of physics-chemical factors of metal working from some given regime of nitrogenization and imposes minimal requirements to experimental data in the process of identification of the mathematical model of the process of obtaining an nitrogenized layer.

Keywords Nonlinear vector regression · Optimization of metal working

22.1 Introduction

A classical view to mathematical modeling implies a descriptive approach characteristic of a physicist: the functions bound up with natural phenomena are subject to definite universal principles (laws), and the problem is to discover them. But the practice of descriptive sciences is different. The central conception sooner presumes that mathematical modeling consists in following the principle: the desired optimal model is simply the most exact model within the limits of a given admissible level of complexity or the least complex model, which approximates the (experimental) data observed with a precision up to a given admissible discoordination.

V. A. Rusanov (✉)
Institute for System Dynamics and Control Theory (ISDCT SB RAS), Lermontova str. 134, Irkutsk, Russia
e-mail: v.rusanov@mail.ru

S. V. Agafonov
Irkutsk State Agricultural Academy (ISAA), Baikalskay str. 257, Irkutsk, Russia
e-mail: agafonov@yandex.ru

A. V. Daneev · S. V. Lyamin
Irkutsk State Railway University (ISRU), Chernishevskogo str. 15, Irkutsk, Russia
e-mail: daneev@mail.ru

S. V. Lyamin
e-mail: slyamin@forus.ru

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_22,

The idea of formalization of considerations of model's complexity, which relates to the theory of identification of systems, was investigated in [1, 2]. From the viewpoint put forward by L. Ljung [3, 4], the idea that identification algorithms (by all means) have interpretation in the language of optimal approximation, is the main one. In the present paper have we essentially employed both of the indicated approaches, i.e. we have outline a combined methodology, which forms the ground of the procedure of optimal nonlinear approximation in the process of mathematical modeling of the process of nitrogenization of a mechanical part's surface to be processed under the conditions of effect of some inversive electrostatic field (with a non-stationary potential), within the frames of some linear-quadratic representation of vector regression equations.

22.2 Statement of the Problem of Synthesis of Optimal Multi-dimensional Regression

In principle, static models of the type "input–output" may be obtained from dynamic ones by applying experimental stationary finite values (or, what is equivalent, for the zero frequency). Unfortunately, the dynamic model is generally linearized, what is inadmissible for the static model, when this model is to be used for the purpose of optimization within a substantial band. Furthermore, the static model must be more detailed than the dynamic one (optimization, which improves the productivity by some 1 %, already represents a substantial interest from the application viewpoint), so, the structural-parametric identification of the multi-dimensional static nonlinear system of the type "input–output" in the absence of complete a priori understanding (knowledge) of the physics-mathematical principles of its functioning, a so called mathematical model of "black box", deserves an attentive deep consideration, especially when we have to ground the admissible level of complexity of the process under scrutiny.

From now on, R is the field of real numbers; R^n is an n-vector space over R (with the Euclidean norm denoted by $\|\cdot\| R^n$; $M_{n,m}(R)$ is the space of all the $n\times m$-matrices (i.e. the matrices of dimension $n\times m$) with the elements from R and with the Frobenius matrix norm $\| D \|_F := \left|(\sum d_{ij}^2)^{1/2}, \right| D = [d_{ij}]$ (what is equivalent to $D \in M_{n,m}(R) \Rightarrow \|D\|_F = (\mathrm{tr}.D^T D)^{1/2}$); as usually, the symbol: = denotes the equality by definition; det is a matrix determinant; tr $G := \sum g_{ii}$ is the trace of quadratic matrix G (the sum of its diagonal elements); "T" is the operation of transposition of a matrix; E_n is a unit $n\times n$-matrix; col $(a_1,\ldots,a_n)$ is a column vector with real elements $a_1,\ldots,a_n$.

A normal approach in the theory of identification of complex systems of the type of "input–output" methodologically consists in [5] *a priori* fixation of some partially parametrized class of stationary models and then, on the basis of fixed a posteriori data, to choose the parameters of the model's equations, which would minimize some formal criterion. In essence, this approach may be considered as application of the first method (denoted in the Introduction), in which "adjustment

of the model's parameters" (under a fixed number of free coefficients in its equations) is conducted. In this case, the criterion is defined by the model's complexity chosen a priori. So, for the purpose of further consideration, let us identify a class of stationary static interconnected nonlinear systems of the type "input–output", which are described by the vector-matrix regression equation of the form

$$y = c + Au + \mathrm{diag}[u^{\mathrm{T}} B_1 u,\ldots,u^{\mathrm{T}} B_n u]\mathrm{col}(1,\ldots,1) + \varepsilon(u); \qquad (22.1)$$

$y \in R^n$ is the vector of *system's output signals*, $u \in R^m$ is the vector of *system's assigning influences*, $c \in R^n$, $A \in M_{n,m}(R)$, $B_i \in M_{m,m}(R)$, $B_i^{\mathrm{T}} = B_i$ $(i = 1,\ldots,n)$ and diag […] is the *diagonal* $n \times n$-matrix of corresponding bi-linear controlling influences (effects) $u^{\mathrm{T}} B_i u$. As far as the vector function $\varepsilon(u)$ is concerned, we presume that the structure of its analytical representation is *a priori* unknown, but on the whole, it inexplicitly depends on the choice of the linear—$c + Au$ and bi-linear—diag $[u^{\mathrm{T}} B_1 u,\ldots,u^{\mathrm{T}} B_n u]$ col(1,…,1) components of the input signal—because the nonlinear component $\varepsilon(u)$ of Eq. (22.1) may always be considered as a residual ("under-modeled") term of the expansion of its right-hand side.

It is clear, the result y, predicted by the *linear-quadratic form* (LQF) $c + Au + \mathrm{diag}[u^{\mathrm{T}} B_1 u,\ldots,u^{\mathrm{T}} B_n u]\mathrm{col}(1,\ldots,1)$ of the right-hand side of Eq. (22.1), shall differ from the real signal, because the nonlinear law $\varepsilon(u)$ introduces some influence. On the other hand, as noted above, the analytical representation of the term $\varepsilon(u)$ depends on the choosing (fixation) of coefficients of the LQF. As a result, on the stage of identification, correction consists in varying the parameters of the LQF so that the results obtained, and those predicted on the basis of the LQF, would maximally coincide with each other. Obviously, new forecasts and parametric correction may then be conducted operatively (furthermore, additional information is used mainly for conducting partial or complete analysis of adequacy of the model on the basis of the latest current measurements). In other words, speaking more formally, the methodological paradigm of the a posteriori-optimal parametric synthesis of LQF shall provide for $\min \| \varepsilon(u) \|_R^n$ on the family of the representative sample of the field experiments conducted. When we proceed to the "language of formulas", this paradigm acquires the form of the following optimization problem.

S t a t e m e n t of the problem of a posteriori-optimal parametric synthesis of LQF for the equation of nonlinear regression: find a vector-matrix solution c, A, B_i, $i = 1,\ldots,n$ bi-criterion problem

$$\begin{cases} \min(\sum_{l=1,\ldots,k}(\| y(l)\text{-}c\text{-}Au(l)\text{-diag}\,[u^{\mathrm{T}}(l)B_1 u(l),\ldots,u^{\mathrm{T}}(l)B_n u(l)]\mathrm{col}(1,\ldots,1)\|_R^n)^2)^{1/2}, \\ min((\| c \|_R^n)^2 + (\| A \|_F)^2 + \sum_{i=1,\ldots,n}(\| B_i \|_F)^2)^{1/2}, \end{cases} \qquad (22.2)$$

where $y(l) \in R^n$, $u(l) \in R^m$ are vectors of experimental data (here $y(l)$ is the "reaction" to the input influence $u(l)$), k is the number of experiments completed; noteworthy, there are no methodological constraints imposed on the value of k.

R e m a r k 1. The first condition—min $\sum\ldots$ in the mathematical statement (22.2) guarantees—by the general sample of k field experiments—the optimal linear-quadratic approximation of the scrutinized physical process in terms of the nonlin-

ear regression model (22.1); the second condition—provides (in the case of *non-uniqueness* of the solution for the first min $\sum\ldots$) for parametric concretization of such a model with the property of the minimal matrix norm.

22.3 Parametric Identification of the LQF-Structure of Equations of Nonlinear Vector Regression

Let us relate the identification algorithm in the multi-criterion problem statement (22.2) for the interconnected stationary nonlinear system "input–output" of class (22.1) to the concept of *normal pseudo-solution* (or, what is equivalent, of canonical solutions by the method of least squares) for the system of linear algebraic equations.

Definition 1 [6, p. 501]. *Vector $x \in R^p$ is called the normal pseudo-solution of the system of linear equations*

$Dx = d$, $D \in M_{q,p}(R)$, $d \in R^q$. *This vector has the smallest Euclidean norm* $\| x \|_R^p$ *among all the vectors, which make minimum the value of* $\| Dx - d \|_R^q$.

Let $D \in M_{q,p}(R)$ and D^+ be the inverted reciprocal (pseudo-inverse) Moore-Penrose matrix [6, p. 500] for matrix D. The asymptotic construction of the pseudo-inverse matrix has the following analytical form:

$$D^+ = \lim\{D^{\mathrm{T}}(DD^{\mathrm{T}} + \tau E_q)^{-1} : \tau \to 0\}.$$

From now on, the mnemonic sign "$^+$" denotes the operation of pseudo-inverting of the respective matrix.

