Christian Constanda
Matteo Dalla Riva
Pier Domenico Lamberti
Paolo Musolino
Editors

Integral Methods in Science and Engineering, Volume 2

Practical Applications

Birkhäuser

Christian Constanda • Matteo Dalla Riva
Pier Domenico Lamberti • Paolo Musolino
Editors

Integral Methods in Science and Engineering, Volume 2

Practical Applications

Birkhäuser

Editors
Christian Constanda
Department of Mathematics
The University of Tulsa
Tulsa, OK, USA

Matteo Dalla Riva
Department of Mathematics
The University of Tulsa
Tulsa, OK, USA

Pier Domenico Lamberti
Department of Mathematics
University of Padova
Padova, Italy

Paolo Musolino
Systems Analysis, Prognosis and Control
Fraunhofer Institute for Industrial Math
Kaiserslautern, Germany

ISBN 978-3-319-59386-9 ISBN 978-3-319-59387-6 (eBook)
DOI 10.1007/978-3-319-59387-6

Library of Congress Control Number: 2015949822

Printed on acid-free paper

This book is published under the trade name Birkhäuser, www.birkhauser-science.com
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The international conferences on Integral Methods in Science and Engineering (IMSE), started in 1985, are attended by researchers in all types of theoretical and applied fields, whose output is characterized by the use of a wide variety of integration techniques. Such methods are very important to practitioners as they boast, among other advantages, a high degree of efficiency, elegance, and generality.

The first 13 IMSE conferences took place in venues all over the world:

1985, 1990: University of Texas at Arlington, USA
1993: Tohoku University, Sendai, Japan
1996: University of Oulu, Finland
1998: Michigan Technological University, Houghton, MI, USA
2000: Banff, AB, Canada (organized by the University of Alberta, Edmonton)
2002: University of Saint-Étienne, France
2004: University of Central Florida, Orlando, FL, USA
2006: Niagara Falls, ON, Canada (organized by the University of Waterloo)
2008: University of Cantabria, Santander, Spain
2010: University of Brighton, UK
2012: Bento Gonçalves, Brazil (organized by the Federal University of Rio Grande do Sul)
2014: Karlsruhe Institute of Technology, Germany

The 2016 event, the fourteenth in the series, was hosted by the University of Padova, Italy, July 25–29, and gathered participants from 26 countries on five continents, enhancing the recognition of the IMSE conferences as an established international forum where scientists and engineers have the opportunity to interact in a direct exchange of promising novel ideas and cutting-edge methodologies.

The Organizing Committee of the conference was comprised of

Massimo Lanza de Cristoforis (University of Padova), chairman,
Matteo Dalla Riva (The University of Tulsa),
Mirela Kohr (Babes-Bolyai University of Cluj-Napoca),
Pier Domenico Lamberti (University of Padova),

Flavia Lanzara (La Sapienza University of Rome), and
Paolo Musolino (Aberystwyth University),

assisted by Davide Buoso, Gaspare Da Fies, Francesco Ferraresso, Paolo Luzzini, Riccardo Molinarolo, Luigi Provenzano, and Roman Pukhtaievych.

IMSE 2016 maintained the tradition of high standards set at the previous meetings in the series, which was made possible by the partial financial support received from the following:

The International Union of Pure and Applied Physics (IUPAP)
GruppoNazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA), INDAM
The International Society for Analysis, Its Applications and Computation (ISAAC)
The Department of Mathematics, University of Padova

The participants and the Organizing Committee wish to thank all these agencies for their contribution to the unqualified success of the conference.

IMSE 2016 included four minisymposia:

Asymptotic Analysis: Homogenization and Thin Structures; organizer: M.E. Pérez (University of Cantabria)
Mathematical Modeling of Bridges; organizers: E. Berchio (Polytechnic University of Torino) and A. Ferrero (University of Eastern Piedmont)
Wave Phenomena; organizer: W. Dörfler (Karlsruhe Institute of Technology)
Wiener-Hopf Techniques and Their Applications; organizers: G. Mishuris (Aberystwyth University), S. Rogosin (University of Belarus), and M. Dubatovskaya (University of Belarus)

The next IMSE conference will be held at the University of Brighton, UK, in July 2018. Further details will be posted in due course on the conference web site blogs.brighton.ac.uk/imse2018.

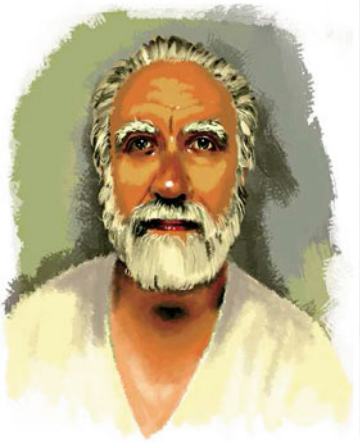
The peer-reviewed chapters of these two volumes, arranged alphabetically by first author's name, are based on 58 papers from among those presented in Padova. The editors would like to thank the reviewers for their valuable help and the staff at Birkhäuser-New York for their courteous and professional handling of the publication process.

Tulsa, OK, USA — Christian Constanda
March 2017

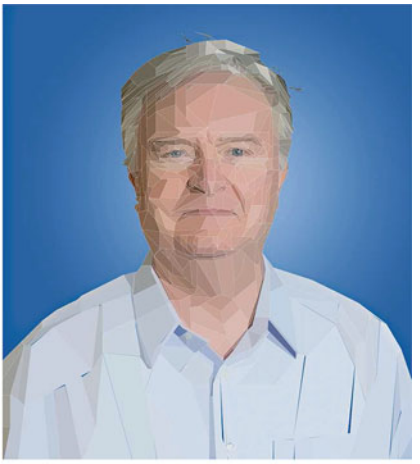
The International Steering Committee of IMSE:

A novel feature at IMSE 2016 was an exhibition of digital art that consisted of seven portraits of participants and a special conference poster, executed by artist Walid Ben Medjedel using eight different techniques. The exhibition generated considerable interest among the participants, as it illustrated the subtle connection between digital art and mathematics. The portraits, in alphabetical order by subject, and the poster have been reduced to scale and reproduced on the next two pages.

Digital Art by Walid Ben Medjedel

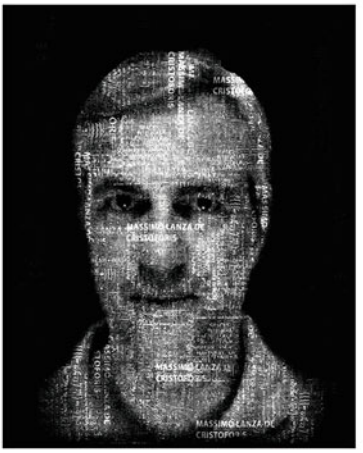


Mario Ahues
Acrilic portrait



Christian Constanda
Polygon portrait

Mirela Kohr
Vector portrait

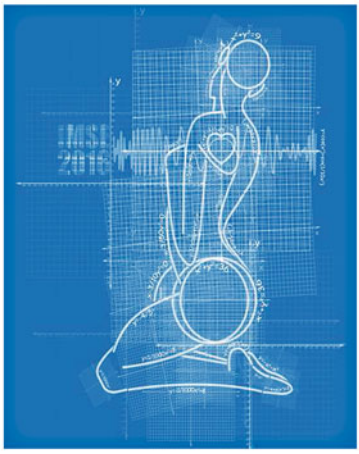


Massimo Lanza de Cristoforis
Text portrait

Flavia Lanzara
Airbrushing portrait

Dorina Mitrea
Pop-art portrait

Ovadia Shoham
Ink-pen portrait



IMSE2016 special poster
Mixed media

Contents

List of Contributors

Nadia M. Abusag Nottingham Trent University, Nottingham, UK

Mario Ahues Blanchait University of Lyon, Saint–Étienne, France

Éder L. de Albuquerque University of Brasilia, Brasilia, Brazil

Filomena D. d'Almeida University of Porto, Porto, Portugal

Francesco Altomare University of Bari, Bari, Italy

Antônio C.A. Alvim Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Tsegaye G. Ayele Addis Ababa University, Addis Ababa, Ethiopia

Luiz F.F.C. Barcellos Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Ricardo C. Barros State University of Rio de Janeiro, Nova Friburgo, RJ, Brazil

Miriam Belmaker The University of Tulsa, Tulsa, OK, USA

Elvise Berchio Polytechnic University of Torino, Torino, Italy

Bardo E.J. Bodmann Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Davide Buoso Polytechnic University of Torino, Torino, Italy

Lucas Campos University of Brasilia, Brasilia, Brazil

Haroldo F. Campos Velho National Institute for Space Research, São José dos Campos, SP, Brazil

Luis P. Castro University of Aveiro, Aveiro, Portugal

David J. Chappell Nottingham Trent University, Nottingham, UK

Nadia Chuzhanova Nottingham Trent University, Nottingham, UK

Alberto Cialdea University of Basilicata, Potenza, Italy

David L. Colton University of Delaware, Newark, DE, USA

Rafael Company Polytechnic University of Valencia, Valencia, Spain

Christian Constanda The University of Tulsa, Tulsa, OK, USA

Jonathan J. Crofts Nottingham Trent University, Nottingham, UK

Jesús Pérez Curbelo State University of Rio de Janeiro, Nova Friburgo, RJ, Brazil

Fateme Daburi Farimani Ferdowsi University of Mashhad, Mashhad, Iran

Panagiotis Dimitrakopoulos The University of Maryland, College Park, MD, USA

Patrizia Donato University of Rouen Normandie, Saint-Étienne-du-Rouvray, France

Dale Doty The University of Tulsa, Tulsa, OK, USA

Maryna V. Dubatovskaya The Belarusian State University, Minsk, Belarus

Tamirat T. Dufera Adama Science and Technology University, Adama, Ethiopia

Mahdi A. Esfahani Ferdowsi University of Mashhad, Mashhad, Iran

Carlos E. Espinosa Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Mohamed Fakharany Tanta University, Tanta, Egypt

Julio C.L. Fernandes Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Rosário Fernandes University of Minho, Braga, Portugal

Gustavo Fernández-Torres National Autonomous University of México, Ciudad de México, México

Milton Ferreira Polytechnic Institute of Leiria, Leiria, Portugal

Fahin Forouzanfar The University of Tulsa, Tulsa, OK, USA

Saulo R. Freitas NASA Goddard Space Flight Center, Greenbelt, MD, USA

Igor C. Furtado Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Carlos R. García Institute of Technology and Applied Sciences, La Habana, Cuba

Filippo Gazzola Polytechnic University of Milano, Milano, Italy

Delfina Gómez University of Cantabria, Santander, Spain

Rita C. Guerra University of Aveiro, Aveiro, Portugal

Paul J. Harris University of Brighton, Brighton, UK

Gabriel Hattori Durham University, Durham, UK

Randy D. Hazlett The University of Tulsa, Tulsa, OK, USA

Adino Heimlich Nuclear Engineering Institute, Rio de Janeiro, RJ, Brazil

Lucas Jódar Polytechnic University of Valencia, Valencia, Spain

Hanane Kaboul University of Lyon, Saint–Étienne, France

Saleh Kargarrazi The Royal Institute of Technology, Kista, Sweden

Abolfazl Karimpour Iran University of Science and Technology, Tehran, Iran

Ali Karimpour Ferdowsi University of Mashhad, Mashhad, Iran

Mostafa Karimpour Ferdowsi University of Mashhad, Mashhad, Iran

Yuri I. Karlovich Autonomous State University of Morelos, Cuernavaca, Morelos, México

Sam H. Kettle Durham University, Durham, UK

Andreas Kleefeld Forschungszentrum Jülich GmbH, Jülich, Germany

Kianoush Kompany Virginia Tech, Blacksburg, VA, USA

Gene Kouba Chevron Energy Technology Company (Retired), Houston, TX, USA

Piotr Kozarzewski University of Warsaw, Warsaw, Poland

Cibele A. Ladeia Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Luca Lanzoni University of San Marino, San Marino, San Marino

Sergio Q. Bogado Leite National Nuclear Energy Commission, Rio de Janeiro, RJ, Brazil

Vita Leonessa University of Basilicata, Potenza, Italy

Angelica Malaspina University of Basilicata, Potenza, Italy

Clelia Marchionna Polytechnic University of Milano, Milano, Italy

Tamires Marotto North Fluminense State University, Macaé, RJ, Brazil

Rebecca Martin Nottingham Trent University, Nottingham, UK

Aquilino S. Martinez Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

José M.A. Matos Politechnic School of Engineering, Porto, Portugal

André Meneghetti Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Abdelaziz Mennouni University of Batna 2, Batna, Algeria

Sergey E. Mikhailov Brunel University West London, Uxbridge, UK

Dorina Mitrea University of Missouri, Columbia, MO, USA

Ram Mohan The University of Tulsa, Tulsa, OK, USA

Mirella Cappelletti Montano University of Bari, Bari, Italy

Hunghu Nguyen The University of Tulsa, Tulsa, OK, USA

Andrea Nobili University of Modena and Reggio Emilia, Modena, Italy

Nicholas H. Okamoto University of Missouri, Columbia, MO, USA

Fernando R. Oliveira Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Stefano Panizzi University of Parma, Parma, Italy

Maria V. Perel St. Petersburg State University, St. Petersburg, Russia

Alvaro M.M. Peres North Fluminense State University, Macaé, RJ, Brazil

Maria Eugenia Pérez University of Cantabria, Santander, Spain

Gary Phillips University of Brighton, Brighton, UK

Adolfo P. Pires North Fluminense State University, Macaé, RJ, Brazil

Alexander V. Podol'skii Moscow State University, Moscow, Russia

Carlos F. Portillo Oxford Brookes University, Wheatley, UK

Hamid-Reza Pourreza Ferdowsi University of Mashhad, Mashhad, Iran

Maryam Pourreza Sharif University of Technology, Tehran, Iran

Enrico Radi University of Modena and Reggio Emilia, Reggio Emilia, Italy

Federica Raimondi University of Salerno, Fisciano, SA, Italy

Habib Rajabi Mashhadi Ferdowsi University of Mashhad, Mashhad, Iran

Ioan Raşa Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Elisabeth Reichwein Heinrich-Heine-Universität, Düsseldorf, Germany

Manuela Rodrigues University of Aveiro, Aveiro, Portugal

Sergei V. Rogosin The Belarusian State University, Minsk, Belarus

Heloisa M. Ruivo National Institute for Space Research, São José dos Campos, SP, Brazil