Lemma 1 [7, p. 35]. *Vector $x = D^+ d$ represents a normal pseudo-solution of the linear system $Dx = d$, $D \in M_{q,p}(R)$, $d \in R^q$.*

For the purpose of "interrelation" between the variables of input effects on the data of the general sample, let us denote by $\hat{u}(l)$ the $(1 + m(m+3)/2)$-vector, which has the following coordinate representation:

$$\begin{aligned}
&\hat{u}(l) := \mathrm{col}(1, u_1(l), \ldots, u_m(l), u_1(l)u_1(l), \ldots, u_r(l)u_s(l), \ldots, u_m(l)u_m(l)) \in R^{m(m+3)/2},\\
&1 \le r \le s \le m,\\
&\mathrm{col}(u_1(l), \ldots, u_m(l)) := u(l) \in R^m,\\
&1 \le l \le k.
\end{aligned} \tag{22.3}$$

Let us call $U := [\hat{u}(1), \ldots, \hat{u}(k)]^T \in M_{K,1+M(M+3)/2}(R)$ *the full matrix of experimental data related to input effects*, respectively, $\beta_i := \mathrm{col}(y_i(1), \ldots, y_i(k)) \in R^k$—*the fill vector of experimental data related to output signal* y_i $(i = 1, \ldots, n)$. Next, orienting to the linear-parametric description of the coefficients for the nonlinear model of the type "input–output" for the output signal y_i, let us write down—due to system (22.1)—the linear-quadratic form of its regression equation

$$c_i + \sum_{1\le j\le m} a_{ij}u_j + \sum_{1\le q\le p\le m} b_{iqp}u_q u_p,\ (i=1,\ldots,n). \tag{22.4}$$

Now introduce the $(1+m(m+3)/2)$-vector of regression model's parameters. Obviously, due to (22.4), any fixed set of n such vectors completely defines the representation of the LQF with respect to some "input–output" model of type (22.1):

$$z_i := \mathrm{col}(c_i, a_{i1},\ldots,a_{im}, b_{i11},\ldots,b_{iqp},\ldots,b_{imm}) \in R^{1+m(m+3)/2}, 1\le q\le p\le m. \tag{22.5}$$

Proposition 1. *The optimization problem* (22.2) *has the solution* (22.5)

$$z_i^* = U^+\beta_i,\quad i=1,\ldots,n;$$

here U is a complete matrix of experimental data related to input effects, β_i is the full vector of experimental data related to output signal y_i ($i=1,\ldots,n$).

Proof. According to relations (22.3) and (22.4), system (22.1) acquires the following compact form for each l-th experiment

$$y_i(l) = \hat{u}^{\mathrm{T}}(l)z_i + \varepsilon_i(l), i=1,\ldots,n. \tag{22.6}$$

Therefore, if the optimization problem of the form (22.2) is reformulated (obviously) in the vector-matrix terms z_i, β_i, U, we arrive at the following multi-criterion problem statement with regard to vectors z_i, $i=1,\ldots,n$:

$$\begin{cases}\min \| \beta_1 - Uz_1 \|_R^k,\\ \min \| z_1 \|_R^{1+m(m+3)/2},\end{cases}$$

$$\begin{cases}\min \| \beta_i - Uz_i \|_R^k,\\ \min \| z_i \|_R^{1+m(m+3)/2},\end{cases}$$

$$\begin{cases}\min \| \beta_n - Uz_n \|_R^k,\\ \min \| z_n \|_R^{1+m(m+3)/2}.\end{cases}$$

Obviously, due to Lemma 1, this multi-criterial system has a unique normal pseudosolution (22.5) with respect to z_i, $i=1,\ldots,n$.

Corollary 1. *Let $z_i^* = U^+\beta_i$, ($i=1,\ldots,n$), hence each vector z of parameters of LQF* (22.4) *is such that $z\ne z_i^*$, satisfies one of the two conditions*

$$a)\quad \| \beta_i - Uz \|_R^k > \| \beta_i - Uz_i^* \|_R^k$$

or

$$b)\quad \| \beta_i - Uz \|_R^k = \| \beta_i - Uz_i^* \|_R^k \ \textit{and}\ \| z \|_R^{1+m(m+3)/2} > \| z_i^* \|_R^{1+m(m+3)/2}.$$

R e m a r k 2. Qualitative estimates *a)*, *b)* from Corollary 1 depend mainly on the volume of a posteriori information (number of experiments k), i.e. if $k > 1+m(m+3)/2$,

then, as a rule, realized is item *a)*; if $k \le 1 + m(m+3)/2$ then it is quite probable that realized is item *b*).

22.4 Modeling of the Linear-Quadratic Structure of Equations of Vector Regression for the Process of Nitrogenization

Without any loss of generality, in the capacity of the initial (zero) position of the vector of input control influences u it is possible to accept some empirically identified (from the general set of experimental data) point ω of space R^m; obviously, in this case, coordinates $u_1,\ldots,u_m$ of vector u shall be considered as deviations with respect to the regime ω.

The process of nitrogenization in the environment of inversive electrostatic field in a series of field experiments ($k=12$) may be described in terms of the following variables:

vector $y = col(y_1, y_2, y_3) \in R^3$ of controlled characteristics of nitrogenization: y_1—Vickers surface hardness number 10^{-1} [HV],

y_2—specific wear 10^{-1} [mg/cm^2],

y_3—depth of the nitrogenized layer 10^2 [mm];

vector $u = col(u_1, u_2, u_3, u_4) \in R^4$ of variations of the regime's parameters $\omega = col(\omega_1, \omega_2, \omega_3, \omega_4)$:

u_1—variation (w.r.t. ω_1) of the degree of dissociation of ammonium 10^{-1} [%],

u_2—variation (w.r.t. ω_2) of the temperature of the process 10^{-1} [°C], u_3—variation (w.r.t. ω_3) of the duration of the process 10^{-1} [h],

u_4—variation (w.r.t. ω_4) of the voltage on the electrodes 10^{-3} [V].

Note, direct application of analytical methods developed above, results in not very complex but bulky computations (below the computation was conducted in the environment of MATLAB [8]); for example, according to Table 22.1, matrix U has the dimension of $k \times 1 + m(m+3)/2 = 12 \times 15$, and the matrix pseudo-inverse with respect to U^+ has, respectively, the dimension of 15×12.

U is the complete matrix of experimental data:

1	0	0	−1	0,4	0	0	0	0	0	0	0	1	−0,4	0,16
1	1	5	−1	0,4	1	5	−1	0,4	25	−5	2	1	−0,4	0,16
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	3,4	1	0	0	3,4	0	0	0	0	0	11,56
1	0	5	0	3,4	0	0	0	0	25	0	17	0	0	11,56
1	1	0	−1	3,8	1	0	−1	3,8	0	0	0	1	−3,8	14,44
1	0	5	−1	3,8	0	0	0	0	25	−5	19	1	−3,8	14,44
1	1	5	0	0	1	5	0	0	25	0	0	0	0	0
1	0,4	3	−0,4	0,18	0,16	1,2	−0,16	0,072	9	−1,2	0,54	0,16	−0,072	0,0324
1	0,3	3,5	−0,3	0,16	0,09	1,05	−0,09	0,048	12,25	−1,05	0,56	0,09	−0,048	0,0256
1	0,2	4	−0,2	0,14	0,04	0,8	−0,04	0,028	16	−0,8	0,56	0,04	−0,028	0,0196
1	0	5	0	0,1	0	0	0	0	25	0	0,5	0	0	0,01

Table 22.1 Experimental data of the process of obtaining the nitrogenized layer are (ω1 = 45 %, ω2 = 500 ºC, ω3 = 25 h, ω4 = − 1,900 V*)*

Experiment no	Assigning influences				Nitrogen layer parameters		
	u_1	u_2	u_3	u_4	y_1	y_2	y_3
1	0	0	−1	0,4	80,3	80,3	6,0
2	1	5	−1	0,4	93,3	93,3	3,4
3-ω	0	0	0	0	97,4	97,4	13,1
4	1	0	0	3,4	84,7	84,7	12,2
5	0	5	0	3,4	79,2	79,2	10,3
6	1	0	−1	3,8	54,8	54,8	42,4
7	0	5	−1	3,8	87,0	87,0	11,9
8	1	5	0	0	89,4	89,4	3,5
9	0,4	3	−0,4	0,18	87,0	87,0	4,2
10	0,3	3,5	−0,3	0,16	92,0	92,0	3,8
11	0,2	4	−0,2	0,14	98,8	98,8	4,0
12	0	5	0	0,1	89,4	89,4	3,6

U^+ is the matrix pseudo-inverse with respect to U:

0	0	1	0	0	0	0	0	0	0	0	0
0,2555	−0,26	−0,291	0,2335	0,2408	−0,231	0,2403	0,2397	5,8631	2,3343	−5,604	2,64
0,0641	−0,152	0,4173	−0,055	0,0398	0,0517	−0,04	−0,04	0,4082	0,3125	0,1758	−0,22
0,4685	−0,002	0,5228	−0,018	0,0204	0,0199	0,0203	0,0202	0,1598	0,0192	0,0999	−0,03
0,2684	0,045	0,2496	0,0522	0,05239	−0,02	−0,053	−0,052	0,7959	0,3416	0,79	−0,38
0,2422	−0,228	0,2058	0,2448	0,2491	−0,241	0,2486	0,248	5,7779	2,3993	5,6402	2,687
0,1	0,1	0,1	−0,1	0,1061	0,0943	−0,1	0,1	2,2982	0,9476	2,2036	−1,25
0,4601	−0,414	0,4467	0,4537	0,4237	−0,438	0,4193	0,4143	0,3385	0,1372	0,3271	−0,15
0,0997	0,2145	0,1332	−0,116	0,1908	0,1544	−0,202	−0,214	0,8461	0,343	0,8177	−0,38
0,0127	0,0299	0,0433	0,0118	0,0096	−0,01	0,0084	0,0084	0,0876	0,0627	0,0261	0,082
0,0804	−0,094	0,0774	−0,122	0,1282	0,1012	−0,105	0,094	0,2097	0,0838	0,2045	0,073
0,0491	0,0149	0,0564	−0,055	0,0705	0,0031	−0,012	−0,015	0,5244	0,2096	0,5112	0,183
0,4819	0,0335	0,4377	0,0298	0,0121	−0,031	−0,012	−0,012	0,0747	0,0842	0,1361	0,081
0,1418	0,1198	0,1545	0,1981	0,1234	−0,146	−0,121	−0,118	0,4192	0,1665	0,3999	−0.19
0,0066	−0,034	0,0093	0,0638	0,0296	0,0013	0,0327	0,0362	0,5241	0,2081	0,4998	−0,24

Taking into account the solution of the parametric optimization problem (22.4)–(22.6) and equation of the model of linear-quadratic vector regression (which describes in terms of a multi-dimensional polynomial approximation the interconnected process of nitrogenization in the environment of inversive electrostatic field, which possesses the variation of the potential due to the parametric representation of the vector structure U^+, and also, according to Table 22.1, of vectors β_i, $i = 1,\ldots,3$) have the form:

Table 22.2 Nonlinear regression model

Number in the forecast no	Forecast for the nonlinear regression model		
	$y_1(u)$	$y_2(u)$	$y_3(u)$
1	80,3	6,0	14
2	93,3	3,4	22
3	97,4	13,1	17
4	84,7	12,2	18
5	79,2	10,3	28
6	54,8	42,4	11
7	87	11,9	25
8	89,4	3,5	33
9	86,226	4,129	21,782
10	94,066	3,989	24,582
11	97,251	3,858	27,564
12	89,658	3,624	34,073

$$
\begin{aligned}
y_1(u) &= 97,4 - 65,075u_1 - 3,706u_2 + 9,369u_3 + 5,991u_4 - 64,313{u_1}^2 + 25,339u_1u_2 + \\
&+ 11,2136u_1u_3 + 7,159u_1u_4 + 0,529{u_2}^2 - 8,346u_2u_3 - 6,161u_2u_4 - 8,607{u_3}^2 + \\
&+ 6,29u_3u_4 + 6,227{u_4}^2; \\
y_2(u) &= 13,1 - 9,098u_1 - 2,232u_2 + 4,435u_3 - 2,235u_4 - 8,648{u_1}^2 + 3,531u_1u_2 - \\
&15,604u_1u_3 + 5,491u_1u_4 + 0,067{u_2}^2 + 2,361u_2u_3 + 0,502u_2u_4 - 3,986{u_3}^2 - \\
&- 5,336u_3u_4 + 0,5{u_4}^2; \\
y_3(u) &= 17 + 0,398u_1 + 0,964u_2 + 1,388u_3 - 0,437u_4 + 0,226{u_1}^2 - 0,424u_1u_2 + \\
&+ 5,264u_1u_3 - 1,84u_1u_4 + 0,507{u_2}^2 + 0,091u_2u_3 - 0,772u_2u_4 - 1,56{u_3}^2 - \\
&0,027u_3u_4 + 0,702{u_4}^2.
\end{aligned}
\tag{22.7}
$$

Critical analysis of the "predicted efficiency" of the proposed model intended for nonlinear mathematical description of the physics-chemical properties of the process of nitrogenization expressed in terms of *quasi-linear vector-matrix* regression equations (22.1), i.e. by the system (22.7), allows to conduct the relative comparison of the latter three columns of Table 22.1 with the following table obtained due to (22.7).

In the next section, we are going down to the multi-dimensional geometric investigation of "minimax" properties of solutions for the nonlinear vector regression, which describes electrostatic nitrogenization of the processed part surface, to the end of finding the regime of wear resistance and corrosion resistance for the geometry of its part. An interesting trait of the analytical results obtained is their explicit algebraic dependence on the parameters of system (22.7).

22.5 Interpolation of the Physics-Technological Characteristics of the Nitrogen Layer. Optimization of the Process of Nitrogenization

After all, the main objective of mathematical modeling is to answer the question "How it can the scrutinized physical process proceed and how it must proceed actually under some external controlling influence?". The answer to the second part of the question gives the solution of the optimization problem (22.9), while the answer to the first presumes the following fact:

Proposition 2. *The indicator of quality of nitrogenization* $J_i(u) := y_i(u)$ $(i = 1,\ldots,n)$ *may have the internal maximum or minimum in the identified LQF-structure of equations of nonlinear regression only at point* $u_i^* \in R^m$:

$$u_i^* = -B_i^{-1} A^T e_i / 2, \tag{22.8}$$

$\{e_1,\ldots,e_n\}$*—basis in* R^n. *Furthermore, when* $u^T B_i u$ *is a negative definite quadratic form, the indicator* $J_i(u)$ *has maximum at point* (22.8)*; when* $u^T B_i u$ *is a positive definite quadratic form, the indicator* $J_i(u)$ *has minimum at* u_i^*. *In the case, when* $u^T B_i u$ *assume both positive and negative values, we encounter the stationary point of more complex type, i.e. the so called saddle point*.

Proof. For the quality indicator $J_i(u)$ on the set of values of the linear-quadratic model (22.1), the necessary condition of local extremum is

$$\mathrm{col}(\partial(e_i^T Au + u^T B_i u)/\partial u_1,\ldots,\partial(e_i^T Au + u^T B_i u)/\partial u_n) = 0 \in R^n,$$

geometric coordinates (22.8) for the stationary point u_i^* with respect to the accepted functional of the quality indicator $J_i(u)$ are defined in the space R^m, while the sign-definiteness of the second differential

$$d^2 J_i(u) = \sum\nolimits_{1\le q\le m} \sum\nolimits_{1\le p\le m} \partial^2 J_i(u)/\partial u_q \partial u_p \,|_{u^*}\, u_q u_p$$

defines sufficient conditions of extremum in the stationary point u_i^*.

Corollary 2. *If matrix* B_i *is positive definite (similarly, negative definite) then the minimum (resp. maximum) value of the quality indicator* $J_i(u)$ *is*

$$c_i - e_i^T A B_i^{-1} A^T e_i / 4.$$

When turning back to the system of quadratic equations of synthesis of the physical structure of the surface ("embedment") content of the nitrogen layer (22.7), we obtain the numerical implementations of matrices A, B_i $(i = 1,\ldots,3)$:

should be used to present the results of investigations and large sets of figures clearly.

$$A = \begin{bmatrix} -65,075 & -3706 & 9,369 & 5,991 \\ -9,098 & -2,232 & 4,485 & -2,235 \\ 0,398 & 0,964 & 1,388 & -0,437 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} -64,313 & 12,67 & 5,607 & 3,579 \\ 12,67 & 0,529 & -4,173 & -3,095 \\ 5,607 & -4,173 & -8,607 & 3,145 \\ 3,579 & -3,095 & 3,145 & 6,227 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} -8,648 & 1,766 & -7,802 & 2,745 \\ 1,766 & 0,067 & 1,18 & 0,251 \\ -7,802 & 1,18 & -3,986 & -2,668 \\ 2,745 & 0,251 & -2,668 & 0,5 \end{bmatrix},$$

$$B_3 = \begin{bmatrix} 0,226 & -0,212 & 2,632 & -0,92 \\ -0,212 & 0,507 & 0,046 & -0,386 \\ 2,632 & 0,046 & -1,56 & -0,013 \\ -0,92 & -0,386 & -0,013 & 0,702 \end{bmatrix}.$$

Now we can solve the analytical problem, which has been the stimulus to investigation of positiveness (or negativity) of quadratic forms from equation (22.7), i.e. to answer the question—when the stationary point (22.8) is the point of relative minimum, maximum or the saddle point.

Speaking more formally, the problem of defining the positive (or negative) algebraic definiteness of the quadratic forms $u^{\mathrm{T}}B_i u$ has been reduced to the geometric problem of rather general type—computing of eigenvalues λ_{ij} ($i=1,\ldots,3$; $j=1,\ldots,4$) of symmetric matrices B_i ($i=1,\ldots,3$):

$\rightarrow\lambda_{11}=-67.5644$, $\lambda_{12}=-9.2743$, $\lambda_{13}=1.8251$, $\lambda_{14}=8.8491$, what speaks about the existence of a stationary saddle point for the goal functional $y_1(u)$: $R^4\rightarrow R$;

$\rightarrow\lambda_{21}=-14.7856$, $\lambda_{22}=-2.6697$, $\lambda_{23}=0.362$, $\lambda_{24}=5.0252$, what speaks about the existence of a stationary saddle point for the goal functional $y_2(u)$: $R^4\rightarrow R$;

$\rightarrow\lambda_{31}=-3.5248$, $\lambda_{32}=0.0847$, $\lambda_{33}=0.8665$, $\lambda_{34}=2.4482$, what speaks about the existence of a stationary saddle point for the goal functional $y_3(u)$: $R^4\rightarrow R$.

The graphic illustration of variations of quality indicators $J_i(u)$, $i=1,\ldots,3$ under a stationary temperature and duration of the process of nitrogenization depending on the scaled (according to data of Table 22.1) variations (with respect to the regime of nitrogenization ω) of the degree of dissociation of ammonium (Q 40 %) and voltage of the electrostatic field ($\pm$1000 V).

While combining previous results, the standard regime of nitrogenization, which provides for maximum hardness, wear resistance and the thickness of the physical structure of nitrogen layer of the processed surface of a mechanical part, let us relate them to the solution of the optimization problem of the following form

$$\max\{F(u) : u \in R^4\},$$
$$F(u) := r_1 J_1(u) + r_2 J_2(u) + r_3 J_3(u), \tag{22.9}$$

where the weighting coefficients r_i, $i=1,\ldots,3$ of the goal functional $F(u)$ must be chosen, while proceeding from the considerations of proper expert assessment of the differentiated effect of the quality indicators $J_i(u)$, $i=1,\ldots,3$ [9]. We have considered the following weighting coefficients: $r_1=0.5$, $r_2=-0.3$, $r_3=0.2$; the sign "–" with the coefficient r_2 means that the problem statement (22.9) actually provides for relative *minimization* (!) of the parameter of specific wear y_2 (what is equivalent, displacement to the point of min $J_2(u)$ in the linear structure of functional $F(u)$).

This allows us to write down the goal functional (22.9) in the following analytical form:

$$F(u) = 48{,}17 - 29{,}729u_1 - 0{,}99u_2 + 3{,}632u_3 + 3{,}579u_4 - 29{,}517{u_1}^2 +$$
$$+11{,}526u_1u_2 + 11{,}341u_1u_3 + 1{,}564u_1u_4 + 0{,}346{u_2}^2 - 4{,}863u_2u_3 -$$
$$-3{,}4u_2u_4 - 3{,}42{u_3}^2 + 4{,}74u_3u_4 + 3{,}1{u_4}^2. \tag{22.10}$$

The geometry of the six variants of the functional dependence for the quality indicator (22.10) depending on the coordinate variation of two identified components (while the other two are "frozen") of the 4-dimentional vector u in terms of deviations from the regime ω. In this case, parameters of the variations have constituted the following intervals (in terms of relative physics units):

$$u_1 = \pm 40\ \%,\ u_2 = \pm 50^\circ\mathrm{C},\quad u_3 = \pm 5\ \mathrm{h},\quad u_4 = \pm 1000\ V.$$

Development of new techniques of alloying metals necessitates existence of an adequate mathematical model, which would be capable of predicting the reciprocal influence of different factors of the physics-chemical environment on the process of metal working, as well as revealing the influence of mechanical and geometric characteristics of the processed part's surface upon the results obtained. As far as the multi-factor process of nitrogenization is concerned, the mathematical model of optimization (22.9) gives such a possibility, i.e. the possibility to reveal the most critical parameters and give the defining directions of improving the exploited and developed technological installations intended for obtaining the nitrogenized layer. Proposition 2, and also formula (22.8), which allow to compute the geometric coordinates of the stationary point for the optimization problem (22.9), define (in terms of system (22.1)) the following highly efficient technological parameters of the regime of nitrogenization:

Proposition 3. *The stationary point* $u^* \in R^4$ *in the problem related to optimization of the regime of electrostatic nitrogenization* (22.9) *has the algebraic solution*

$$u^* = -(r_1B_1 + r_2B_2 + r_3B_3)^{-1}((e_1 + e_2 + e_3)^{\mathrm{T}}\,\mathrm{diag}\,[r_1, r_2, r_3]A)^{\mathrm{T}} / 2, \tag{22.11}$$

in this case, the sufficient condition (*that the given point ensures satisfaction of* max $\{F(u)\colon u\in R^4\}$) is *the requirement that it is elliptic:*

$$\det[b_{ij}]_q < 0, \quad q = 1,\ldots,4 \tag{22.12}$$

or, what is equivalent, for the eigen-numbers λ_i *of matrix* $(r_1B_1+r_2B_2+r_3B_3)$ *we have* $\lambda_i<0$, $i=1,\ldots,4$; here $[b_{ij}]_q\in M_{q,q}(R)$ are the main sub-matrices [6, c. 30] of matrix $(r_1B_1+r_2B_2+r_3B_3)$.

22.6 Discussion

Let us start from the remark that if condition (22.12) is not satisfied the stationary point (22.11) is possibly the saddle (hyperbolic) point of functional $F(u)$ and, consequently, additional analysis of coordinates (22.11) is required; when speaking more formally, the availability of the saddle point is guaranteed by the replacement—at least in one relation (but not in all relations)—of the inequality "<" from (22.12) with ">". In this case, a similar replacement of "<" with "≤" possible provokes the structure of the parabolic point.

Due to system (22.1) (or, what is equivalent, due to equation (22.10)) the stationary point (22.11) in the coordinate representation (of the vector-row) writes

$$u^{*\mathrm{T}} = [0{,}1761 \quad 3{,}7794 - 0{,}5622 \quad 1{,}8787],$$

or, the same, in terms of physical dimensions and "counting" from the regime ω, we have:

$$u^{*\mathrm{T}} = [46{,}76\% \quad 537{,}794^{\mathrm{o}}\mathrm{C} \quad 19{,}378\ \mathrm{h} \quad -21{,}3\ \mathrm{V}].$$

Let us show that the mathematical result (in particular, the coordinates of the stationary point of the regime of nitrogenization (22.11)) obtained above are in good correspondence with the logic of our physics related reasoning. Since the eigen-numbers of matrix $(r_1B_1+r_2B_2+r_3B_3)$ are, respectively, $\lambda_1=-31.8762$, $\lambda_2=0.5298$, $\lambda_3=-3.276$, $\lambda_4=5.1355$, this gives evidence that functional $F(u)$ has a stationary saddle point: $R^4\to R$ for the weighting coefficients r_i, $i=1,\ldots,3$, chosen above.

According to (22.10), at the stationary point u^* obtained the functional $F(u)$ reaches its "max" with respect to variables u_1 and u_3 and "min", respectively, with respect to u_2 and u_4. The physical sense of this proposition implies the following: as far as the structure $F(u)$ is concerned, it is not possible to exceed (make larger) the degree of dissociation of ammonium by more than 46,76%, and the duration nitrogenization by more than 19,378 h, and, furthermore, in this case, simultaneously, it is better not to decrease the temperature of the gas mixture below the level of 537.794 °C, it is also better not to make the general potential of the electrostatic field smaller than 21.3 V. Otherwise, violation of these parameters shall provoke the reduction of efficiency of the process of nitrogenization in the aspect of reaching

the technological indicator $F(u)$, which provides for the maximum surface hardness and the depth of the nitrogenized layer side by side with minimization of specific wear of the part processed.

If computed (predicted) coordinates of the stationary point (22.11) go beyond the confidence region of adequacy of the mathematical model (22.7) in virtue of some physics-technological factors-parameters, than it is necessary to conduct an additional practical experiment bound up with nitrogenization, which is "maximally close" to the coordinates (22.11) and introduce (in the capacity of the regime ω) the data of this experiment into the extended matrix of experimental data U, after what it is possible to conduct recomputation of all the stages of optimization of the process of obtaining the nitrogenized layer, which is described above (if there is the need, such an experiment and the process of identification of model (22.1) are to be repeated); this important improvement, in essence, methodologically extends the standard [10] procedure of planning the experiment.

22.7 Conclusions

We have described the process of constructing a nonlinear mathematical model of the type "input–output" for the process of nitrogenization in the distributed environment of electrostatic field. This model is used for technological computation of hardness parameters for the material of the metal part, whose surface is processed. It can be used for assessment of the specific mechanical wear, the depth of the nitrogenized layer, etc. This regression model uses the identified (on the basis of experimental data obtained) multi-dimensional quadratic equations, what allows the researcher to adequately describe the process of nonlinear diffusion in the process of "nitrogen-alloying" within a wide band of variations of (I) the degree of dissociation of ammonium, (II) the temperature, (III) the duration of the process and (IV) the electric voltage at the pair "anode–cathode".

Deviations in the computed (predicted) values of the synthesized nitrogenized layer and experimental data revealed are hardly ever of principal character. This has given us the opportunity to propose an efficient mathematical technique ("a finite chain" of algebraic formulas) for computing optimal properties and parameters of nonlinear multi-factor regime of nitrogenization.

The ideas explicated in the present paper may be developed in several directions of theoretical-applied investigations oriented to improvement of the algorithms of computing an optimal technology of nitrogenization in an electrostatic field proposed above, as well as to extending the frames of adequacy of regression equations of nitrogenization at the expense of additional investigation of the factors of its nonlinearity; these can be oriented to:

determination and algorithmization of the procedure of choosing the weighting coefficients ri, $1 \le i \le 3$ in (22.9), while proceeding from satisfaction of the algebraic conditions (22.12), which provide for the elliptic character of the stationary point of the goal functional (22.9); extension of the linear-quadratic form of regression

equations (22.1) by the "Taylor expansion" of the vector function y of higher order; account (in the capacity of extended coordinates of the vector function y of the regression model) of such physics-mechanical parameters of the synthesized nitrogenized layer in the environment of some electrostatic field, such as the coefficient of dry friction for the surface processed and for the brittle nitrogenized layer obtained; constructing the process of identification of a nonlinear a posteriori–adaptive mathematical model of nitrogenization with an additional condition of presence of high-frequency electromagnetic field; determination (under such a problem statement) of high technological multi-factor parameters of the process of nitrogenization, and also obtaining optimal values for the length and the amplitude of the waves of electromagnetic oscillations.

Acknowledgement Thank you for your cooperation and contribution.

References

1. Caines P.E. On the scientific method and the foundation of system identification.—In: Modelling, Identification and Robust Control (Byrnes C.I., Lindquist A., eds.).—North Holland, Amsterdam, 1986, pp. 563–580
2. Rissanne J. Stochastic complexity and statistical inference.—Unpublished manuscript, I.B.M. Research K54/282, San Jose, California, 1985
3. Ljung L., Söderström T. Theory and Practice of Recursive Identification.—MIT Press, Cambridge, Massachusetts, 1983
4. Ljung L. A non-probabilistic framework for signal spectra.—In: Proc. 24th Conf. Decis. Control, Ft Lauderdale, Florida, December, 1985, pp. 1056–1060
5. Ljung L. System Identification—Theory for the User.—Moscow: Nauka. Publ., 1991 (in Russian)
6. Horn R.A., Johnson C.R. Matrix Analysis.—Moscow: Mir. Publ., 1989 (in Russian)
7. Gantmacher F.R. Theory of Matrices.—Moscow: Nauka Publ., 1988 (in Russian)
8. Andreyevsky B.R., Fradkov A.L. Elements of Mathematical Modeling in Programming Environments MATLAB and SCILAB.—St. Petersburg: Nauka Publ., 2001 (in Russian)
9. Makarov I.M., Vinogradskaya T.M., Rubchinsky A.A., Sokolov V.B. The Theory of Choosing a Decision and Decision Making.—Moscow: Nauka Publ., 1982 (in Russian)
10. Adler Yu.P., Markova E.V., Granovsky Yu.V. Planning Experiment in the Process of Finding Optimal Conditions.—Moscow: Nauka Publ., 1976 (in Russian)

Chapter 23
Self-Localization by Laser Scanner and GPS in Automated Surveys

V. Barrile and G. Bilotta

Abstract Our contribution is based on a research aimed to a "quick" resolution of an integrated problem oriented towards the self-localization and perimetration through mobile devices. The adopted methodology is applied on a real case study by using the following surveying tools: a kinematic Global Positioning System (GPS) and a Laser Scanner supporting a "mobile platform". A GPS receiver provided by Leica Geosystem and a two-dimensional Laser Scanner provided by the Automation and Control Laboratory of the University "Mediteranea" of Reggio Calabria were positioned on an experimental mobile system specifically designed to simulate the behaviour of a future and fully automated platform. The research is aimed to conduct the traditional land surveying through a Laser Scanner alongside with GPS receivers in a three dimensional centimetric resolution within one single system of reference made up of individual scans operated by a "Stop-and-Go" device.

Keywords Laser scanner · Self-localization · GPS · Survey

23.1 Introduction—The Experimental Campaign

As part of a collaboration between the Geomatics Lab and the Automation and Controls Laboratory of the Mediterranean University of Reggio Calabria, aimed to the possible development and implementation of an algorithm based on the use of a laser-scanner sensor for applications mobile robotics, we carried out a first experiment in the yard behind the university (Fig. 23.1).

This experiment was aimed to an automated kinematic perimetering of the area under investigation with simultaneous auto-location detection sensor through the integration of laser scanner and GPS measurements.