Renata S.R. Ruiz National Institute for Space Research, São José dos Campos, SP, Brazil

Javad Sadeh Ferdowsi University of Mashhad, Mashhad, Iran

Aleksey V. Setukha Lomonosov Moscow State University, Moscow, Russia

Tatiana A. Shaposhnikova Moscow State University, Moscow, Russia

Ovadia Shoham The University of Tulsa, Tulsa, OK, USA

Mikhail S. Sidorenko Ioffe Institute, St Petersburg, Russia

Fernando C. Silva Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Odair P. da Silva State University of Rio de Janeiro, Nova Friburgo, Brazil

Adel Soheili Ferdowsi University of Mashhad, Mashhad, Iran

Stanislav L. Stavtsev Institute of Numerical Mathematics of the Russian Academy of Sciences, Moscow, Russia

Olaf Steinbach Graz University of Technology, Graz, Austria

P. Stpiczyński Maria Curie-Skłodowska University, Lublin, Poland

Jon Trevelyan Durham University, Durham, UK

Marcelo S. Trindade University of Porto, Porto, Portugal

Nguyen M. Tuan National University of Viet Nam, Hanoi, Vietnam

Paulo B. Vasconcelos University of Porto, Porto, Portugal

Vladimir B. Vasilyev National Belgorod Research State University, Belgorod, Russia

Jenny Venton University of Brighton, Brighton, UK

Bruno J. Vicente North Fluminense State University, Macaé, RJ, Brazil

Nelson Vieira University of Aveiro, Aveiro, Portugal

Marco T.B.M. Vilhena Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Darko Volkov Worcester Polytechnic Institute, Worcester, MA, USA

Maria E.S. Welter National Institute for Space Research, São José dos Campos, SP, Brazil

Elvira Zappale University of Salerno, Fisciano (SA), Italy

Sergey Zheltukhin Rifiniti, Inc., Boston, MA, USA

Barbara Zubik-Kowal Boise State University, Boise, ID, USA

Chapter 1
On a Continuous Energy Monte Carlo Simulator for Neutron Transport: Optimisation with Fission, Intermediate and Thermal Distributions

L.F.F. Chaves Barcellos, B.E.J. Bodmann, S.Q. Bogado Leite, and M.T. Vilhena

1.1 Introduction

Neutron transport is relevant in a variety of applications as, for instance, in medicine, industrial applications, radiation protection and nuclear energy production among others. In this context, the present work reports on the development of a simulator for neutron transport considering continuous energy dependence of cross sections [CaEtAl11, CaEtAl13]. As a progress in comparison to other implementations, the cross sections are parametrisations in the range between 0 MeV and 20 MeV, including resolved and unresolved resonances, and with a maximum deviation smaller than $\sim$ 1% from measured data. Other implementations may be found in the literature such as Serpent [Le15], MCNP [Mo03], Tripoli [BoEtAl03], OpenMC [RoFo13], Keno [PeCo75], GEANT [AgEtAl03], MCBend [CoEtAl13], where cross sections are determined from interpolation of cross section from databases.

In the present contribution we report on an optimisation of a Monte Carlo simulator based on the interaction and tracking philosophy also found in GEANT. In the former neutrons are classified according to three overlapping energy distributions (fission, intermediate and thermal). Neutrons from fission and during slowing down suffer predominantly down-scattering, whereas in the thermal region neutrons may gain kinetic energy from collisions with nuclei and molecules due to their thermal motion. To circumvent simulating thermal up- and down-scattering that do not significantly change properties of the thermal neutron population, we introduce a statistical treatment reducing the problem by considering reaction rates only.

L.F.F.C. Barcellos (✉) • B.E.J. Bodmann • M.T. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: luizfelipe.fcb@gmail.com; bardo.bodmann@ufrgs.br; mtmbvilhena@gmail.com

S.Q.B. Leite
National Nuclear Energy Commission, Rio de Janeiro, RJ, Brazil
e-mail: sbogado@ibest.com.br

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_1

The fission and the thermal distributions preserve shape, but their respective integrals may vary with time, whereas the intermediate distribution has unknown shape and integral. It is noteworthy that all distributions are continuous over the whole energy range and thus one faces the challenge to determine for a neutron with given kinetic energy to which distribution it belongs.

1.2 Neutron Transport by a Monte Carlo Method

The underlying philosophy of the present simulator follows the paradigm of the GEANT platform, which besides efficient geometry resource makes use of tracking and interaction algorithms. The present simulator considers the same degrees of freedom as the Boltzmann transport equation, namely position, time, propagation direction and kinetic energy. Different than deterministic models found in the literature such as diffusion theory, the P_N and S_N approximation for the transport equation [Sj13] and the references therein, the method employed here to attain physical information of the transport phenomenon is by sampling of a sufficiently high number of neutron histories from a physical Monte Carlo procedure that allows to determine quantities such as the spectral neutron population, the angular or scalar flux depending on the specific tags that are being used either in the simulation or in a posterior data evaluation. With the present contribution we simulate a simplified reactor neutron problem and focus on the question of identifying the distribution a neutron with a specific energy belongs to. The problem of identification arises due to the fact that two adjacent distributions overlap significantly in certain energy regions.

1.3 Program Description

The C++ Monte Carlo simulator in development features sectionally analytical functions for the energy dependent microscopic cross sections in the range from 0 MeV to 20 MeV. In the present case 200 executions were performed, each starting with 5000 neutrons, and ending up with 10^6 neutron histories. For tallying reasons linked to computer hardware constraints each execution was limited to 5000 Monte Carlo steps, and these were segmented in 50 intervals of 100 steps each, i.e. after 100 steps the simulation reached a checkpoint, where it was halted and the respective dataset was saved. The subsequent run then used these data as the initial condition for the following 100 steps.

At the beginning of each Monte Carlo step, neutrons created by fission are given two random angles (between $[0, 2\pi]$ and $[-\pi/2, \pi/2]$, respectively) that define their direction and further a random energy that obeys the fission distribution and the positions are given by coordinates of the fission reaction.

$$\chi(E) = 0.453\, e^{-1.036\, MeV^{-1}\, E} \sinh \sqrt{2.29\, MeV^{-1}\, E} \tag{1.1}$$

Note, the tracking and the interaction scheme was optimised in the sense that each Monte Carlo step has an interaction, which increases computational efficiency, but at the cost of loosing a unique relation between Monte Carlo step and corresponding time interval. Thus, a Monte Carlo step may be related only to an average of a time interval distribution that may be reconstructed from the tallies. After the displacement of the neutron its position is checked in order to evaluate whether it remains still in the reactor core volume or whether it escaped, where in the latter case the history of the neutron ends and a new neutron is selected. Finally, the type of neutron interaction is selected, which is based on both region and neutron energy.

In the case of radiative capture the procedure is the same as for escape, the neutron's history ends and a new neutron is chosen. In case that fission occurs, the history of the fission inducing neutron ends and a multiplicity of new neutrons is generated. In case of scattering the energy and the direction angles are updated for the next step. The main structure of the program is shown on the flowchart in Figure 1.1.

As a simplified case study, we consider the geometry of the reactor by a cube with edges of dimensions $400\,cm \times 400\,cm \times 400\,cm$. The inner part contains three regions, where region 1 measures $250\,cm \times 250\,cm \times 400\,cm$ and contains a homogeneous mixture of water and uranium dioxide enriched to 0.73% and the latter occupies 25% of the respective volume. Around the central box there is a hollow box, i.e. region 2, with extensions $350\,cm \times 350\,cm \times 400\,cm$ and is composed of water. There is a second hollow box, allocated in region 3 with a homogeneous mixture of water and uranium dioxide, but with completely depleted uranium dioxide which occupies 45% of the respective volume. For convenience we adopted periodic boundary conditions in the vertical direction (aligned with the $z-axis$). The program executes the tracking and interaction of neutrons in the whole volume.

The position in which a reaction will occur at the end of a Monte Carlo step depends on the kinetic energy of the neutron, its position at the beginning of the step, the direction of movement and the total macroscopic cross sections of the chemical composition of the reactor core material along the trajectory. The final position of the track will then be determined by a stochastic selection for the length of the travelled path. To this end a multiple S of the mean free path is generated by a random number, following a standard procedure $S = -\ln(1-a)$ and $a \in [0,1]$. Consequently the length of the path is $L = S\Sigma_t^{-1}$ where Σ_t is the microscopic total cross section characteristic for the path. In case a neutron crosses a boundary between sub-domains the path length is calculated by a weighted sum of cross section contributions characteristic for the respective regions $L = \sum_i P_i S\, \Sigma_t^{-1}$, where P_i is the fraction of S that corresponds to the trajectory segment within the i-th sub-domain. After updating the position, a verification checks whether the neutron remains inside the boundaries of the reactor core volume, and thus whether the neutron tally continues inside or terminates outside the domain.

After the position of the interaction is defined, the target involved in the reaction is chosen. In Region 2 the target is a water molecule or one of its constituents (H and O), whereas in regions 1 or 3 a random number is generated and compared

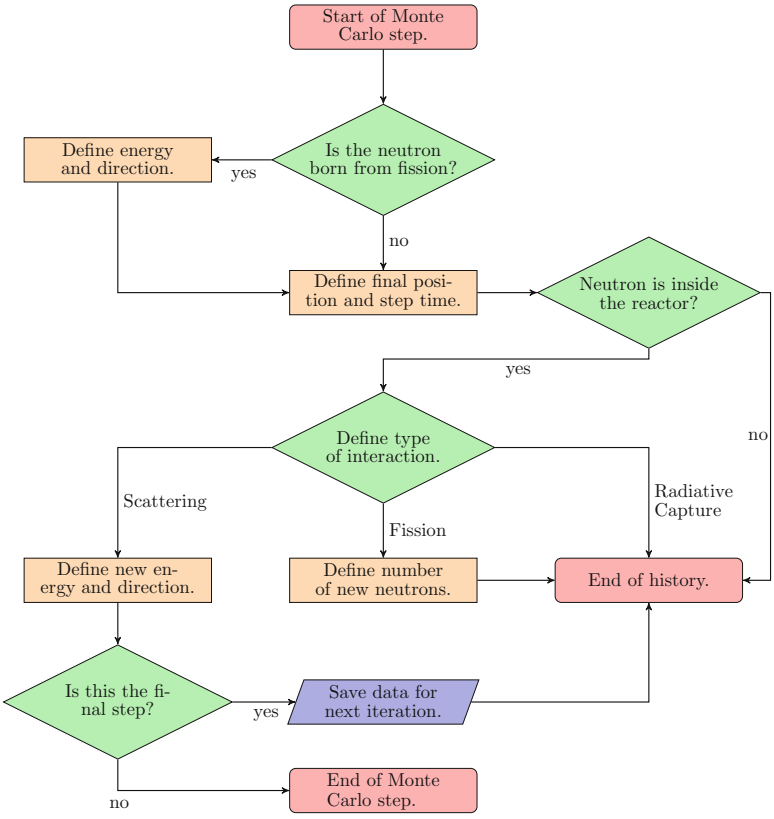


Fig. 1.1 Program flowchart with its principal instances

to the volume proportions of water and uranium dioxide, more specifically to its constituents. The subsequent step is then to select the type of interaction by generating a random number which is compared to the stoichiometric ratio of the cross section of all possible targets. As an example, the probability of a reaction in uranium dioxide (UO_2) is given by $p_i = \frac{\sigma_i}{2\,\sigma_{t,O} + e\,\sigma_{t,U-235} + (1-e)\,\sigma_{t,U-238}}$, in which e is the enrichment, $\sigma_{t,O}$ is the total cross section of oxygen-16, $\sigma_{t,U-235}$ the total cross section of uranium-235, $\sigma_{t,U-238}$ the total cross section of uranium-238 and σ_i is the cross section of a specific neutron reaction in one of the nuclei.

1.4 Nuclear Reactions

If the chosen reaction is fission two stochastic operations are in order. The first one is to decide the number of neutrons born from fission, and the second is to define their energies. In order to define the number of neutrons from fission a random

number ν is generated between 0 and 0.972. The upper limit was chosen such as to guarantee that the average number of neutrons created in fission coincides with the expected value of $\nu = 2.48$ for fission induced by thermal neutrons and nuclear fuel U-235. It is noteworthy that the huge bulk of fission reactions releases either two or three neutrons, so that to a good approximation only these two cases are taken into account. These neutrons have energies roughly in the range between 10^0 MeV up to 10^1 MeV as given by Equation (1.1). The position of the fission reaction is also recorded for it is the initial position of the next Monte Carlo step of the newly generated neutrons. At the present state of developments no contributions due to delayed neutrons are considered, this pertinent issue will be included in the next version of the simulator.

In case of scattering, a new energy and a new direction in agreement with energy and momentum conservation must be given to the neutron. A simplification of the program is that it considers the scattering as isotropic in the centre of mass system. Strictly speaking, scattering is isotropic for low kinetic energies and small nuclei, however, as the collision energies become higher and/or target nuclei become larger, anisotropy increases. So far the described processes treat down-scattering only, i.e. energy loss of neutrons in their interactions with their respective targets [GlSe94].

$$\frac{E'}{E} = \frac{A^2 + 2A\,cos(\theta) + 1}{(A+1)^2}$$

Here E is the energy of the neutron in the *Laboratory* system before the collision, E' is the energy of the neutron in the *Laboratory* system after the collision and θ is the angle of scattering measured in the *Centre-of-Mass* system and that is in the plane that contains both vectors of incident direction and scattered direction of the neutron. However, the closer neutrons approach thermal energies also up-scattering is important, due to the thermal motion of the target nuclei which is no longer negligible. As soon as neutrons may be classified as thermal they are in equilibrium with the environment which allows to simplify the tracking and interaction procedure. Equilibrium implies conservation of the respective energy distribution, so that the only relevant stochastic quantity that shall be determined is the reaction rate.

The procedure to determine whether a neutron belongs to the thermal distribution is as follows. A random number between 0 and 1 is generated and compared to the Maxwell-Boltzmann cumulative distribution with an equilibrium temperature of 568 K. Should the random number be larger than the cumulative distribution, then the neutron is considered to be in thermal equilibrium with the moderator. Neutrons that are part of the thermal population are assigned a new energy, sampled from the Maxwell-Boltzmann probability distribution.