V. Barrile (✉)
DICEAM Department, Mediterranean University of Reggio Calabria,
via Graziella Feo di Vito, 89100 Reggio Calabria, Italy
e-mail: vincenzo.barrile@unirc.it

G. Bilotta
Ph.D. NT&ITA, Planning Department, IUAV University of Venice,
Santa Croce 191 Tolentini, 30135 Venice, Italy
e-mail: giuliana.bilotta@gmail.com

N. Mastorakis, V. Mladenov (eds.), *Computational Problems in Engineering,*
Lecture Notes in Electrical Engineering 307, DOI 10.1007/978-3-319-03967-1_23,

Fig. 23.1 Survey area behind the university building

Fig. 23.2 Mobile platform

In particular, we used a rudimentary "moving platform" (trolley mobile), equipped with a laser-scanner (which currently allows to perform scans only within the planimetric) mounted on a trolley with wheels (Fig. 23.2); on the same carriage, above the laser sensor, was placed the GPS receiver (Fig. 23.3).

The sensor is connected to the USB port of a laptop that sends to the LRF instructions to be executed through the use of the programming language Matlab (programming language used for all the algorithms implemented for the management and implementation of the system).

It should be noted preliminarily that the automation of the procedure is not yet currently available and that today the operations are carried out manually.

In particular, there has been a 360° rotation of the basket by making the acquisitions at regular intervals of time trying to ensure the continuity of motion, simulating a behavior as much as possible regular.

Before of the integration operations between the different survey methods, was independently carried out a perimeter of the study area through GPS survey in classic mode Real Time Kinematic; processing of the acquired data performed with the commercial program of the Leica LGO allowed to obtain the coordinates of the

Fig. 23.3 Survey operations

robot.txt

1011	558340.830	4219463.856	71.060
1010	558351.745	4219464.071	70.979
1009	558353.802	4219462.929	71.011
1008	558353.486	4219460.726	71.041
1007	558348.486	4219448.349	71.161
1006	558357.896	4219447.128	71.199
1005	558362.532	4219455.262	71.131
1004	558364.995	4219463.866	71.055
1003	558360.563	4219463.981	71.065
1002	558362.141	4219464.797	71.061
1001	558361.499	4219467.919	71.014

Fig. 23.4 GPS data

points shown in the diagram of Fig. 23.4 representing the perimeter of the study area.

The same data were subsequently reported on georeferenced map; these data, connected each other, allowed therefore to delimit the perimeter of interest (Fig. 23.5).

These data are considered as data "reliable" to be used for comparison with the survey methods later proposed. In particular, it has been positioned in this regard (integrated laser scanner—GPS—mobile cart) on the platform above the laser scanner sensor, a GPS antenna (Fig. 23.6) in such a way to obtain simultaneous measurements [1–3].

23.2 Measurement by Laser Scanner

We made seven scans with the "equipped mobile trolley", manually moving it (360°), with a view to its future and complete automation [4, 5].

Scans are shown below (Fig. 23.7).

Fig. 23.5 GPS data in map

Fig. 23.6 Simultaneous positioning laser scanner and GPS on "mobile equipped trolley"

For each scan was carried out at the same position detected by GPS measurements useful for linking the different scans through the measurement of external targets.

Single scans were processed and linked together by means of an algorithm implemented in Matlab (lab AeC), in the testing phase and the subsequent development by deriving a series of segments that describe, with little margin for error, the geometry of the square (Fig. 23.8).

23.3 The Algorithm

The algorithm implemented in Matlab and used in this experiment does not use the common return target detected externally but makes a connection of several scans through statistical autocorrelation methods by using the distinctive features that the

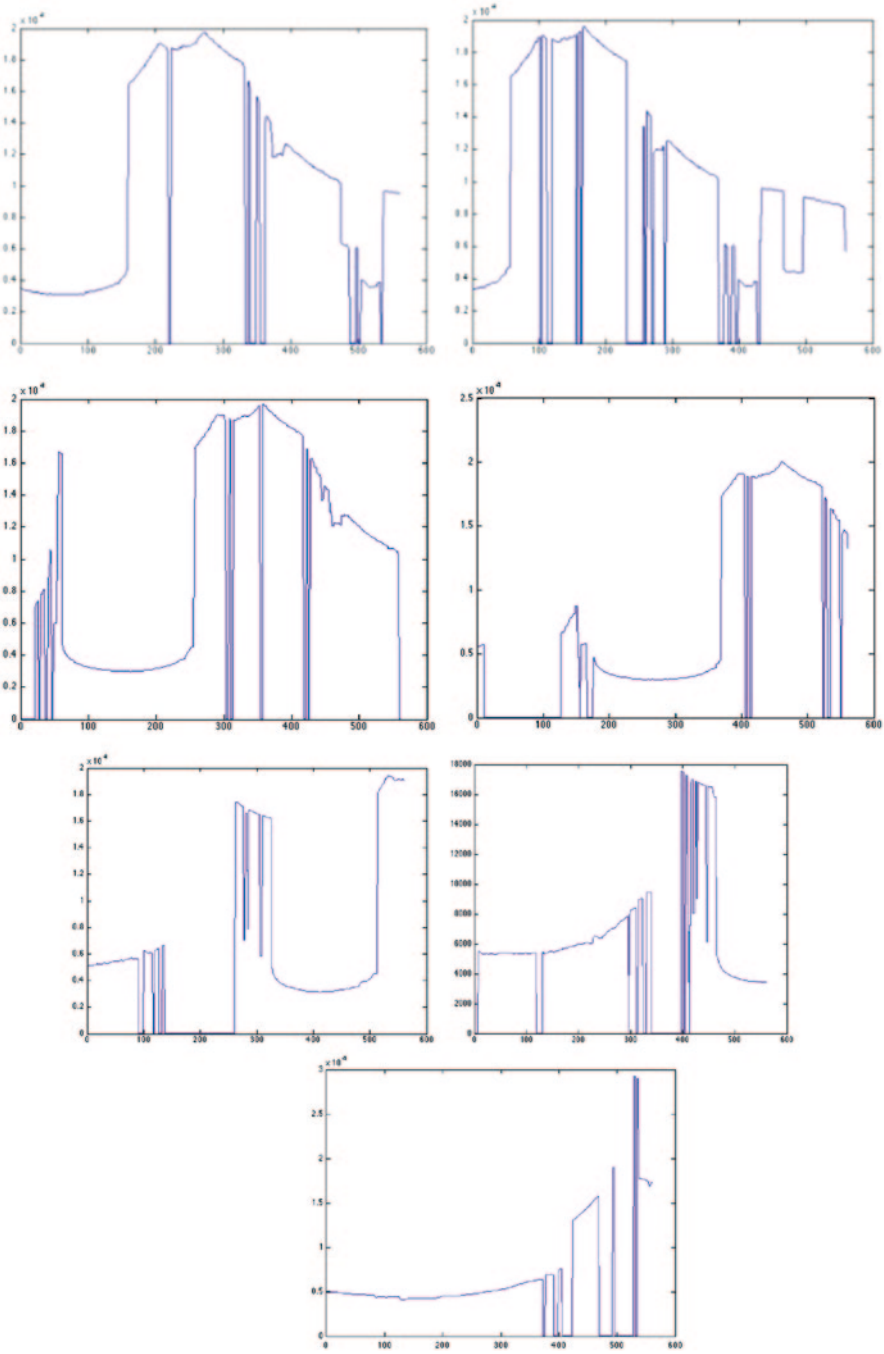


Fig. 23.7 Laser scanner scans nn. 1–7

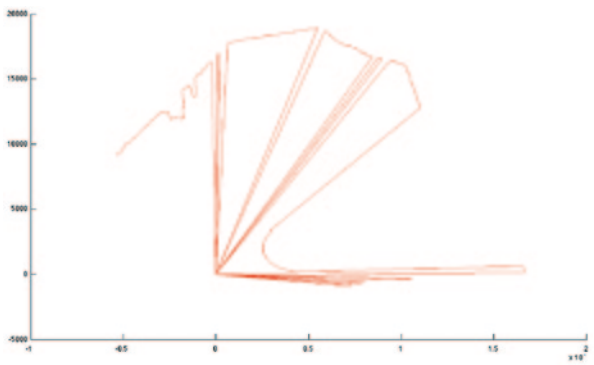


Fig. 23.8 Result of localization algorithm

robot (mobile equipped trolley) is able to perceive the environment through the use of the laser scanner sensor [6–8].

These characteristics may be the geometric shapes, such as edges, circles or rectangles, or additional data such as barcodes. The features must have a precise and fixed position within the environment and should be easily detectable by the sensor [9].

The methodology used can be divided into two phases: extraction of features from the measurements made by the sensors; coupling between features belonging to different measures so as to determine the deviation between the two measures in terms of a shift (Dx, Dy) and a rotation Dα.

We thus applied an algorithm of "SLAM" [10–12] based exclusively on information from a laser scanner. This algorithm introduces a new model for the prediction of the future state (described in Fig. 23.9).

The methods of location-based laser odometry differs depending on what data are used to search the correspondence between scans. The algorithm that will be described below is based on matching through the use of features and is shown schematically in Fig. 23.10. The implementation of the algorithm created for testing builds on the following general considerations (Feature extraction—Matching between features—Optimization Process) performing particularly well-known in the literature in closed environments and adapted in the present work for a trial in open environments.

Fig. 23.9 Prediction model of future state

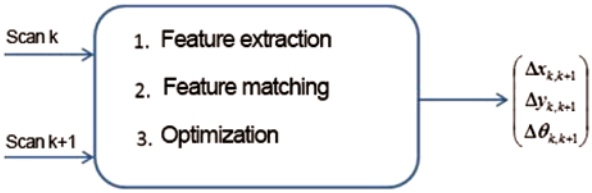


Fig. 23.10 Diagram of localization algorithm based on use of features

From the knowledge of the current pose of the robot, xk, its covariance, cov(xk), the extracted features to scan the k-th and k+1-th scan and the covariance associated with the features you want to calculate the pose of the robot to the next step, xk+1, and its covariance, cov(xk+1). To do this you must perform three steps:

- Extraction of set of features belonging to scan S_1 and of set of features F_2 belonging to the scan S_2 subsequent respect to S_1;
- Matching between features of the two scans that will be a subset of those extracted;
- Optimization process: calculation of the deviation between the two scans through the calculation of the transformation excellent in terms of rotational translation.

23.3.1 Feature Extraction

As is known in the literature the matching techniques through the use of features presuppose a preliminary phase concerning the extraction of features from the scan. The features are divided into two types: "jump-edges" and "corners".

To detect the features jump-edges, a scan is divided into groups (called "clusters") of consecutive scan points. In this way, each cluster consists of a starting point, p_i, and an end point, p_j, and the k-th cluster is defined in the following way:

$$c_k = \{p_m \mid p_m \in S, i \le m \le j\} \tag{23.1}$$

The start and end points of each cluster are candidates to become features jump-edges as long as these points are invariant with respect to the movement of the robot.

To extract the features "corners" within a scan is instead necessary to extract lines from each cluster using an algorithm such as "split-and-merge".

Each line extracted is characterized by the following parameters: $l_q=[\alpha_q, n_q, len_q]$, where α_q is the angle between the line and the x-axis; n_q is the number of points that constitute the line and len_q is the length of l_q.

If the intersection of two successive lines is such that $|\alpha_{q+1}- \alpha_q|>\Delta\alpha_{th}$ and that, for both for l_q that l_{q+1}, or $len>len_{th}$ or $n_q>n_{th}$ (where len_{th} is the minimum length and n_{th} is the minimum number of points of the lines that make up the corner) then p_{cc}, which is the end point of l_q, is a candidate to become a feature corner.

A problem that could cause a wrong operation of the algorithm of localization is the presence of a small number of features between scans.

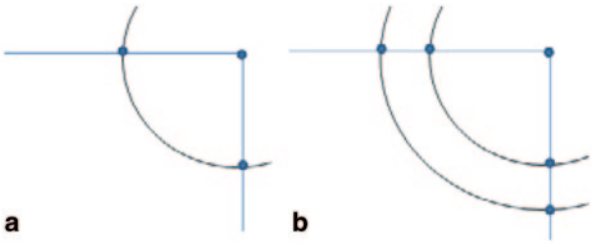


Fig. 23.11 Enrichment of the number of features through the use of one (a) and two (b) circumferences

Consider the only feature that is extracted from a corner of the environment. Starting from this feature are created two other features by making of the compass opening centered in the feature to start. The two new features are the intersections of this circle with two lines that form the angle (Fig. 23.11a). A further enrichment of the features can be carried out by considering two circles centered in the feature starting having two different radii (Fig. 23.11b). In this case the number of features extracted is quintupled.

23.3.2 Matching Between Features

Once extracted, by two successive scans, the features that represent the same physical point of the environment, is necessary to couple. We use a matching algoritm well known in the literature and adapted to the particular testing in open environments, which is based on a function of dissimilarity, *d*. We define this function for two points p_i and p_j, belonging to two successive scans:

$$d(p_i, p'_j) = \left\| p_i, p'_j \right\|_2 + B \tag{23.2}$$

If $|\alpha_{next} - \alpha'_{next\,j}|$ o $|\alpha_{pre,\,i} - \alpha'_{pre,\,j}|$ exceeds a certain threshold, p_i and p_j are not coupled and B becomes equal to infinity, otherwise B is equal to zero. Once constructed the matrix containing all the functions of dissimilarity (called dissimilarity matrix), the smallest value of this matrix is eliminated and the corresponding features are coupled. This is done at each step, until all the elements of the matrix are eliminated or until the remaining elements have a value above a certain threshold.

23.3.3 Optimization Process

Particularly useful for optimization process is the implementation of the various steps below, realized, under updating and further experimentation that take their cues from what is known in the literature and shown below. To calculate the new installation of the robot in an optimal way is necessary to establish a model of uncertainty for the features extracted, i.e. a model which takes account of errors (such as the noise of the measurement process, e_{ob}, and the quantized nature of the

angles of the rays, e_q); it could cause that the actual position of the feature differs from that calculated.

The position of the k-th feature must therefore be written as:

$$f_k = p_i + e_{ob_i} + e_{qi} \tag{23.3}$$

And the expected value of the position of features, f_k, is given by:

$$\hat{f}_k = E(f_k) = p_i + E(\mathrm{e}_{\mathrm{ob}}) + E(e_q) \tag{23.4}$$

where E(·) is the expected value operator.

At this point it is necessary to calculate the covariance of f_k:

$$Cov(\hat{f}_k) = E\left(\left(f_k - \hat{f}_k\right)^T . \left(f_k - \hat{f}_k\right)\right) = Cov(\tilde{e}_{ob_i}) + Cov(\tilde{e}_{q_i}) \tag{23.5}$$

Using the measurement of the features and their corresponding covariance, the algorithm calculates the displacement (defined in terms of translation, T, and rotation, R) effected by the robot between the two scans. To find the optimal values of T and R the following error function must be minimized:

$$E = \sum_{j=1}^{m} (\hat{f}_{j,pre} - (R\hat{f}_{j,new} + T))^{\mathrm{t}} C_j^{-1} (\hat{f}_{j,pre} - (R\hat{f}_{j,new} + T)) \tag{23.6}$$

Where m is the number of features coupled, $f_{j,\,pre}$ and $f_{j,\,new}$ are two new features coupled refer respectively to the previous scan and the current one; $v_j = (f_{j.pre} - (Rf_{j,\,new} + T))$ is the j-th vector innovation and C_j is its covariance.

Assuming that the errors in the scans are independent, we can write:

$$C_j = \mathrm{cov}(e_{ob}^{pre} + e_q^{pre}) + R\,\mathrm{cov}(e_{ob}^{new} + e_q^{new})R^t \tag{23.7}$$

There is the possibility of writing the variables in vector form. In such form, the displacement of the robot can be indicated in the following way:

$$X = (q_1 \quad q_2 \quad t_1 \quad t_2)t \tag{23.8}$$

Where t_1 and t_2 are respectively the translations along the x direction and the y direction. The rotation R and translation T matrices become defined as follows:

$$R = \begin{pmatrix} q_1^2 - q_2^2 & -2q_1q_2 \\ 2q_1q_2 & q_1^2 - q_2^2 \end{pmatrix} \qquad T = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \tag{23.9}$$

The optimization problem is solved using the SQP method, “Sequential Quadratic Programming”.

Assuming that X* is the deviation between the two scans that minimizes the function E described above, for calculating the covariance of X* should exist Jacobian, J, projecting the uncertainty of the features in the uncertainty of X*. If there is an explicit function, g, which relates X* to F, which is the vector of all the features coupled, we have X* = g(F). The Taylor series expansion of g in the neighborhood of E(F) will be:

$$X^* = g(\hat{F}) + (F - \hat{F})\frac{\partial X^*}{\partial F} + O(F - \hat{F})^2 \tag{23.10}$$

The last summand represents the higher order terms.

The Jacobian between X* and F projects the uncertainty of X* in F, namely:

$$\text{cov}(X^*) = J\,\text{cov}(F)J^T \tag{23.11}$$

However, there is an explicit relationship between F and X*, then they are related by an implicit function, I(X*,F) = 0, which is derived from ∂E/∂X = 0. You can obtain this Jacobian using the equation:

$$J = -\left(\frac{\partial \gamma}{\partial X^*}\right)^{-1}\left(\frac{\partial \gamma}{\partial F}\right) \Rightarrow J = -\left(\frac{\partial^2 E}{\partial X^2}\right)^{-1}\left(\frac{\partial^2 E}{\partial F \partial X}\right) \tag{23.12}$$

with X = X*.

Conducting additional steps and substitutions you can get the desired Jacobian matrix. The independence of the features of a scan brings to obtain a total diagonal covariance matrix.

Furthermore, assuming that the features extracted by two successive scans are independent, the covariance of each pair will be:

$$\text{cov}(F) = \begin{bmatrix} \text{cov}(F_1) & 0 \\ 0 & \text{cov}(F_m) \end{bmatrix}, \text{cov}\left(F_j\right) = \begin{bmatrix} \text{cov}(f_{j,new}) & 0 \\ 0 & \text{cov}(f_{j,pre}) \end{bmatrix} \tag{23.13}$$

Substituting the expressions of J and cov(F) we get cov(X*), i.e. the uncertainty of the deviation.

The pose of the robot at the generic instant k can be defined as:

$$x_{r,k} = [x_{r_1 k}, x_{r_2 k}, x_{r_3,k}] \tag{23.14}$$

where $x_{r1,k}$ e $x_{r2,k}$ represent the translations along the x and y axes, while $x_{r3,k}$ is the orientation. The new robot pose is then determined by the equation:

$$\hat{x}_{r,k+1} = \hat{x}_{r,k} + \begin{pmatrix} \cos\hat{\theta}_k & -\sin\hat{\theta}_k & 0 \\ \sin\hat{\theta}_k & \cos\hat{\theta}_k & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} t_1 \\ t_2 \\ \Delta\theta \end{pmatrix}, \hat{\theta}_k = \hat{x}_{r_3,k} \tag{23.15}$$

where t_1 and t_2 are the translations along the x and y between the instants k and k+1 while $\Delta\theta$ is the rotation in the same time interval.

The calculation of the covariance of $X_{r,\,k+1}$ requires its differentiation with respect to the random parameters of the right side of the equation just written:

$$J_P = \frac{\partial x_{r,k+1}}{\partial(x_{r,k}, q_1, q_2, t_1, t_2)} \tag{23.16}$$

Given the independence between $x_{r,\,k}$ and X, the covariance of the parameters of the right side of the equation $x_{r,\,k+1}$ will be:

$$p' = \begin{pmatrix} \mathrm{cov}(p_k) & 0_{3x4} \\ 0_{4x3} & \mathrm{cov}(X^*) \end{pmatrix} \tag{23.17}$$

The covariance of $x_{r,\,k+1}$ can be calculated in the following way:

$$\mathrm{cov}(x_{r,k+1}) = J_P P' J_p^t \tag{23.18}$$

The state vector x_k of the system is composed of the state of the robot, $x_{r,\,k}$, and the state of all the features, $x_{f,\,k}$. The state vector and its covariance before the prediction will be:

$$\hat{x}_k = \begin{pmatrix} \hat{x}_{r,k} \\ \hat{x}_{f,k} \end{pmatrix}, \mathrm{cov}(x_k) = \begin{pmatrix} \mathrm{cov}(x_{r,k}) & \mathrm{cov}(x_{r,k}, x_{f,k}) \\ \mathrm{cov}(x_{f,k}, x_{r,k}) & \mathrm{cov}(x_{f,k}) \end{pmatrix} \tag{23.19}$$

The movement of the robot does not affect the status of the features, so we have:

$$\hat{x}_{k+1} = \begin{pmatrix} \hat{x}_{r,k+1} \\ \hat{x}_{f,k} \end{pmatrix}, \mathrm{cov}(x_{k+1}) = \begin{pmatrix} \mathrm{cov}(x_{r,k+1}) & \mathrm{cov}(x_{r,k}, x_{f,k}) J_P'^T \\ J_P' \,\mathrm{cov}(x_{f,k}, x_{r,k}) & \mathrm{cov}(x_{f,k}) \end{pmatrix} \tag{23.20}$$

J'_p is the truncated form of J_p and includes only the differentiation of $x_{r,\,k+1}$ with respect to $x_{r,\,k}$.