The procedure to find the new neutron direction after scattering is determined in a complete three-dimensional fashion, although cylinder symmetry would allow a reduction into a plane. Let the unit vector $\vec{\Omega}_i$ be the direction of the incoming neutron with $\alpha \in [0, 2\pi]$ and $\beta \in [-\pi/2, \pi/2]$ angles with respect to the laboratory reference frame (see Equation (1.2)). For convenience one may construct one

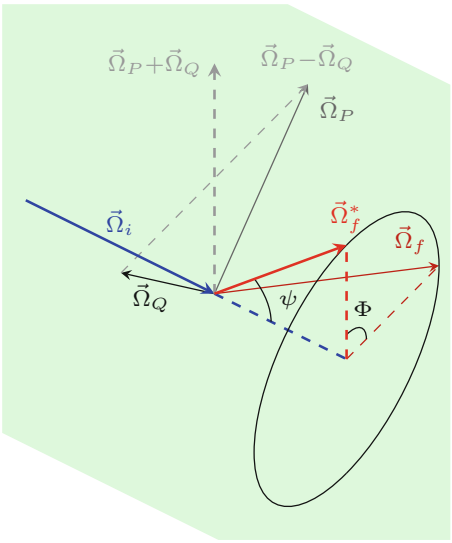


Fig. 1.2 Sketch of the neutron scattering scheme

possible final direction $\vec{\Omega}_f^*$ (see Figure 1.2). All remaining possible final vectors in agreement with cylinder symmetry may be generated with two auxiliary orthogonal vectors $\vec{\Omega}_P$ and $\vec{\Omega}_Q$ that by construction are symmetrical on either side of the scattering plane defined by $\vec{\Omega}_i$ and $\vec{\Omega}_f^*$ (see Equation (1.3)). The vector $\vec{\Omega}_P + \vec{\Omega}_Q$ lies then in the scattering plane, whereas $\vec{\Omega}_P - \vec{\Omega}_Q$ is perpendicular to the latter. The plane by $\vec{\Omega}_P$ and $\vec{\Omega}_Q$ defines the rotation plane that contains the circle with all possible outcomes for the final direction $\vec{\Omega}_f$ of the neutron after scattering (see Figure 1.2).

$$\vec{\Omega}_i = \begin{pmatrix} \cos(\alpha)\,\cos(\beta) \\ \sin(\alpha)\,\cos(\beta) \\ \sin(\beta) \end{pmatrix}, \qquad \vec{\Omega}_f^* = \begin{pmatrix} \cos(\alpha)\,\cos(\beta+\psi) \\ \sin(\alpha)\,\cos(\beta+\psi) \\ \sin(\beta+\psi) \end{pmatrix} \tag{1.2}$$

$$\begin{aligned} \frac{\sin(\psi)}{\sqrt{2}}\,(\vec{\Omega}_P - \vec{\Omega}_Q) &= \vec{\Omega}_i \times \vec{\Omega}_f^* \\ \frac{\sin(\psi)}{\sqrt{2}}\,(\vec{\Omega}_P + \vec{\Omega}_Q) &= \vec{\Omega}_f^* - \cos(\psi)\,\vec{\Omega}_i \end{aligned} \tag{1.3}$$

$$\vec{\Omega}_f = \cos(\psi)\,\vec{\Omega}_i + \sin(\psi)\,(\cos(\Phi)\,\vec{\Omega}_P + \sin(\Phi)\,\vec{\Omega}_Q)$$

Here $\Phi \in [0, 2\pi]$ is a random angle. A necessary feature of scattering not implemented yet is due to the fact that approximately below 1 *eV* instead of a single free nuclide one has to consider whether the atom is a constituent of a molecule or solid state, so that in the previous case molecular degrees of freedom such as rotation and vibration shall be considered, whereas in a solid state phonon degrees of freedom shall be taken into account.

1.5 Coupled Distributions

Several features that characterise the simulator are new and different to the other aforementioned neutron transport codes. Since none of them makes use of properties such as shape preservation of distributions or the fact that in the thermal regime consequences of thermal equilibrium may be explored, as a consistency test of the present implementation we compare a linearised model for the coupled distributions with direct findings from the simulation. To this end the following system of differential equations is considered and solved.

$$\frac{\partial}{\partial t}\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} = \begin{pmatrix} -\lambda_{f,c,e,1} & \lambda_{t(2,1)} & 0 \\ 0 & -\lambda_{f,c,e,2} - \lambda_{t(2,1)} & \lambda_{t(3,2)} \\ \nu\lambda_{f,1} & \nu\lambda_{f,2} & \nu\lambda_{f,3} - \lambda_{f,c,e,3} - \lambda_{t(3,2)} \end{pmatrix}\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix}$$

$$\lambda_{f,c,e,i} = \lambda_{f,i} + \lambda_{c,i} + \lambda_{e,i} \qquad \text{for } i \in 1, 2, 3$$

Here D is the total number of particles in each distribution, λ is the mean rate of each interaction per Monte Carlo step, ν is the mean number of neutrons emitted by fission, the subscripts 1, 2 and 3 represent, respectively, the thermal, intermediate and fission distribution, and the subscripts f, c, e and $t(i,j)$ represent, respectively, the fission reaction, radiative capture reaction, neutron leakage and the transition of a neutron from distribution i to distribution j. The aforementioned interaction rates are computed after the simulation is completed.

1.6 Results

Results were obtained for a starting population of 10^6 neutrons. The behaviour of the total population along all 5000 Monte Carlo steps is shown in Figure 1.3.

By inspection of Figure 1.3 one identifies a sub-critical regime, this will also be supported by the computation of the neutron multiplication factor. It is also possible to note an increase of the number of neutrons during the first steps of the simulation. This behaviour, apparently in contrast to the sub-critical tendency, is attributed to the fact that the simulation is started with a fission distribution only. Criticality can be evaluated by dividing the number of fissions caused by neutrons of one generation by the number of fissions caused by neutrons of the previous generation, in such a way that each time neutrons are created by fission they belong to a generation that follows the generation of the neutron that induced the fission reaction. Along the 5000 Monte Carlo step 252 neutron generations were identified. The resulting neutron multiplication factor is presented in Figure 1.4.

Figure 1.4 shows that after generation 200 the neutron multiplication factor deviates from the behaviour presented in previous steps. This can be explained by the fact that neutrons of different generations are present in the same Monte

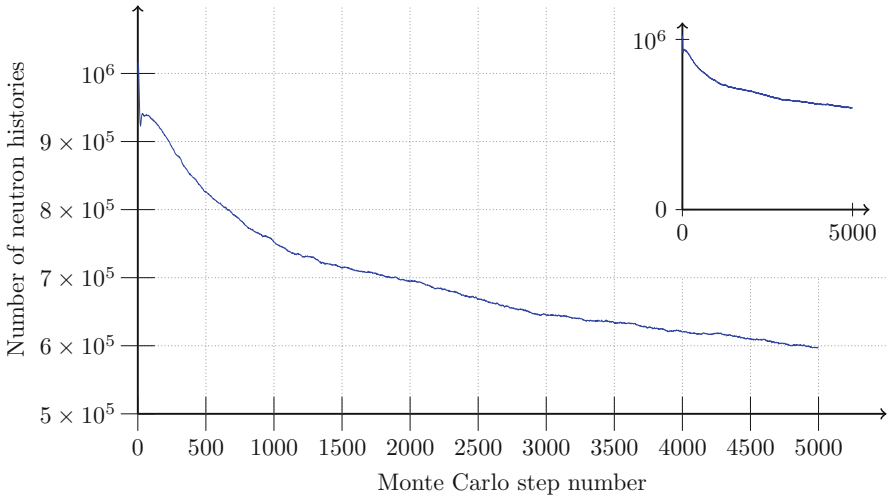


Fig. 1.3 Total number of neutrons per Monte Carlo step

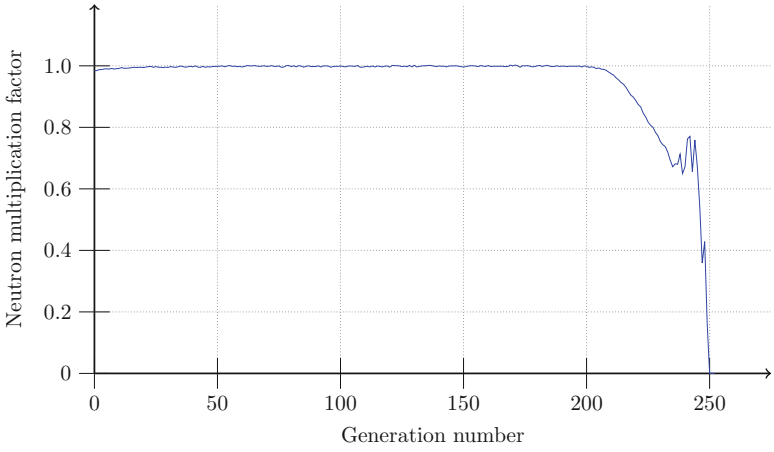


Fig. 1.4 Multiplication factor from the neutron life cycle

Carlo step and, as the simulation is stopped at step 5000, the number of neutrons in a generation that causes fission diminishes for the subsequent generations, and thus the decay of the multiplication factor is an artefact of the way the simulation terminates. The first generations are also less representative, for they are influenced by the specific conditions that define initialisation. For the generations 20 to 200 the geometric mean of the neutron multiplication factor was calculated with numerical value $k_{eff} = 0.998417$.

In Figure 1.5 the ratios of the populations of each of the three distributions by the total population for each step are shown.

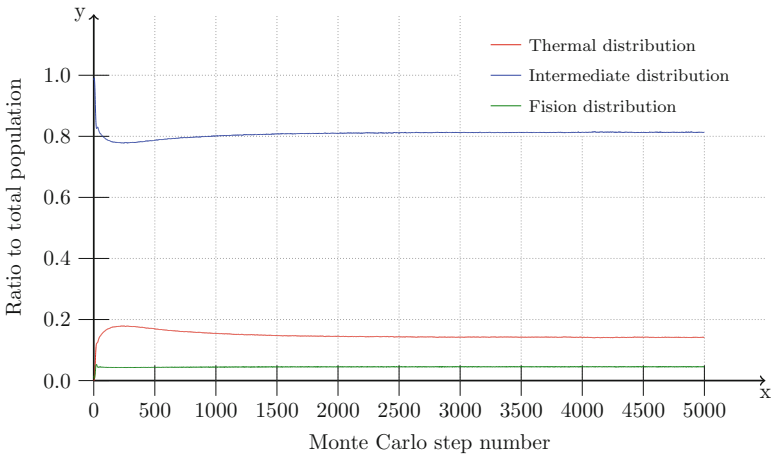


Fig. 1.5 Ratios of each distribution population to the total population from the simulation

In Figure 1.5 it is perceivable that, even though the total population varies, the proportions of the distributions are relatively constant for steps larger than $\sim$ 500. The preservation of these ratios is due to the fact that the neutron spectrum remains stationary along the Monte Carlo steps. The rates of the different interactions were computed and this information was used in the differential equation system. Due to the fact that the computed rates are mean values per step and that initialisation of the program induces a bias in the results, the initial condition for the system was taken from the respective population of each distribution at step 500. The solution is presented in Figure 1.6 as ratios to the total population, comparing the ratios from the linearised coupled distribution equation system to findings from the Monte Carlo simulation, which shows fairly good agreement and thus shows consistency of the implementation.

1.7 Conclusions and Future Work

The simulator in development is able to solve neutron transport problems in the complete phase space of the Boltzmann equation without simplifications or discretisations. It can successfully track and tag particles and their respective properties, allowing thus for the construction of physically relevant probability distribution functions, such as neutron density, angular and scalar fluxes. We showed results for the proportions of the fission, intermediate and thermal distributions, where a results from Monte Carlo simulated ratios were compared to the analytical result of the linearised model and showed fairly good agreement thus confirming consistency of the implementation. As a next step in the development of the simulator we will use resources of power computing to accelerate the simulation execution runs and prepare the playground for more complex future extensions.

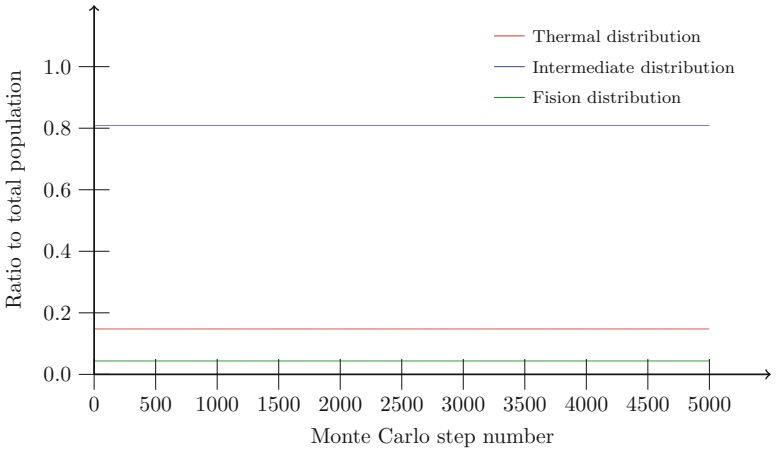


Fig. 1.6 Ratios of each distribution population to the total population from the linearised model

References

[AgEtAl03] Agostinelli, S., Allison, J., Amako, K., Apostolakis, J., Araujo, H., Arce, P., Asai, M., Axen, D., Banerjee, S., Barrand, G., Others: Geant4–a simulation toolkit. Nucl. Instrum. Methods Phys. Res. Sect. A Accelerators Spectrom. Detect. Assoc. Equip. **506**(3), 250–303 (2003)

[BoEtAl03] Both, J.P., Mazzolo, A., Peneliau, Y., Petit, O., Roesslinger, B.: User manual for version 4.3 of the Tripoli-4 Monte-Carlo method particle transport computer code (2003)

[CaEtAl11] de Camargo, D.Q., Bodmann, B.E.J., Vilhena, M.T., Bogado Leite, S.Q.: A novel method for simulating spectral nuclear reactor criticality by a spatially dependent volume size control. In: Constanda, C., Harris, J.P. (eds.) Integral Methods in Science and Engineering: Computational and Analytic Aspects, pp. 35–45. Birkhäuser, Boston (2011)

[CaEtAl13] de Camargo, D.Q., Bodmann, B.E.J., Vilhena, M.T., Bogado Leite, S.Q., Alvim, A.C.M.: A stochastic model for neutrons simulation considering the spectrum and nuclear properties with continuous dependence of energy. Prog. Nucl. Energy **69**, 59–63 (2013)

[CoEtAl13] Cowan, P., Dobson, G., Martin, J.: Release of MCBEND 11. In: Proceedings of the 12th International Conference on Radiation Shielding (ICRS-12) and 17th Topical Meeting on Radiation Protection and Shielding (RPSD-2012) (2013)

[GlSe94] Glasstone, S., Sesosnke, A.: Nuclear Reactor Engineering: Reactor Design Basics, vol. 1. Springer, New York (1994)

[Le15] Leppänen, J., Pusa, M., Viitanen, T., Valtavirta, V., Kaltiaisenaho, T.: The serpent Monte Carlo code: status, development and applications in 2013. Ann. Nucl. Energy **82**, 142–150 (2015)

[Mo03] X-5 Monte Carlo Team: MCNP: a general Monte Carlo N-particle transport code. Los Alamos National Laboratory, Los Alamos (2003)

[PeCo75] Petrie, L.M., Cross, N.F.: KENO IV: An Improved Monte Carlo Criticality Program. Oak Ridge National Lab., TN (1975)

[RoFo13] Romano, P.K., Forget, B.: The OpenMC Monte Carlo particle transport code. Ann. Nucl. Energy **51**, 274–281 (2013)

[Sj13] Sjenitzer, B.L.: The Dynamic Monte Carlo Method for Transient Analysis of Nuclear Reactors. Delft University of Technology, Delft (2013)

Chapter 2
The Use of Similarity Indices in the Analysis of Temporal Distribution of Mammals

M. Belmaker

2.1 Introduction

Ecological and paleontological studies look at changes in species composition between spatially and temporally distinct regions. In such cases, data are organized in a numerical $n \times p$ matrix or data frame, where n corresponds to different sampling times or sites (in this study assemblages) and p denotes each of the different variables that describe the locality studied. These may include the biological community (measured in species incidence or relative abundance), or other variables that connote the physical or chemical environment [Le98]. Thus, ecological datasets are multidimensional and represent a geometric hyperspace.

We identify change or stasis in such a system by means of β diversity, which is the variation in species composition between assemblages. It can be used to test for turnover (antonym inertia) and is measured as change in community structure from one assemblage to another along a gradient. If inertia is present in the system, then variability in species incidence or abundance between assemblages may result from sampling bias or other stochastic processes. However, if there is a positive correlation between community structure and a directional change along a vector, this may be related to monotonous changes in the abiotic or biotic environment.

In paleontological assemblages, on the scale of $10^6 - 10^7$ years, studies have described various patterns of recurring fossil assemblages [Mi93]. The best documented and known pattern is "coordinated stasis," which describes an empirical pattern of community level stasis coupled with an abrupt change in community structure of fossil assemblages. This inertia is present in spite of independent evidence for climate change. In contrast, in younger assemblages that date to the Quaternary (2.6 mya to present), community structure shifts predictably in response

M. Belmaker (✉)
The University of Tulsa, Tulsa, OK, USA
e-mail: miriam-belmaker@utulsa.edu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_2

to environmental change. The discrepancy in the faunal response between the pre-Quaternary (Phanerozoic) coordinated stasis and the Quaternary pattern [Ho96] was termed the "Pleistocene Paradox" [St01].

Little research has been done on the intermediate time scale ($10^4 - 10^6$). Here, we present a case study to illustrate the use of (dis)similarity indices and the Mantel's test in the temporal scale, to investigate changes in β diversity among seven mammal communities at the $4X10^5$ year scale.

2.2 The Case Study

The model system used in this study is the paleoanthropological site of 'Ubeidiya, central Jordan Valley, Israel. The site has been dated to approximately 1.6–1.2 million years ago (ma), and the site exhibits early human remains [Be02], as well as rich lithic and faunal assemblages [Ba93] (Figure 2.1).

A unique method of excavation was used at the site due to the extensive post-depositional tectonic faulting of the sediments. Four trenches, numbered I–IV, were excavated. Within each trench, the archaeological strata were numbered in Arabic numerals from oldest to youngest. For example, stratum III 12 is the 12th geological layer of trench III. The site exhibits over 113 uncovered archaeological strata [Ba93]. However, fossil remains of large enough samples have been found in seven strata only (Figure 2.2).

We studied the large (> 10 kg live weight) mammal communities of the seven strata of 'Ubeidiya from all seasons of excavations (1959–2001). Specimens identified to other taxa (small mammals, birds, reptiles, turtle, fish, and invertebrates) were not included in this study.

Bones were identified to the lowest taxonomic level possible. We report species and genus level analysis, eliminating bones identified only to family [Be06].

We calculated two dependent variable data matrices. The first was an incident-based matrix with 1 for presence and 0 for absence. The second was a relative (percent) abundance. We transformed abundances at each of the sites by adding one and taking logarithms.

In 'Ubeidiya, local hydrological conditions suggest a monotonic change (although not linear) in local environment from humid to dryer conditions from strata 7 through 1 [Ma06]. Therefore, an environmental grade was given to each stratum from 1 (driest) through 5 (wettest), as detailed in Table 2.1. For clarity, in lieu of using the original strata names, we labeled the fossil communities from 1 (youngest) through 7 (oldest) (see Table 2.1).

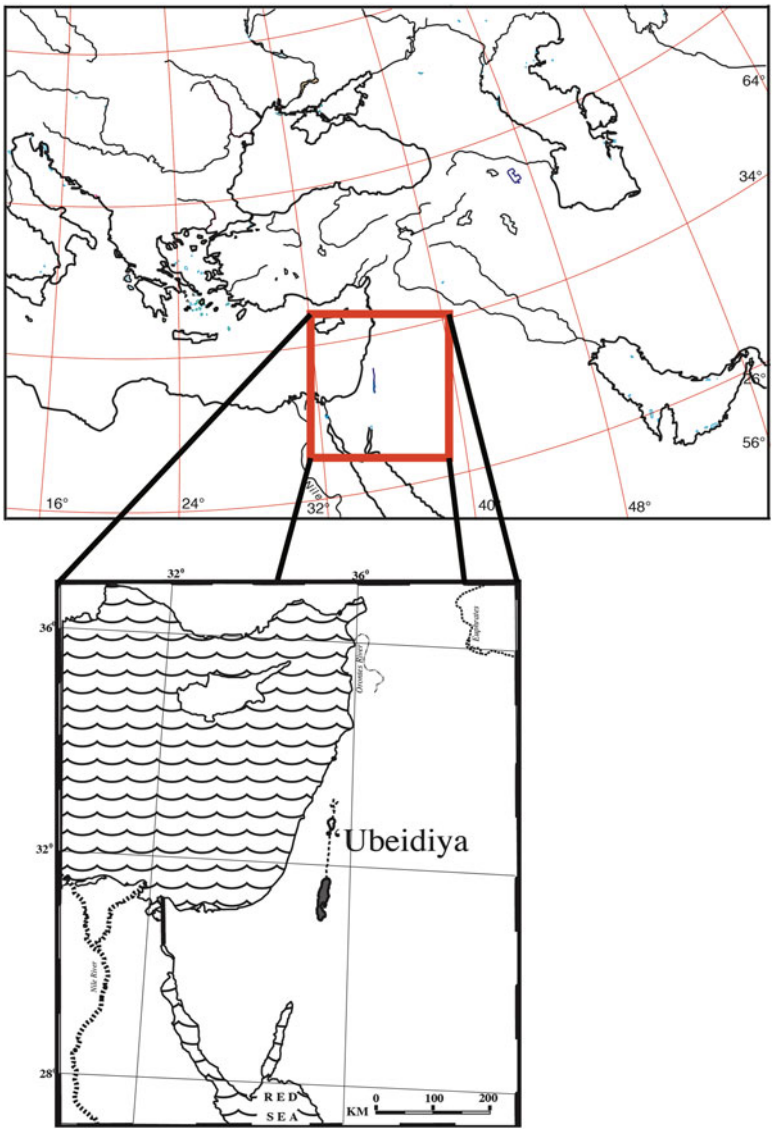


Fig. 2.1 Location of ‘Ubeidiya in the Southern Levant

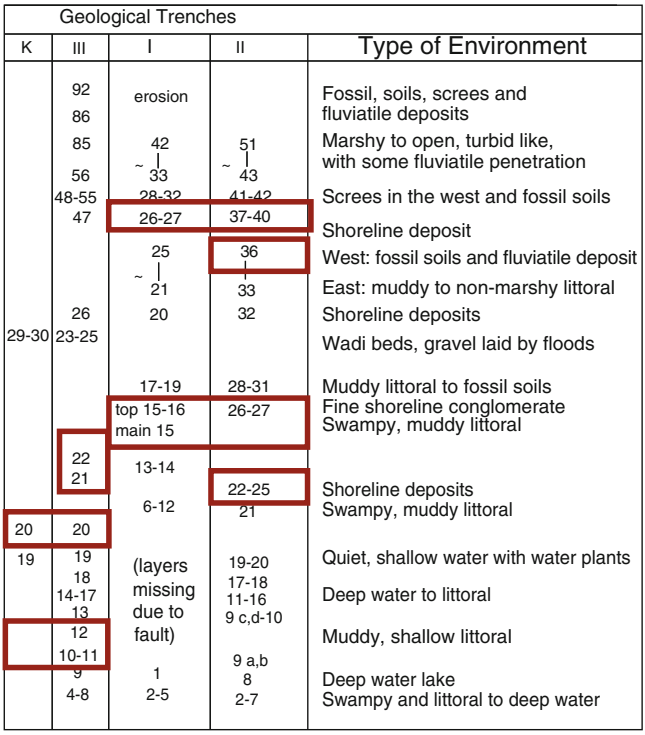


Fig. 2.2 Stratigraphic sequence of 'Ubeidiya

Table 2.1 Environmental gradient in relation to stratigraphic sequence

Strata	Ranked Stratum	Environment
III 11–13	7	5
III 20	6	4
III 21–23	5	4
II 23–25	4	3
II 26–27	3	3
II 36	2	1
II 37	1	1

2.3 The Statistical Model

Two criteria are important when we choose the (dis)similarity coefficient [Le98]:

1. The index used should be appropriate for the data. A presence–absence matrix can be transformed to a similarity with a binary coefficient, while for abundance data we need to use quantitative coefficients.
2. Double zeros: If a species is absent from two sites (double zero), we do not know if this is because it is truly absent or because of sampling bias. Thus,

it is preferable to exclude double zeros using an asymmetrical coefficient. The converse is a symmetrical coefficient, in which zeros are treated like any other value. In this study all coefficients used are asymmetrical.

It is beyond the scope of this study to discuss the numerous similarity indices that have been developed. The reader is referred to the work of Legendre [Le98] for an overview of these methods.

Here, we had three dissimilarity matrices: Two dependent variable matrices were calculated from multidimensional community structure, one using presence–absence data and the other using log-transformed relative abundance data. In addition, we had an explanatory (independent variable) vector: the *environment*.

The presence–absence community matrix was transformed into an asymmetrical binary similarity matrix by means of the Sneath and Sokal index [So63]:

$$S_{X_1,X_2} = \frac{a}{a + 2b + 2c}, \tag{2.1}$$

where a is the number of species common to both assemblages, b is the number of species present in assemblage one but absent from assemblage two, and c is the number of species present in assemblage two but absent from assemblage one.

The relative (log-transformed) abundance *community structure* was converted to a similarity matrix by means of the asymmetrical quantitative Gower coefficient [Go71]. This coefficient, used on normalized abundances, is defined by

$$S_{X_1,X_2} = \frac{\sum_{j=1}^{p} W_{12j} S_{12j}}{\sum_{j=1}^{p} W_{12j}}, \tag{2.2}$$

where $S_{12j} = 1 - [|y_{1j} - y_{2j}|/R_j]$.

The value W_j is called *Kronecker's data*, and its values are $W_j = 0$ when y_j is missing for either one of the objects or both, and $W_j = 1$ when information is present for both objects.

Both community structure similarity matrices were converted to distance matrices by means of the formula $d = \sqrt{1 - s}$, so each similarity index s was converted to a distance value d [Le98].

The *environment* vector (see Table 2.1) was converted to a dissimilarity matrix using the Euclidean distance [Le98]

$$D_{X_1,X_2} = \sqrt{\sum_{j=1}^{p} (Y_{1j} - Y_{2j}^2}. \tag{2.3}$$

Testing if there is a correlation between the similarity of sites structured in space or time, we come across a problem of autocorrelation. Specifically, if there are n objects that are temporally distinct and the matrix is symmetrical (so the time period spanning from object a to object b is the same as the time/distance from b to a), such a matrix contains $n(n-1)/2$ distances that are not independent, as changing the time of one object would change $n-1$ of these distances. Thus, we cannot assess the relationship between two matrices using the parametric correlation coefficient.

Mantel's test is a solution to this type of problem [Ma71] as it takes autocorrelation into account. It is a regression in which the variables are (dis)similarity matrices instead of raw data, and so it allows us to test hypotheses regarding the correlation between distances among objects in matrices X and Y.

The basic form of the Mantel's statistic is calculated [Le98] as

$$z_M = \sum_{i-1}^{n-1} \sum_{j=i+1}^{n} x_{ij} y_{ij}, \tag{2.4}$$

where i and j are row and column indices. This is based on the non-normalized Pearson product moment correlation coefficient [Di83]. However, the test is conditional on the dis(similarity) index used.

To standardize the Mantel's statistic, it was suggested [Le98] that we should take

$$r_M = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i=1}^{n} \left(\frac{x_{ij} - \bar{x}}{S_x}\right)\left(\frac{y_{ij} - \bar{y}}{S_y}\right), \tag{2.5}$$

where $d = \left[\frac{n(n-1)}{2}\right]$ is the number of distances in the upper triangle part of each matrix.

Mantel's r statistic is similar to the coefficient of linear correlation, known as Pearson's r statistic. To test rank-transformed data, a ranked Mantel's statistic, ranked M, may be computed by converting within-matrix rank distances into ranks before computing rM. The correlations are comparable to the nonparametric Spearman correlations r_s.

In this study, the *environment* vector is ordinal, therefore, we used the ranked M (rM) in lieu of r_s.