The next step is the association of the data and update the map. For data binding, the positions of features belonging to the map must be predicted relative to the robot. This is done by a model of the observation that, for the i-th feature, is:

$$f_i^r = h_i(x_r, f_i^{map}) \tag{23.21}$$

The superscript *r* and *map* refer respectively to the coordinates of the robot and global ones.

In the present case, the model h_i is:

$$\begin{pmatrix} f_{i1}^r \\ f_{i2}^r \end{pmatrix} = \begin{pmatrix} \sqrt{(f_{i1}^{map} - x_{r1})^2 (f_{i2}^{map} - x_{r2})^2} \\ a\tan\left(\dfrac{f_{i2}^{map}}{f_{i1}^{map}}\right) - \theta_r \end{pmatrix} \tag{23.22}$$

The total observation model, h, is obtained considering all the features in a single vector. The features that are not coupled with any feature in the map are added to the latter through data binding. The features which would be coupled with map features create new relationships between persistent objects in the map. In this case the state vector of the system and the covariance matrix do not increase in size, but are updated [11, 13].

The obtaining of information from sensors in the current scan is described by a function of measurement:

$$\begin{aligned} \hat{F}^r &= h(\hat{x}_r, \hat{F}^{map}) = h(\hat{x}) \\ \mathrm{cov}(F^r) &= H_x \,\mathrm{cov}(x) H_x \end{aligned} \tag{23.23}$$

where:

$$H_x = \left. \frac{\partial h(x)}{\partial x} \right|_{x=\hat{x}_{k+1}^{(-)}}$$

The models of the process and observation are not linear, so the noise variables are assumed to be taken from normal distributions.

For the filtering step is chosen the iterated extended Kalman filter (IEKF), ie:

$$\begin{aligned} \hat{x}_{k+1,i+1}^{(+)} &= \hat{x}_{k+1}^{(-)} + K_{k+1,i}[F^{k+1} - (h(\hat{x}_{k+1,\mathrm{i}}^{(+)}) + H_x(\hat{x}_{k+1}^{(-)} - \hat{x}_{k+1,\mathrm{i}}^{(+)}))] \\ \mathrm{cov}(\hat{x}_{k+1,i+1}^{(+)}) &= \mathrm{cov}(\hat{x}_{k+1}^{(-)}) - K_{k+1,i} H_x \,\mathrm{cov}(x_{k+1}^{(-)}) \\ K_{k+1,i} &= \mathrm{cov}(x_{k+1}^{(-)}) H_x^T [H_x \,\mathrm{cov}(x_{k+1}^{(-)}) H_x^T + \mathrm{cov}(F^{K+1})]^{-1} \end{aligned} \tag{23.24}$$

where:

$$H_x = \left. \frac{\partial h(x)}{\partial x} \right|_{\hat{x}_{k+1,i}^{(-)}}, \hat{x}_{k+1,0}^{(+)} = \hat{x}_{k+1}^{(-)}$$

23.3.4 Checks and Comparison

The algorithm, implemented on the basis of the above points and adapted to the specific experiment in open environments, has allowed us to analyze the first four scans, while the last three we had difficulties due to external phenomena of noise that prevented proper data collection.

In any case, after a "cleaning" of the data from any nuisance parameters (GPS and laser scanner), overlaying the drawing of the survey to cartography is obtained as shown in Fig. 23.12.

To check the validity of these results, we do a comparison between the results Laser Scanner and those GPS (red line on maps considered as "certain"), preferring

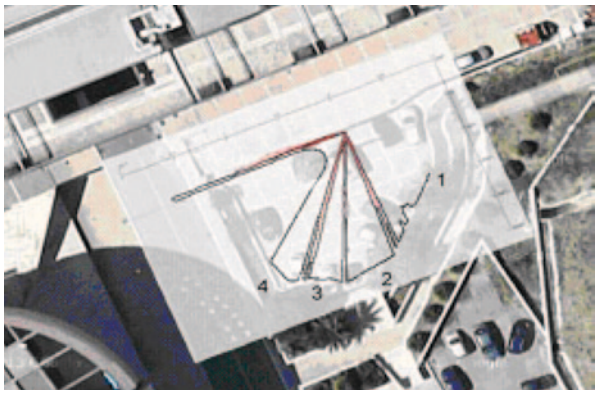


Fig. 23.12 Overlaying result of algorithm on mapping

Fig. 23.13 Comparison between the two methods

the graphic display able to better show the differences between the two methods (Fig. 23.13) rather than the creation of complex tables and graphs summarizing and/or various statistical parameters on the accuracy of the processing, because the aim of "expeditious" of this proposal.

Although there is the same precision of the GPS data in terms of return, however, is highlighted as the algorithm proposed for the processing of the given laser scanner is able to provide by itself discrete results, as evidenced by the partial planimetric correspondence of the two tracks GPS and Laser Scanner shown in Fig. 23.13.

This is a good omen for the continuation of the trial.

23.4 Integration of GPS and Laser Scanner for Connecting Subsequent Scans

As known, the main problem for laser scanner data is the assembly of the scans in order to determine a unique reference system in which "immerse" the obtained model [14–17]. The acquisition of the scans results in an immediate point cloud

ordered in the plane, whose coordinates are known with respect to the center of "taking".

The scan is then locally oriented with respect to a reference system that derives from the arbitrary choice of the pickup point, which will be taken as the origin of the reference system of the scan. The assembly of multiple scans thus requires the knowledge of the parameters of rototranslation: these parameters can be calculated if the position of the origin of the reference system of each scan with respect to a single system is known through the measurement of the external "target". Such a problem for geo—topographic applications is solved by having remarkable points (targets), of which the coordinates are known, in all the scans: in this way each scan can be oriented independently of the other. Their georeferencing can be done by using the techniques of GPS tracking [18].

From the above considerations, the idea of experimenting with a rudimentary expeditious survey able to repeat what has already been experienced with the vehicle fully equipped (equipment includes two GPS, a laser scanner and a target all mounted on a vehicle in motion) that, by combining the two receivers GPS with the sensor laser scanner and a target audience, can overcome the issues raised; the whole mounted on a moving body that allows easy movement between the measurement sessions.

By performing measurements laser scanner and GPS simultaneously with stationary body is thus ensured a high quality of fit and positioning into a single reference system.

The system is to mount on movable equipped trolley (rigidly and coaxially) the laser scanner surmounted by a GPS and connect the trolley through a rigid arm (adjustable in length) to a "target" coaxially surmounted by other GPS reference (which will serve as the orientation of the scan), left free to rotate anyway so as to guide the laser target to the sensor. In this way, the problem of defining the coordinates of the acquisition point (Laser Scanner) and target orientation is overcome by fitting precisely coaxially two GPS receivers, respectively, the Laser Scanner and the target [19, 20].

The receivers, while the laser sensor scans, acquire measurements from GNSS satellite constellations providing coordinates, both geographic both local coordinates of the laser sensor and the target orientation into a single reference system.

Once we have defined the ideal location for the first scan, we must place and stop the mobile equipped trolley at the point defined by performing both those measures GPS and Laser Scanner with the characteristics of density required by the survey. After a few minutes we must close the measures and shall move the trolley equipped cabinet in the next position chosen for the second major station, operating as before and repeating the process until completion of the survey. The processing of GPS data will allow to obtain homogeneous coordinates for all points of outlet (station laser scanner) and for all orientation target with sub-centimeter accuracy. These coordinates are assigned to stations and targets thereby allowing the software used for the management of the scans to unite and georeference all the scans made even in the absence of homologous points or targets positioned on the ground [21].

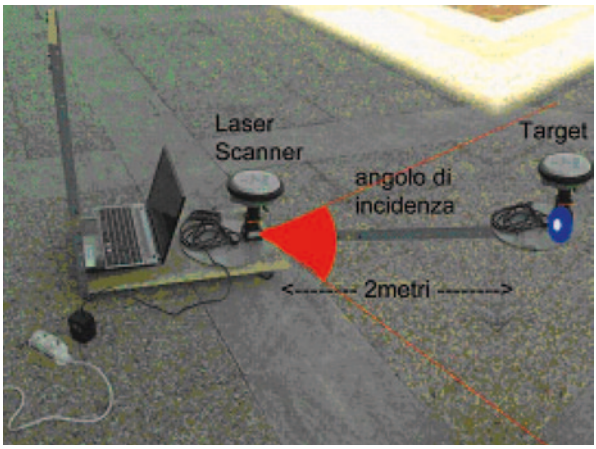


Fig. 23.14 System with target and dual GPS

Arm (distance between the coaxial position of the laser scanner and that of the target) of:	Deviations "average" positioning expressed in percentage
3 meters	11%
2 meters	19%
1.5 meters	37%
1 meters	Determinabile only occasionally with higher % than 65%
0.5 meters	Location not determined

Fig. 23.15 Variation of the percentage error compared to GPS method (in the test simulated) by varying the arm in question

In this way, in addition to speed up and facilitate the steps of the survey in the field by eliminating the need of affixed targets and the necessity of their internal visibility between a measurement session and the other, will be easier georeferencing also individual scans with no points in common, decreasing processing time of "point clouds" resulting from the scans.

Taking into account what was said above, namely we have tried to make an initial experimentation in order to achieve "coarse" and "expeditious" what has already been experimented on equipped machine (cf. Leica experiment reported in bibliography).