We employed one tailed hypothesis calculated for positive test statistics. To assess the significance of a departure from zero correlation, the rows and columns of one of the matrices are subjected to random permutations 10,000 times, with the statistic recalculated after each permutation. The significance of the observed statistic is the proportion of the permutations that lead to a higher correlation coefficient.

We can formulate two hypotheses: $H_0 : r_s = 0$ and $H_1 : r_s > 0$.

We define A to be the presence–absence matrix, B the relative abundance matrix, and C the environment matrix. We predict that both presence–absence and relative

abundance will change as a function of environmental change so that $r_s(AC) > 0$ and $r_s(AB) > 0$.

We are aware that multiple comparisons may increase the type I error of the statistics of significance for each comparison. In [Fe02] it is suggested that the use of adjusted p-values should be reconsidered since it increases the chance of making type II errors and requires an increase in sample size. The latter point is of particular importance in paleontological studies. Following the suggestions described in [Fe02], we present unadjusted p-values and combine the study's statistical significance with the magnitude of the effect, the quality of the study, and findings from other studies instead of adjusted p-values.

2.4 Results

Correlating relative abundance with the local environment change observed in 'Ubeidiya resulted in a significant correlation between the two variables; specifically, $r_s = 0.543$ and $p = 0.005$. This would suggest that the faunal community changed over time due to the environmental change observed by the geomorphological analysis.

However, contrary to expectations, there is no correlation between *presence–absence community structure* and *environment* ($r_s = 0.295$, $p = 0.121$). This suggests a pattern of inertia in community presence–absence across the $4x10^5$ years represented by the sequence in 'Ubeidiya (Figure 2.3).

2.5 Discussion and Conclusion

The question of identification of stasis or change in community structure has implication for understanding the tempo and mode of ecological and evolutionary processes. The site of 'Ubeidiya is dated within this time period of the "Pleistocene paradox." It is younger than most sites that exhibit coordinated stasis (greater than circa. 100 Ma), yet older than the glacial Pleistocene sites (that is, less than 0.8 Ma).

Applying (dis)similarity indices and the Mantel's test, we concluded that *relative abundance community structure* correlated with *environment*. A detailed observation of species distribution throughout the sequence shows that fallow deer *Dama* sp., row deer *Capreolus* sp., the large extinct deer *Praemegaceros obscurus*, and the North Africa ass *Equus tabeti* shift their abundance among the strata. Over time, there is a shift from a high to low proportion of woodland taxa (fallow and roe deer) with a concomitant increase in open grassland taxa (the *Praemegaceros* and ass). Thus, changes in the large mammalian fauna in 'Ubeidiya reflect a local environmental shift towards greater aridity [Be06].

In contrast to the pattern of relative abundance, the *presence–absence community structure* did not correlate with environmental change, which implies inertia over the

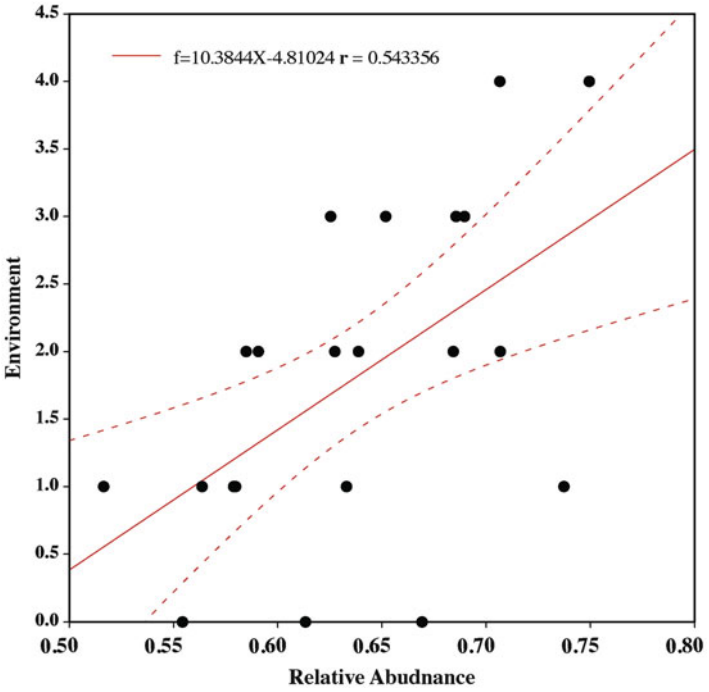


Fig. 2.3 Cluster diagram showing the correlation between similarity in large mammal relative abundance and similarity between environmental gradient

entire sequence. This persistence in species incidence across the 'Ubeidiya sequence is interpreted in accordance with the "recurrent assemblages" model suggested in [Mi93] for paleontological assemblage and supported for modern ecological communities [Ra90].

The response of taxa to climatic shift is dependent on the amplitude and frequency of climatic change [Ra90]. During low and medium amplitudes of environmental shift, taxa may be able to tolerate the change. Thus, despite independent evidence for a climatic shift, no change is observed in the fossil record. In higher amplitudes of environmental shift, taxa will shift their range. This is often observed in the fossil record as a change in abundance. In very high amplitudes of climatic shift, taxa will become extinct. This may be observed in the fossil record as faunal turnover. Relative frequency of species may fluctuate, whereas species presence–absence may remain constant over time [Ra90].

Consequently, the different pattern of turnover and stasis for species presence–absence and relative abundance, which are apparent in 'Ubeidiya, may be attributed to persistence that occurred during periods of low amplitude environmental change. While species changed in their relative abundance, the amplitude of climate change was not enough to evoke change in species presence–absence pattern. A similar pattern was found in the middle Pleistocene site of Atapuerca, Spain [Ro11], which

may be attributed to medium amplitudes of climatic change, large enough to result in faunal turnover but sufficiently low to maintain a similar ecological structure of the community.

The use of similarity indices and the Mantel's test allows us to illustrate the utility of similarity matrices in the study of paleontological community structure in time.

References

[Ba93] Bar-Yosef, O., Goren-Inbar, N.: The Lithic Assemblages of 'Ubeidiya, a Lower Paleolithic Site in the Jordan Valley. The Institute of Archeology/The Hebrew University of Jerusalem, Jerusalem (1993)

[Be02] Belmaker, M., Tchernov, E., Condemi, S., Bar-Yosef, O.: New evidence for hominid presence in the Lower Pleistocene of the Southern Levant. J. Hum. Evol. **43**, 43–56 (2002)

[Be06] Belmaker, M.: Community structure through time: 'Ubeidiya, a Lower Pleistocene site as a case study. Ph.D. Dissertation, The Hebrew University of Jerusalem (2006)

[Di83] Diets, E.J.: Permutation tests for association between two distance matrices. Syst. Zool. **32**, 21–26 (1983)

[Fe02] Feise, R.J.: Do multiple outcome measures require p-value adjustment? BMC Med. Res. Methodol. **2**, 8–12 (2002)

[Go71] Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics **27**, 857–871 (1971)

[Ho96] Holterhoff, P.F.: Crinoid biofacies in Upper Carboniferous cyclothems, midcontinent North America: faunal tracking and the role of regional processes in biofacies recurrence. Palaeogeogr. Palaeoclimatol. Palaeoecol. **127**, 47–81 (1996)

[Le98] Legendre, P., Legendre, L.: Numerical Ecology. Elsevier, Amsterdam (1998)

[Ma06] Mallol, C.: What's in a beach? Soil micromorphology of sediments from the Lower Paleolithic site of 'Ubeidiya, Israel. J. Hum. Evol. **51**, 185–206 (2006)

[Ma71] Mantel, N.: The detection of disease clustering and a generalized regression approach. Cancer Res. **27**, 209–220 (1967)

[Mi93] Miller, W.: Models of recurrent fossil assemblages. Lethaia **26**, 182–183 (1993)

[Ra90] Rahel, F.J.: The hierarchical nature of community persistence: a problem of scale. Am. Nat. **136**, 328–344 (1990)

[Ro11] Rodríguez, J., Burjachs, F., Cuenca-Bescós, G., García, N., Van der Made, J., Pérez González, A., Blain, H.-A.: One million years of cultural evolution in a stable environment at Atapuerca (Burgos, Spain). Quat. Sci. Rev. **30**, 1396–1412 (2011)

[So63] Sokal, R.R., Sneath, P.H.A.: Principles of Numerical Taxonomy. W.H. Freeman, San Francisco (1963)

[St01] Sterelny, K.: The reality of ecological assemblages: a palaeo-ecological puzzle. Biol. Philos. **16**, 437–461 (2001)

Chapter 3
The Method of Superposition for Near-Field Acoustic Holography in a Semi-anechoic Chamber

D.J. Chappell and N.M. Abusag

3.1 Introduction

Near-field acoustic holography (NAH) is the process of reconstructing the vibrational behaviour of a structure from measurements of the acoustic field generated by these vibrations. Traditionally NAH was applied to planar regions where Fourier methods can be used to reconstruct the structural vibrations, even at frequencies beyond the sampling resolution limit [Wi99]. The Method of Superposition (MoS) and the inverse boundary element method (IBEM) are relatively popular alternative methods for reconstructing the vibrational properties of structures with more general geometries [VaWi06, ChHa09]. Recent work has shown that the MoS can also be effectively combined with sparse ℓ_1 regularisation to generate solutions using only a small number of terms in the superposition [AbCh16]. In this work we discuss a reformulation of the MoS for NAH experiments in a semi-anechoic chamber; experiments in fully anechoic chambers can often prove impractical. In particular, we propose a modified Green's function approach for a semi-infinite domain with a hard reflecting boundary using the Method of Images, and present the results of some supporting numerical experiments.

3.2 Method of Superposition

Consider a three-dimensional half-space of the form $H = \{\mathbf{x} \in \mathbb{R}^3 : x_3 > 0\}$, and let $\Omega \subset H$ be a finite domain with boundary surface $\Gamma \subset \bar{H}$. In order to help visualise the set-up, one can think of H as the space represented by a semi-anechoic chamber

D.J. Chappell (✉) • N.M. Abusag
Nottingham Trent University, Nottingham, UK
e-mail: david.chappell@ntu.ac.uk; nadia.abusag2007@my.ntu.ac.uk

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_3

with a rigid floor and fully absorbing walls and ceiling that behave approximately as an infinite half-space. The object Ω represents an acoustically radiating object placed in H, such as a loudspeaker. We decompose Γ into two parts, $\Gamma_H \subset H$ and $\Gamma_0 = \Gamma \cap \{\mathbf{x} \in \mathbb{R}^3 : x_3 = 0\}$ so that $\Gamma = \Gamma_0 \cup \Gamma_H$ and $\Gamma_0 \cap \Gamma_H = \emptyset$. In a practical setting, Γ_0 corresponds to the part of the acoustically radiating object that is in contact with the floor of the semi-anechoic chamber, and is thus assumed to be non-empty. Let $\Omega_+ = H \backslash \bar{\Omega}$ denote the unbounded domain exterior to the object Ω, which is assumed to be filled with a homogeneous compressible acoustic medium with density ρ and speed of sound c. For a time-harmonic disturbance of frequency ω, the sound pressure p satisfies the homogeneous Helmholtz equation in Ω_+

$$\Delta p + k^2 p = 0, \tag{3.1}$$

where $k = \omega/c$ is the wavenumber. Since this work considers an unbounded domain within the half-space H, then p must also satisfy the Sommerfeld radiation condition

$$\lim_{R \to \infty} R \left\{ \frac{\partial p}{\partial R} - ikp \right\} = 0 \tag{3.2}$$

for $\mathbf{x} \in H$, with $R = \|\mathbf{x}\|_2$.

The superposition method approximates p at some point $\mathbf{x} \in \bar{\Omega}_+$ using a basis expansion of the form

$$p(\mathbf{x}) \approx \sum_{j=1}^{n} \sigma_j G_H(\mathbf{x}, \mathbf{y}_j), \tag{3.3}$$

where G_H is the half-space Green's function for Helmholtz equation in three dimensions given by

$$G_H(\mathbf{x}, \mathbf{y}) = \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|} + \frac{e^{ik|\mathbf{x}-\mathbf{y}'|}}{4\pi|\mathbf{x}-\mathbf{y}'|}. \tag{3.4}$$

Here $\mathbf{y}' = (y_1, y_2, -y_3)$ corresponds to the reflection of the point $\mathbf{y} = (y_1, y_2, y_3) \in \Omega$ in the plane $\partial H = \{\mathbf{x} \in \mathbb{R}^3 : x_3 = 0\}$. The points $\mathbf{y}_i \in \Omega$, $i = 1, \ldots, n$ are the source locations and σ_i are the source strengths, which are determined by application of the method. Note that the half-space Green's function G_H corresponds to the Neumann Green's function with

$$\frac{\partial G_H}{\partial y_3}(\mathbf{x}, \mathbf{y}) = 0$$

whenever $\mathbf{y} \in \partial H$. Hence, the Green's function G_H satisfies the rigid floor boundary condition as a function of the second variable.

3.3 Near-Field Acoustic Holography in a Half-Space

In the NAH problem we are given values of the acoustic pressure p at a discrete set of points in the acoustic near field within Ω_+. We will assume that the data points $\mathbf{x}_i$, $i = 1, \ldots, m$ lie on a surface $\Gamma^* \subset \Omega_+$. Note that the pressure data is usually obtained from measurements using a microphone array. However, in this work we only generate the pressure data numerically as described in Section 3.5. The NAH problem in the half-space H is to use the given pressure data to recover the Neumann boundary data on Γ_H. Solving this problem via the method of superposition is then a matter of finding the set of source strengths σ_j, $j = 1, \ldots, n$, that reproduce the acoustic pressure data to some desired accuracy in the least squares sense. That is, σ_j are chosen so that the ℓ_2 norm of the residual vector $\mathbf{r}$, with entries given by

$$r_i = p(\mathbf{x}_i) - \sum_{j=1}^{n} \sigma_j G_H(\mathbf{x}_i, \mathbf{y}_j) \tag{3.5}$$

for $i = 1, .., m$, is smaller than a desired error tolerance. Once the source strengths have been obtained, then the Neumann boundary data can be recovered from

$$\frac{\partial p}{\partial \mathbf{n}}(\mathbf{x}) \approx \sum_{j=1}^{n} \sigma_j \frac{\partial G_H}{\partial \mathbf{n}}(\mathbf{x}, \mathbf{y}_j), \tag{3.6}$$

where $\mathbf{n}$ is the outward unit normal to Γ.

Regularisation is always required in general, even for $n = m$, since NAH is an ill-posed inverse problem (see, for example, [ChHa09]). For experimental problems, the pressure measurements will contain errors and the ill-posedness of the problem means that these errors are amplified in the (unregularised) solutions. In the next section, we describe a regularisation scheme designed to promote sparsity in the solution as suggested in [ChEtA12] for two-dimensional planar problems, and in [AbCh16] for three-dimensional problems.

3.4 Regularisation and Sparse Reconstruction

Here we give a brief presentation of the strategy employed in [AbCh16], for more details, see [AbCh16] and [ChEtA12]. The sparse regularisation approach is designed to minimise $|\boldsymbol{\sigma}|_0$, the number of non-zero entries of $\boldsymbol{\sigma}$, for a fixed acceptable discrepancy level indicated by $\|\mathbf{r}\|_2$. As noted in [ChEtA12], the possibility of a sparse reconstruction is highly dependent on the basis functions used to represent the solution. In the superposition method, these basis functions are the fundamental solution of the Helmholtz equation at a set of distinct interior charge points.

Directly minimising $|\boldsymbol{\sigma}|_0$ is often intractable because of non-convexity (see [ChEtA12]). We therefore instead seek to minimise the ℓ_1 norm

$$\|\boldsymbol{\sigma}\|_1 = \sum_j |\sigma_j|. \tag{3.7}$$

The use of the ℓ_1 norm allows one to apply powerful convex optimisation algorithms and still promotes sparsity by making many of the components of $\boldsymbol{\sigma}$ negligibly small, meaning that they can be well approximated by zero without degrading the reconstructed solution. The following procedure will be applied to find a sparse representation $\hat{\boldsymbol{\sigma}}$ of the source strengths $\boldsymbol{\sigma}$

$$\hat{\boldsymbol{\sigma}} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\sigma}\|_1 \quad \text{subject to} \quad \|\mathbf{r}\|_2^2 \leq \epsilon. \tag{3.8}$$

This procedure, which will be implemented using the convex optimisation toolbox CVX [GrBo15], requires a data fidelity constraint ϵ to be specified. Choosing the parameter ϵ involves a trade off between allowing sparser solutions with larger values of ϵ and achieving more accurately reconstructed solutions with smaller values of ϵ. A good choice of ϵ will depend on how noisy the pressure data is and hence will be problem dependent.

3.5 Numerical Results

Numerical results will be computed for acoustic radiation from a cuboid of similar dimensions to a typical loudspeaker cabinet (0.28m*times*0.28m $\times$ 0.42m). The base of the cuboid Γ_0 lies in the plane $z = 0$. Although the method of superposition is a mesh free method, we will use a triangulation of Γ_H to generate the points at which the pressure data is computed, as well as the internal charge points and the points at which we reconstruct the solution on Γ_H. In particular, for a given triangulation of Γ_H we reconstruct the Neumann boundary data at the centroid of each triangle and project (from each centroid) a distance δ along the normal vector to Γ into Ω_+ to obtain the points where the exterior pressure data is recorded. The internal charge points are positioned inside Ω, on a scaled down version of Γ_H with scaling factor $\alpha \in (0, 1)$. For example, a value of $\alpha = 0.5$ corresponds to a surface of internal charge points whose dimensions are exactly half those of Γ_H.

We will reconstruct the boundary data generated by a point source at $\mathbf{x}_0 \in \Omega$. The pressure data is then constructed using the half-space Green's function, and is hence of the form

$$(\mathbf{p}_0)_j = a\left(\frac{e^{ik|\mathbf{x}_j - \mathbf{x}_0|}}{|\mathbf{x}_j - \mathbf{x}_0|} + \frac{e^{ik|\mathbf{x}_j - \mathbf{x}_0'|}}{|\mathbf{x}_j - \mathbf{x}_0'|}\right), \qquad j = 1, \ldots, m. \tag{3.9}$$

Here, $a \in \mathbb{C}$ is the strength of the source, which in these examples is arbitrarily taken to be $a = 3 - i$. The boundary data generated at $\mathbf{y} \in \Gamma_H$ may also be obtained for the case of a point source at $\mathbf{x}_0$ by replacing $\mathbf{x}_j$ in (3.9) by $\mathbf{y} \in \Gamma_H$, differentiating in the direction of $\mathbf{n}_y$ and evaluating at the centroids of the triangulation $\mathbf{y} = \mathbf{y}_j$ for $j = 1, \ldots, m$ to give

$$\begin{aligned}\frac{(\mathbf{v})_j}{a} =& \frac{\mathbf{n}_{y_j} \cdot (\mathbf{x}_0 - \mathbf{y}_j)}{|\mathbf{y}_j - \mathbf{x}_0|^3}(1 - ik|\mathbf{y}_j - \mathbf{x}_0|)e^{ik|\mathbf{y}_j - \mathbf{x}_0|} \\ &+ \frac{\mathbf{n}_{y_j} \cdot (\mathbf{x}_0' - \mathbf{y}_j)}{|\mathbf{y}_j - \mathbf{x}_0'|^3}(1 - ik|\mathbf{y}_j - \mathbf{x}_0'|)e^{ik|\mathbf{y}_j - \mathbf{x}_0'|}.\end{aligned}$$

Using this calculation it is possible to verify the accuracy of the regularised approximate solutions with different wavenumbers and point source positions $\mathbf{x}_0 \in \Omega$. We will also investigate the behaviour of the method at irregular frequencies of the volume enclosed by the interior charge points, and the dependence on the dimensions / location of the interior charge point surface controlled by the parameter α.

Uniformly distributed and additive white noise will be applied to $\mathbf{p}_0$ in order to more closely replicate experimental observations. The use of Gaussian noise was also considered and, in general, led to slightly more accurate reconstructions than uniformly distributed noise. However, the quality of the reconstructions also fluctuated more widely when using different Gaussian noise vectors (of the same norm) than for uniformly distributed noise, and so we present the results for uniformly distributed noise since we believe they give a more indicative and repeatable measure of the performance of our reconstruction methods. We denote the added noise vector as $\mathbf{w}$ and specify the ratio

$$w = \frac{\|\mathbf{w}\|_2}{\|\mathbf{p}_0\|_2}, \tag{3.10}$$

referring to w as the level of added noise in the sequel.

For our sparse reconstruction method, we use the following criteria to determine whether the jth charge point is dominant [AbCh16]:

$$\log\left(\frac{|\sigma_j|}{\min_i |\sigma_i|}\right) > \beta \log\left(\frac{\max_i |\sigma_i|}{\min_i |\sigma_i|}\right).$$

We will use the notation $N^*(\beta)$ for the number of dominant charge points satisfying this condition, taking $\beta = 0.5$ by default and so we denote $N^* = N^*(0.5)$. The ℓ_2 percentage error in the reconstructed solution $\hat{\mathbf{v}}$ will be calculated using

$$\frac{\|\hat{\mathbf{v}} - \mathbf{v}\|_2}{\|\mathbf{v}\|_2} \times 100\%. \tag{3.11}$$

For these experiments the number of charge points, the number of measurement points and the number of points at which we reconstruct the solution are all equal to 504. This is achieved by triangulating the internal source surface in an identical way to Γ_H and taking the charge points at the triangle centroids. The data fidelity parameter ϵ appearing in Equation (3.8) is chosen as

$$\epsilon = (\max\{\epsilon_{\min}, w\})^2 \|\mathbf{p}_0\|_2^2, \tag{3.12}$$

where w is the level of noise added to the pressure data as before. Larger choices of ϵ permit sparser solution representations. However, it only makes sense to choose a larger ϵ for noisy data, otherwise it leads to less accurate reconstructions. The parameter $\epsilon_{\min} \geq 0$ is included as a tolerance level that is used for the low or zero noise case. A relatively large choice of $\epsilon_{\min}$ will lead to sparser reconstructions at the expense of accuracy, and the converse is true for small $\epsilon_{\min}$. The results in this work have been obtained with $\epsilon_{\min} = 1\text{E-}6$.

First consider the case $k = 1$ and $\mathbf{x}_0 = (0, 0, 0.1)$, where the frequency is relatively low, is not close to an irregular frequency and $\mathbf{x}_0$ is relatively close to the origin and will lie inside the surface on which the interior charge points are located. Under such conditions the superposition method is expected to work well. The pressure data are specified at a distance $\delta = 0.035$m from Γ_H and the internal source surface is scaled down to have dimensions $\alpha = 1/3$ the size of Γ. We note that these choices should lead to good results based on the fact that δ should be chosen small enough to capture evanescent contributions to the pressure field, but still large enough to be a practical distance for taking experimental measurements. For the choice of the parameter α, we observed in [AbCh16] that too small a value will lead to severe ill conditioning as the charge points become very close together, but choosing too large a value of α will also give poor results, and a choice in the range $\alpha \in (0.1, 0.6)$ generally gives the best results.

Figure 3.1 shows the sparse reconstruction of the Neumann data with noise level $w = 5\%$. The exact solution is also shown for reference and appears almost identical to the sparse reconstruction. The right sub-plot shows the charge point strengths σ_j, $j = 1, \ldots, 504$ given by the sparse reconstruction algorithm. Note that many of the $\sigma_j, j = 1, \ldots, 504$ are suppressed and are close to $O(10^{-6})$, but 13 dominant terms can be picked out which are close to $O(10^{-1})$. The sparse reconstruction shown in the left sub-plot was created using only these 13 values.

We now investigate the behaviour of the method for some potentially problematic choices of the wavenumber k when $w = 15\%$ noise is added to the sampled pressure data. First we look at the case when the frequency is increased, including when the Nyquist frequency is exceeded. Since our measurements are taken at triangle centroids then the resulting measurement grid is irregular and so the Nyquist frequency is not well-defined. We therefore choose the Nyquist frequency associated with the regular grid given by the triangle vertices, as a value approximately representative of the Nyquist frequency. For the discretisation considered in the previous section with 504 triangles, the grid spacing is $\Delta x = 0.04667$, meaning that the wavenumber corresponding to the Nyquist frequency is $k_{\text{nyq}} = \pi/\Delta x = 67.32$.

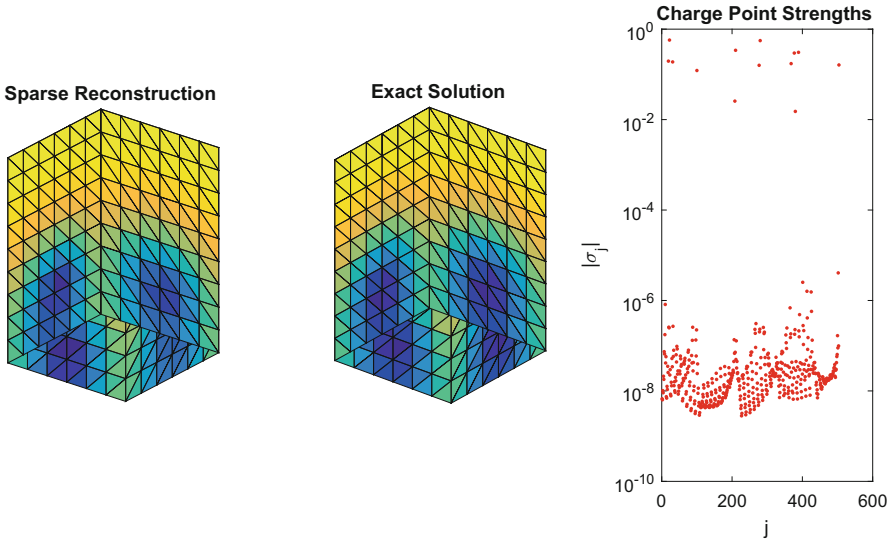


Fig. 3.1 Neumann boundary data on a cuboid generated by a point source at $\mathbf{x}_0 = (0, 0, 0.1)$ with wavenumber $k = 1$ and $w = 5\%$ added noise. The plots compare the exact solution against the ℓ_1 reconstruction approach using only the $N^* = 13$ dominant charge points of largest magnitude shown in the right sub-plot

We also investigate the performance of the method close to other typical threshold frequencies for numerical solution approaches, such as the six grid points per wavelength rule of thumb for finite and boundary element methods, which gives a maximum wavenumber of $k = 22.44$ for the grid described above. The performance of the method at irregular frequencies will also be investigated. For the method of superposition these irregular frequencies are the resonances of the region enclosed by the interior source surface. We set $\alpha = 1/3$; numerical studies indicate that in this case, one such frequency approximately corresponds to the wavenumber $k = 48.907$. The maximum wavenumber studied corresponds to the wavelength being close to (but still greater than) the exterior measurement distance $\delta = 0.035$m.

The left plot of Figure 3.2 shows that both irregular and high frequencies lead to a degradation in the accuracy of the reconstruction, and lead to a loss of sparsity in the reconstructions. We note that accurate and reasonably sparse reconstructions can be generated up to the Nyquist frequency $k = 67.32$ (except for at characteristic frequencies), since we can reconstruct the solution with a smaller error than the level of added noise (15%) and using only around 10% or less of the total number (504) of charge points. We note that if the surface of interior charge points includes the monopole generating the acoustic field then one would obtain exact representations for arbitrarily high frequencies.