Specifically, it was built by placing a measuring system on the mobile trolley equipped (rigidly and coaxially) the laser scanner superimposed by a GPS and connecting the trolley through a rigid arm (simulating the modulation length through the ability to extend and contract) to a "target" coaxially superimposed by other GPS reference.

In particular, measures have been simulated with arms of 3, 2, 1.50, 1, 0.5 m (Fig. 23.14)

The overall reconstruction of the data, although simulated, is very interesting in particular for the test carried out with the arms of 3 and 2 m (note in this regard the result of the perimeter displayed in color and overlaid on the map as reported in Fig. 23.16). Instead, less accurate appear the results obtained with simulated arm of 1.5 m, while it was not possible to make reliable reconstructions with simulated arm of 1 m or less. (Fig. 23.15).

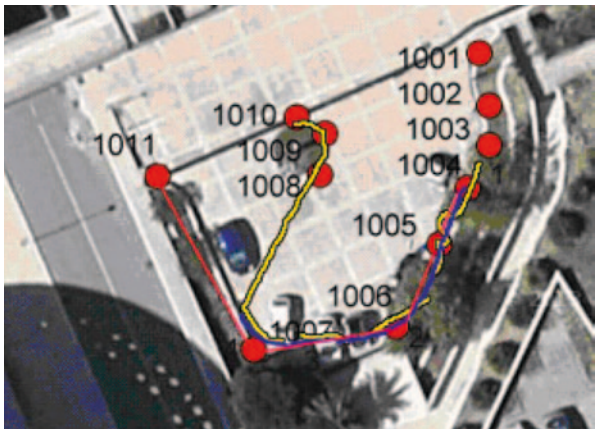


Fig. 23.16 Integration of the two different methodological approaches

The algorithm we used is then enriched in such a way as to be able to solve a problem of mapping.

For all the scans in input will be repeated the following steps:

- Initialization;
- Calculation of the i-th pose;
- Updating the Map.

The initialization is carried out between a number of scans by Laser Scanner defining the initial pose, related to the first scan performed, which is taken as reference pose. In this phase, are initialized the vectors x_{tmp}, y_{tmp} and θ_{tmp} that will contain the offset of the i-th scan with respect to the initial reference pose.

Thereafter, for the calculation of the i-th pose is essential the algorithm chosen to solve the problem of the localization that is used for the calculation of all deviations.

Finally, to update the map of the environment is sufficient to plot the i-th scan (of which we know the absolute pose) in the reference system of the first scan.

As known, the drift phenomenon that affects the localization consists in the fact that the errors made in the estimation of the robot pose tend to be additive in time. This means that, if the segment of the route taken by the robot between successive updates of the pose is sufficiently large, even after a short period of time after the start, the error of the pose estimation is high compared to its real location. This phenomenon can be found primarily in cases in which a localization is performed based only on odometric sensors and is due to systematic errors, such as the presence of wheels with different diameters or the misalignment between the wheels, and non-systematic errors such as slippage of the wheels or the presence of irregular contact surfaces [22].

We could then face the drift problem by comparing a highest possible number of laser scans. There are various solutions to the drift problem, most of which are based on the "sensor fusion", doing measurements by multiple sensors that interact in order to obtain an estimate of the pose that is as close as possible to the real one.

There are several methods to achieve this goal, among which the most popular are the so-called “Bayesian filters” that estimate a state x from noisy sensory measurements. This category includes the “Kalman Filter” (with its extensions) and “particle filters”. Looking at the problem from a probabilistic point of view, the robot does not have, instant by instant, the certainty of where he is, but can believe (“belief”) to be in a certain position with a certain uncertainty. On the basis of this statement, the localization problem consists in the estimation of the probability density related to all possible positions, with the aim of obtaining as much knowledge as possible accurate position. Ideally, this occurs when the “belief” has a single peak at the position of the robot and is zero elsewhere.

Returning to the “Kalman Filter”, recursive algorithm that estimates the state of a linear dynamic system affected by noise, this has access to the measurements of sensors which have a linear dependence with the state of the system. It is shown that the Kalman filter converges to the optimal estimation, the one that minimizes the variance of the error of the estimate, assuming the linearity of the system model and measurement, and the corresponding noise is Gaussian with zero mean. Therefore we can say that the Kalman filter calculates the so-called “belief” (which is supposed to have a gaussian) of the state through two phases: the prediction, which calculates the “a priori belief”, i.e. the conditional probability of being in state x_k known the measures until the time k-1, while in the correction phase calculates the “belief a posteriori”, i.e. the conditional probability of being in state x_k known measures up to the instant k.

We are currently working on “particle filters” that allow to derive the estimate of the state (typically a function of the probability density not Gaussian and multimodal) in a system characterized by a nonlinear model.

The algorithm of the particle filter is recursive and consists of two phases: the prediction and updating. Following each action performed by the robot starts the prediction phase in which each particle is modified according to the existing model with the addition of noise to the variable of interest. During the upgrade, any weight of each particle is evaluated according to the new measurements from the sensors. The goal yet to be achieved is to get to the implementation of occupancy grid, which involves the construction, starting from the knowledge of the pose of all scans referred to the reference scan, an occupancy grid map.

The method used for the construction of the grid will be the one already described and the result of this operation will be the partition of the map in a grid in which each element of the grid itself is associated with a probabilistic value of occupancy. By using the occupancy grid more information from different sensors can be integrated in the same representation of the environment, even if they use different methods of data acquisition.

23.5 Conclusions

Of course, although we must emphasize that the results obtained from the integration are to now only been achieved in a "simulated" way and the automation of the procedure is still under study and implementation (having now moved to the cart only by hand), yet the results seem encouraging in view of the realization of a "expeditious" process for the auto positioning and perimetering by using mobile and automated tools.

The results certainly push to further study both in terms of actual full realization of the experiment, both in terms of optimization of the algorithms used for the compensation of the integrated data.

Acknowledgements We thank the laboratory of automatic of Mediterranean University and the Leica Geosystems for the support and the provision of necessary equipment for the data acquisition for the experiment presented.

References

1. Weiß G, Wetzler C, von Puttkamer E (1994) Keeping track of position and orientation of moving indoor systems by correlation of range-finder scans. Proc Intern Conf on Intelligent Robots and Systems, pp 595–601
2. Lu F, Milios E (1997) Robot Pose Estimation in Unknown Environments by Matching 2D Range Scans. J Intelligent and Robotic Systems 18:pp. 249–275
3. Thrun S (2002) Robotic Mapping: A Survey. In: Exploring artificial intelligence in the new millennium. Morgan Kaufmann, Pittsburgh
4. Liu JS, Chen R, Logvinenko T (2001) A theoretical framework for sequential importance sampling and resampling. In: Doucet A, de Freitas N, Gordon NJ (eds) Sequential Monte Carlo in Practice. Springer-Verlag, New York
5. Pirjanian P, Karlsson N, Goncalves L, Di Bernardo E (2003) Low-cost visual localization and mapping for consumer robotics. In: Industrial Robot: An International Journal 30:pp 139–144.
6. Rekleitis IM (2004) A particle filter tutorial for mobile robot localization. (Technical Report TR-CIM-04-02)
7. G. Bekey (2005) Autonomous Robots: From Biological Inspiration to Implementation and Control. The MIT Press, Cambridge, MA
8. Lingemann K, Nüchter A, Hertzberg J, Surmann H (2005) High-Speed Laser Localization for Mobile Robots. J Robotics and Autonomous Systems (JRAS), Elsevier Science 51:pp 275–296
9. Garulli A, Giannitrapani A, Rossi A, Vicino A (2005) Simultaneous localization and map building using linear features. In: Proc. 2nd European Conf Mobile Robots, Ancona (Italy)
10. Garulli A, Giannitrapani A, Rossi A, Vicino A (2005) Mobile robot SLAM for line-based environment representation, Decision and Control. 2005 European Control Conference. CDC-ECC '05, pp 2041–2046
11. Aghamohammadi AA, Taghirad HD, Tamjidi AH, Mihankhah E (2007) Feature-Based Laser Scan Matching For Accurate and High Speed Mobile Robot Localization. In: European Conf on Mobile Robots (ECMR'07)
12. Aghamohammadi AA, Tamjidi AH, Taghirad HD (2008) SLAM Using Single Laser Range Finder. In: Proc. 17th World Congress, The International Federation of Automatic Control, Seoul, Korea

13. Secchia M, Uccelli F (2012) Laser Scanner e GPS -Stop&Go. In: FIG Working Week 2012, Knowing to manage the territory, protect the environment, evaluate the cultural heritage, Rome, Italy
14. Wahde M (2012) Introduction to autonomous robots. Department of Applied Mechanics, Chalmers University of Technology, Goteborg, Sweden
15. Barrile V, Meduri GM, Bilotta G (2009) Laser scanner surveying techniques aiming to the study and the spreading of recent architectural structures. In: Recent advances in signals and systems, Proc 9th WSEAS Intern Conf on Signal, Speech and Image Processing, pp 92–95
16. Barrile V, Meduri GM, Bilotta G (2011) Laser scanner technology for complex surveying structures. In: WSEAS Trans Signal Processing 7:pp 65–74
17. Barrile V, Bilotta G, Meduri GM (2013) Least Squares 3D Algorithm for the Study of Deformations with Terrestrial Laser Scanner. In: Rec Adv in Electronics, Signal Processing and Communication Systems, Proc Intern Conf Electronics, Signal Processing and Communication Systems, Venice, Italy, pp 162–165
18. Bailey T, Nebot E (2001) Localisation in large-scale environments. In: Robotics and Autonomous Systems, 37:pp 261–281
19. Siegwart R, Nourbakhsh IR (2004) Introduction to Autonomous Mobile Robots. A Bradford Book, The MIT Press, Cambridge, MA, London, England
20. Borenstein J, Everett HR, Feng L, Wehe D (1997) Mobile Robot Positioning—Sensors and Techniques. In: J Robotic Systems, Mobile Robots 14:pp 231–249
21. Barrile V, Cacciola M, Cotroneo F, Morabito FC, Versaci M (2006) TECMeasurements through GPS and Artificial Intelligence. In: J Electromagnetic Waves and Applications *20:*pp 1211–1220
22. M. Postorino N, Barrile V (2004) An integrated GPS-GIS surface movement ground control system. In: Management of Information Systems, WIT Press (GBR), pp 3–12