In addition to the general trend of increased errors for higher frequencies, one also observes a local peak in the error at the characteristic wavenumber $k = 48.907$. This suggests that sparse reconstructions are not feasible at higher frequencies or

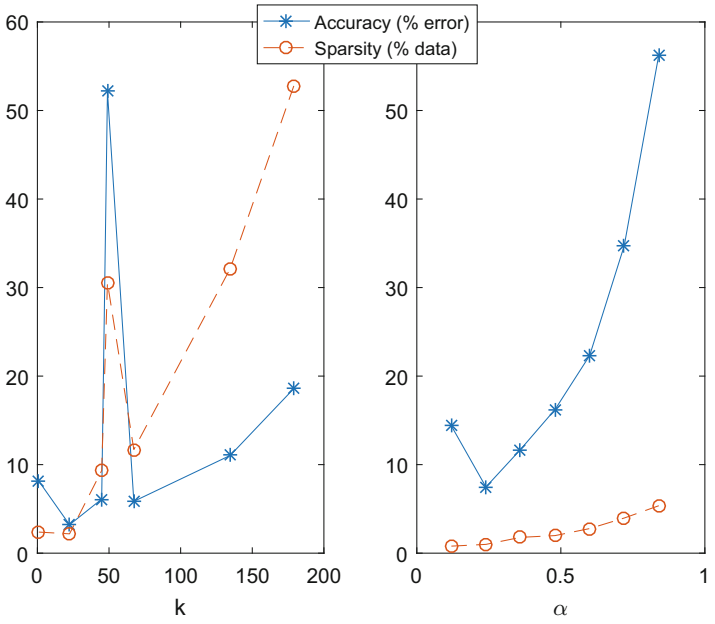


Fig. 3.2 The accuracy and sparsity of the reconstructed solutions with $w = 15\%$ added noise and with $\mathbf{x}_0 = (0, 0, 0.1)$ as before. The solid lines with star markers indicate ℓ_2 percentage (relative) errors. The dashed lines with circle markers indicate the percentage of the 504 charge points used in the reconstruction. Left: The plots show the effect of changing the wavenumber k with a fixed interior charge point surface corresponding to $\alpha = 1/3$. Right: The plots show the effect of using a range of different sized interior charge point surfaces controlled by the parameter α for a fixed wavenumber $k = 1$

at irregular frequencies. However, the reconstruction error is lower than the noise level for all frequencies tested up to twice the Nyquist frequency. The results of this section therefore suggest that the method of superposition with ℓ_1 regularisation can provide excellent reconstructions for frequencies up to around twice the Nyquist frequency, and that sparse reconstructions are feasible up to the Nyquist frequency. Irregular frequencies degrade both accuracy and sparsity. However, if a more accurate and sparsely reconstructed solution was required at $k = 48.907$, then we could change the scaling of the internal source surface (i.e. change α), which would move the location of the irregular frequency as demonstrated in [AbCh16].

We now consider how the accuracy and sparsity of our reconstructed solutions depends on the relative size/position of the internal charge point surface controlled by the parameter α. The source point generating the external pressure data is taken at $\mathbf{x}_0 = (0, 0, 0.1)$ as before, and the wavenumber is fixed to be $k = 1$. The right plot of Figure 3.2 shows both the percentage errors for the sparse reconstructions and the percentage of the 504 charge points used in the reconstruction for different sized interior charge point surfaces. These quantities have been computed for values of α between 0.12 and 0.84. In all cases the added noise level is 15%. We notice that

the error is minimised when the size of the interior source surface is such that it intersects the positive $z-$axis close to $z = 0.1$, where the source point generating the external pressure data is located. This corresponds to the choice $\alpha = 0.24$, since Γ_H intersects the positive z-axis at $z = 0.42$ and $z = 0.24 \times 0.42$ is very close to $z = 0.1$. Likewise, the number of charge points N^* needed to obtain a sparse reconstruction is also minimal close to $\alpha = 0.24$.

In general, the solutions are reasonably accurate (i.e. the reconstruction error is comparable to or less than the data error) for source surfaces with α between 0.12 and 0.48. Choosing $\alpha = 0.84$ gave the worst results. Interestingly, the results of this section suggest that it does not seem to be critical whether or not the surface of interior charge points encloses any singularities in the modelled wave field. Furthermore, the results also point to important potential applications of the sparse superposition method developed in this work for source identification problems in general.

3.6 Conclusions

The method of superposition has been combined with a sparse ℓ_1 reconstruction algorithm and applied to the problem of near-field acoustic holography in a half-space. The developed sparse superposition method is able to reconstruct the normal velocity of a vibrating object using only a very small number of charge points in many cases, and is suitable for experimental verification in a semi-anechoic chamber with a hard reflecting floor.

Acknowledgements N.M. Abusag gratefully acknowledges financial support from the Libyan Ministry of Higher Education and Scientific Research. D.J. Chappell gratefully acknowledges financial support from the European Union (FP7-PEOPLE-2013-IAPP grant no. 612237 (MHiVec)).

References

[AbCh16] Abusag, N.M., Chappell, D.J.: On sparse reconstructions in near-field acoustic holography using the method of superposition. J. Comput. Acous. **24**(3), 1650009 (2016)

[ChEtA12] Chardon, G., Daudet, L., Peillot, A., Ollivier, F., Bertin, N., Gribonval, R.: Nearfield acoustic holography using sparsity and compressive sampling principles. J. Acoust. Soc. Am. **132**(2), 1521–1534 (2012)

[ChHa09] Chappell, D.J., Harris, P.J.: A Burton-Miller inverse boundary element method for near-field acoustic holography. J. Acoust. Soc. Amer. **126**(1), 149–157 (2009)

[GrBo15] Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1, (2015). http://cvxr.com/cvx, last visited 06/02/2015

[VaWi06] Valdivia, N.P., Williams, E.G.: Study of the comparison of the methods of equivalent sources and boundary element methods for near-field acoustic holography. J. Acoust. Soc. Am. **120**(6), 3694–3705 (2006)

[Wi99] Williams, E.G.: Fourier Acoustics. Academic, London (1999)

Chapter 4
Application of Stochastic Dynamic Programming in Demand Dispatch-Based Optimal Operation of a Microgrid

F. Daburi Farimani and H. Rajabi Mashhadi

4.1 Introduction

Uncertainty is an inseparable feature of power systems due to the general uncertainty of the system, uncertain behavior of power consumers, uncertainty of renewable energy resources, and uncertain prices of power market. In power system studies, several uncertainty modeling approaches have been utilized. These methods are categorized into probabilistic methods, possibilistic methods, combined probabilistic and possibilistic methods, and information gap theory. Probabilistic methods are divided into numerical approaches and analytical approaches. Sequential Monte Carlo simulation, non-sequential Monte Carlo simulation, and pseudo-sequential Monte Carlo simulation are numerical probabilistic approaches. Analytical approaches apply mathematical expressions like PDFs to analyze the system and its inputs. Analytical approaches are categorized into two groups. First, linearization-based methods like convolution method, Cumulant method, Gram-Charlier A series, Edgeworth expansion, Cornish-Fisher expansion, Taylor series, and first-order second-method. Due to shortcomings of linearization in the aforementioned methods, the second group methods of analytical approaches are applied. These methods are based on PDF approximation. The point estimation method, Unscented Transformation method, and scenario-based decision making methods are examples of such methods [AiEtAl16]. Categorization of possibilistic methods and combined probabilistic and possibilistic methods are not mentioned due to the focus on probabilistic analytic methods.

Power system operation is one of the most essential programs in power system studies. Power system structural changes and smart grids in recent decades made

F.D. Farimani • H.R. Mashhadi (✉)
Faculty of Engineering, Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: fateme.daburifarimani@stu.um.ac.ir; h_mashhadi@um.ac.ir

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_4

many fundamental changes in distribution system and especially in demand side. In conventional power systems, conventional generation units which are dispatchable, follow nondispatchable conventional loads, through supply dispatch program, based on load following, actually, a real time control of output of power plants to follow system load by several control loops. Today the power system grids in the world are moving toward smart grids. Integration of smart homes, smart meters, smart microgrids, and penetration of renewable generation systems and plugged in electric vehicles in the new and future power systems will be remarkable. Power system structural changes like deregulation of market, highly penetrated DGs, and renewable resources make some challenges in the power system operation. Smart grids provide the required infrastructure of a new optimal operation paradigm in demand side. According to these changes and challenges, it seems to be better to change the operation paradigm from the conventional economic dispatch (ED) to demand dispatch. DD is Remote Control of Dispatchable loads by the grid operator. By dispatchable loads we mean EVs, air conditioners, washing machines, dishwashers, dryers, water heaters which are suitable to be remotely dispatched from the microgrid dispatching center. These DLs participate in DD by giving their commitment period. A standby (waiting) period, which they dedicate their plugged-in appliance to the operator to draw power from the grid. This is actually Load Commitment by giving waiting period.

DD was firstly introduced in [BrEtAl10] by the help of smart grid and development of communication and control technologies in demand side. In the new power systems, loads might be equipped with communication and control technologies and remotely receive the dispatch/control command from the operator. In [BrEtAl10], DD is compared with demand response. Moreover, a precise definition of dispatchable loads is presented. The electric vehicles are aggregated by the aggregator to provide ancillary services like frequency regulation as an example of DD. In [BoEtAl13] unit commitment is reformulated considering DD and a powerful probabilistic wind power forecasting method to handle the uncertainty of wind power. It is illustrated that DD improves reserve requirement and diminishes load and wind power curtailment. Reference [BeEtAl11] suggested a generic formulation for economic dispatch of three buildings along with DR purposes from the perspective of end user without considering topology and constraints of distribution system. In [DaRa13], DD is employed on an autonomous hybrid PV-wind-battery-diesel system. It is generally concluded that DD improves the required battery capacity and diesel generation capacity since the dispatchable loads contribute in load generation balance. Smart charging of electric vehicles by applying DD is performed in [WuEtAl12]. A priority list algorithm for implementation of DD is suggested in [DaRa13] which provides a high correlation between small wind turbines power and dispatchable loads and thus, reduces the total operation cost. A comprehensive report on DD is prepared by DOE/NETL [NETL11] including the definition of DD, a comparison between DD and supply dispatch, the benefits and barriers of implementing DD. In [DaRa15], the DD problem is accurately modeled in details and implemented on a smart microgrid. But it has not taken into account the probabilistic behavior of the components.

DD problem could be actually modeled by optimal control of a discrete-time dynamic system with an additive cost function over a finite horizon called inventory control problem. Our problem is to optimally operate a microgrid by using DD applying analytic probabilistic method. The operator of the microgrid should make optimal decisions in stages, dealing with stochastic situations. The objective is to minimize the total cost of the microgrid operation. The decisions should be made in a way that consider the future costs besides the present cost. So we should not view the decisions individually. The stated idea is captured in dynamic programming whereby at every stage a decision is made that minimizes the sum of the current stage cost and the best expected cost of the future stages [Di87]. Since the problem is faced with uncertainties in generation and consumption, stochastic dynamic programming (SDP) is used to solve the problem.

In this research for the first time we seek to present DD problem through clear formulations based on SDP of dispatchable loads and apply it on some case studies step by step. A broadly applicable model of stochastic optimal control of a dynamic system over a finite number of stages is applied to model DD problem. If the operation horizon, here a day, is divided by N time periods, the operator tries to find the chain of optimal control law so as to minimize the expected total cost. DD problem is very close to the popular example of SDP called inventory control problem [Di87]. This paper firstly presents a brief description to the problem in Section 4.2. After that in Section 4.3 the SDP is introduced briefly. Inventory control problem is presented in Section 4.4. In Section 4.5, problem formulation by SDP (inventory control model) is presented. The solution approach is performed step by step in Section 4.6. The summary and conclusion are presented in the last part of the chapter.

4.2 Problem Description

Consider a stand-alone microgrid consisting of the following components:

1- Wind turbine
2- Diesel generator
3- Storage system (battery)
4- Dispatchable/controllable loads (CL) equipped by smart meters, and undispatchable loads (UL).

Figure 4.1 illustrates the microgrid components. The capabilities of the microgrid operator are: (1) receiving the consumer data about energy consumption scheduling from the smart meters, and (2) sending turn on control signal to the consumers through smart meters. The microgrid operator goal is to minimize the operation cost over the operation horizon, here a day. The operation cost over N stages in the period is formulated in (4.1) by ignoring the power loss. Energy balance constraint is presented in (4.2). Diesel generation limitations are mentioned in (4.3). Dispatchable loads constraints are presented in (4.4). Storage system

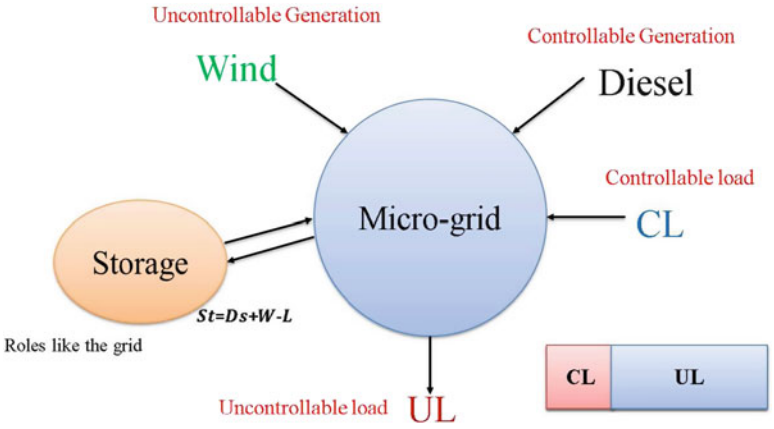


Fig. 4.1 The microgrid components

constraints which are storage evolution constraint, storage limitation, and charge constraint are mentioned in (4.5)–(4.7). X_k is energy storage at the end of period k. The microgrid operator decides what components to be dispatched according to the state of the renewable generations, charge state of the storage, and the grid constraints to minimize the expected total cost of the microgrid operation. By dispatchable components (DICOM) in the microgrid we mean dispatchable loads, storage, and diesel generator. The day is divided into 4 time intervals. The consumers submit their day-ahead consumption schedule to the operator.

$$Min_{Ds} \sum_{k=0}^{N-1} C(Ds_k) + Terminalcost \tag{4.1}$$

$$Ds_k + W_k - St_k - CL_k - UL_k = 0 \tag{4.2}$$

$$Ds_{min} \leq Ds_k \leq Ds_{max} \tag{4.3}$$

$$\sum_{k=1}^{N} CL_k = CL_{max} \tag{4.4}$$

$$X_{k+1} = X_k + St_{k+1} \tag{4.5}$$

$$0 \leq X_k \leq St_{max} \tag{4.6}$$

$$-X_k \leq St_{k+1} \leq (St_{max} - X_k) \tag{4.7}$$

4.3 Stochastic Dynamic Programming

DP is a very useful tool in situations where decisions are made in stages and have to be viewed in the whole trajectory instead of individually. The goal is to minimize the total cost of the stages. To model a problem with DP, two main features are needed: (1) a discrete-time system and (2) an additive cost function. The dynamic system form is presented in (4.8). The total cost over the period is presented in (4.9). First term represents terminal cost and second term is sum of costs in stages. Since w_k is a random variable, the cost will be a random variable, and minimization of a random variable is not meaningful. So, the expected cost should be minimized as presented in (4.10) [Di87]. In deterministic DP problems, u_k chain is obtained at the beginning of the period for all stages. But for SDP problems, the u_k of every stage is obtained when we reach to that stage. A popular example of DP is the inventory control problem which is of interest to this research.

$$X_{k+1} = f_k(x_k, u_k, w_k) \qquad \forall k = 0, 1, \ldots, N-1 \tag{4.8}$$

$$Cost = g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) \tag{4.9}$$

$$min_{u_k}\, E(Cost) = E_{w_k}(g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, w_k)) \tag{4.10}$$

4.4 Inventory Control Model

The inventory control problem is the problem of how to meet a stochastic demand by ordering an optimal quantity of a certain item at the beginning of each of stages. According to Figure 4.2, if k represents the stage number, x_k denotes the available stock at the beginning of the period which presents the state of the system. u_k is the decision variable which is selected at time k with knowledge of the state x_k. w_k is a random parameter representing the stochastic demand, also called disturbance or noise. N is the horizon of the problem, number of time stages which control is applied [Di87]. The system stock evolution equation is presented in (4.11).

4.5 Problem Formulation by SDP (Inventory Control Model)

In our problem, the stock which is going to be controlled (X_k) is the energy stored in the storage system. Equation (4.12) presents the storage balance equation. x_k in the inventory control problem is here denoted by X_k. To extract u_k and w_k, we rearrange the terms of Equation (4.12) into (4.13). Therefore, by comparing (4.13) with (4.11),

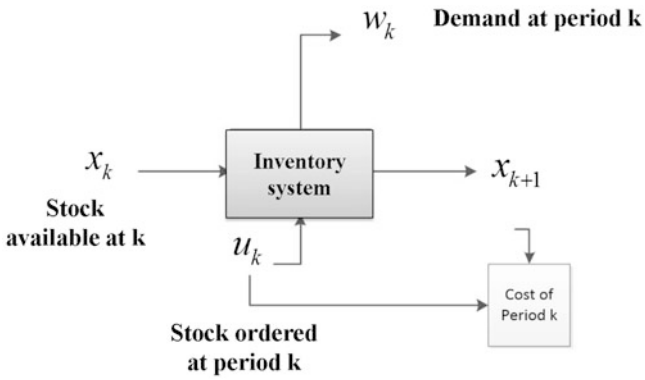


Fig. 4.2 Inventory control diagram

u_k and w_k will be defined. Equations (4.14) to (4.16) present the inventory control parameters in our problem. We can decompose the Equation (4.13) into two different equations representing two different inventory systems in order to decompose the operation paradigm into load following and generation following concepts. This is not the goal of this paper and will be discussed in our future works.

$$x_{k+1} = x_k + u_k - w_k \tag{4.11}$$

$$X_{k+1} = X_k + (Ds_k + W_k) - (CL_k + UL_k) \tag{4.12}$$

$$X_{k+1} = X_k + (Ds_k + CL_k) - (UL_k - W_k) \tag{4.13}$$

$$x_k = X_k \tag{4.14}$$

$$u_k = Ds_k + CL_k \tag{4.15}$$

$$w_k = UL_k - W_k \tag{4.16}$$

To define the total cost of the problem in the form of inventory control system, first we define the cost of every stage k as shown in (4.17). The cost from stage k to the final stage is mentioned by (4.18). The optimal cost as shown in (4.19) is the minimum value of the cost by selecting the best control parameters of Ds_k and CL_k. The terminal cost here is the cost of charging the storage on half capacity as shown in (4.20). So, our problem is to minimize the total cost on two controls u_1 and u_2 as shown in (4.21). Therefore, the problem has two control variables with different natures, first $u_1 = CL_k$ and second $u_2 = Ds_k$. Considering some simplifications, it is shown that by using the following lemma, the first control variable can be eliminated by levelizing the load profile. Firstly we present the proof of the lemma to demonstrate the load levelization and after that we levelize the load and reduce the controls of the problem into one control.

$$g(x_k, u_k, w_k) = C(Ds_k, CL_k) \tag{4.17}$$

$$J_k(x_k) = g(x_k, u_k, w_k) + J^*_{k+1}(x_{k+1}) \tag{4.18}$$

$$J^*_k(x_k) = Min_{Ds_k, CL_k}\, E(J_k(x_k)) \tag{4.19}$$

$$g_N(x_N) = (x_N - 1/2cap_{St})^2 \tag{4.20}$$

$$J^*_k(x_k) = Min_{u_1, u_2}\, E(J_k(x_k)) \tag{4.21}$$

4.6 Lemma

The goal is to minimize the total cost over N stages as in (4.22). By considering the constraints (4.23) to (4.26) of the optimization problem, Lagrangian function of the problem is produced as in (4.27).

To solve the problem, the differentiation of the Lagrangian function with respect to Ds_k which is actually the incremental cost is equaled to zero as stated in (4.28). So, the incremental cost equals with λ as stated in (4.29). Therefore Ds_k is a constant. On the other hand, Ds_k is the difference between load and expected wind power as shown in (4.30). So the load should be levelized.

$$Min_{Ds_k, L_k} Cost = \sum_{k=1}^{N} g(X_k, Ds_k, L_k, W_k) \tag{4.22}$$

$$L_k = CL_k + UL_k \tag{4.23}$$

$$UL_k \leq L_k \tag{4.24}$$

$$0 \leq Ds_k \leq \overline{Ds} \tag{4.25}$$

$$\sum_{k=1}^{N} L_k = L_0 \tag{4.26}$$

$$L.F. = \sum_{k=1}^{N} g(X_k, Ds_k, L_k, W_k) + \lambda(\sum_{k=1}^{N} L_k) \tag{4.27}$$

$$\frac{\partial L.F.}{\partial Ds_k} = \frac{\partial g}{\partial Ds_k} - \lambda = 0 \tag{4.28}$$

$$IC(Ds_k) = \lambda \tag{4.29}$$

$$Ds_k = L_k - E(W_k) \tag{4.30}$$

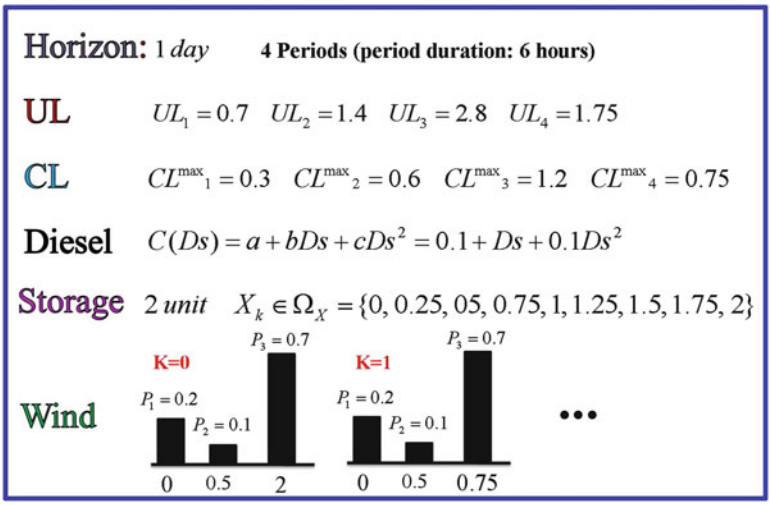


Fig. 4.3 The microgrid components data

4.7 Solution Approach: Step by Step

In this part, we will present the system data and solve the problem of microgrid operation by addition of the microgrid components step by step. The system data is aggregated in Figure 4.3. As shown in Figure 4.3, diesel generator has a quadratic cost function. The capacity of the storage system is assumed to be 2 units. For every stage, a discrete PDF for wind power with three samples is considered.

Case 1
If the microgrid consists of only diesel generation and uncontrollable loads as shown in Figure 4.4, there is no degree of freedom for the operator. In this case, the total load is uncontrollable. It means that diesel generation must exactly follow the UL variations. According to the diesel cost function presented in Figure 4.3, the total cost of the microgrid operation approximately equals with 75 units. In this case the generation is controllable and the load is uncontrollable and the only applicable operation paradigm is load following.

Case 2
If in the microgrid system of case 1, a part of the total load is considered controllable, the problem will have two control parameters. We use the aforementioned lemma to reduce the controls and simply solve the problem. The result of the lemma was to levelize the load. We have used a quadratic programming to levelize the load. As shown in Figure 4.5, the total cost has been decreased due to the degree of freedom provided by controllable loads. It should be noted that the controllable loads are assumed to be costless and the operator is not going to pay for consumers motivation in DD commitment.

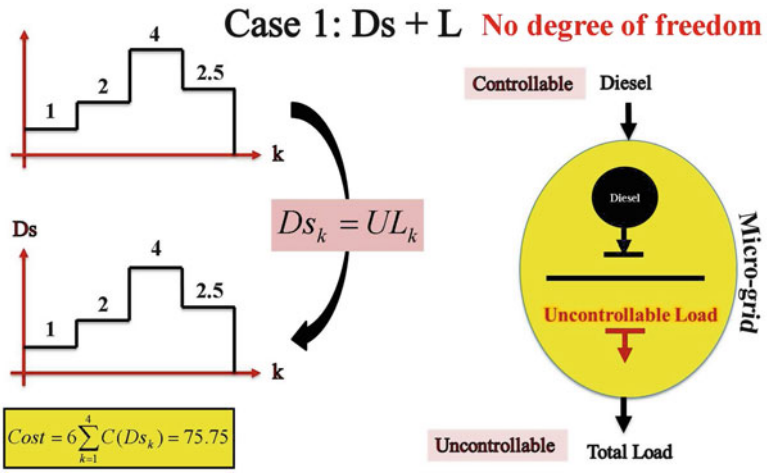


Fig. 4.4 The microgrid components in case 1

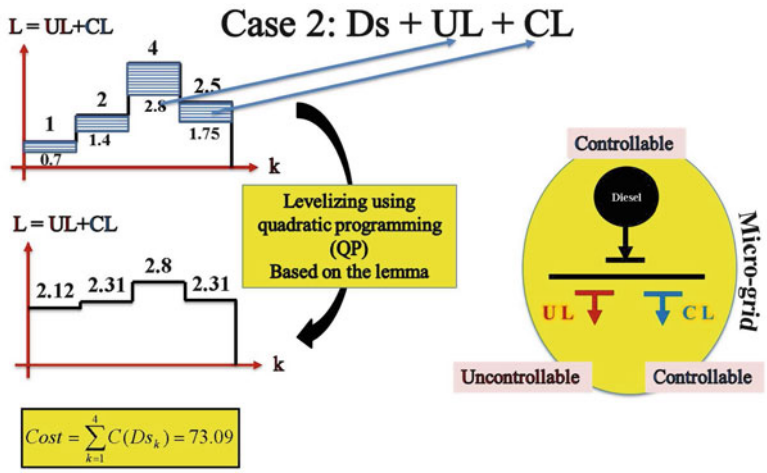


Fig. 4.5 The microgrid components in case 2 by adding controllable loads

Case 3

If the storage system is added to the microgrid system of case 2, as shown in Figure 4.6, we face with a DP problem as explained in the previous section. By solving the DP problem with $N = 4$, the optimal control chain of diesel generation will be determined as illustrated in Figure 4.7. In Figure 4.7, there are five rows of circles. Each row belongs to a certain value of the initial charge state of the storage system. The initial charge takes the values 0, 0.5, 1, 1.5, and 2. For example, if the initial charge is 1 unit, the control chain which represents the diesel generation command is (2.1, 2.3, 2.3, 2.3) which is the blue trajectory in Figure 4.7.

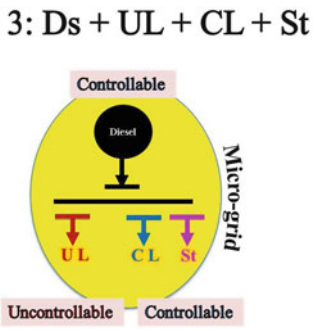


Fig. 4.6 The microgrid components in case 3 by adding storage system

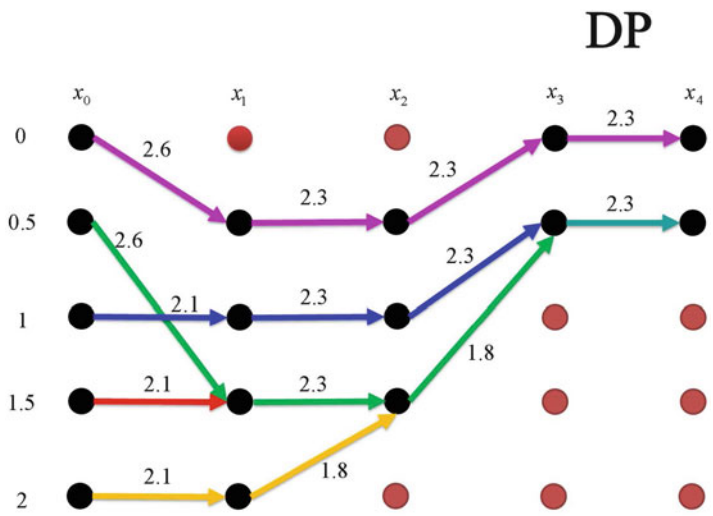


Fig. 4.7 The optimal control chain

Figure 4.8 presents the levelized load, the diesel generation, and storage charge state when the initial storage charge equals with 1 unit. The total operation cost which is shown in Figure 4.8 has been decreased by the use of storage system.

Case 4

If we add wind power generation to the microgrid system of case 3 as shown in Figure 4.9, we face with an SDP problem due to the stochastic behavior of wind power. The PDF of wind power for the stages is used to solve the SDP problem. By solving the SDP problem with $N = 4$, the optimal control of the first stage will be determined as illustrated in Figure 4.10. But since the problem is stochastic, the optimal control of the next stages will be determined when we reach that stage by solving an SDP problem with updated information.

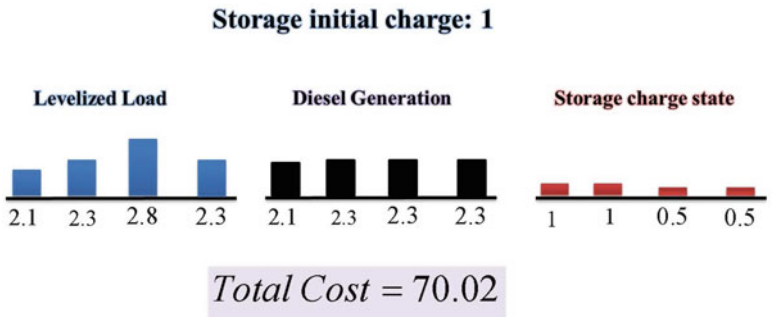


Fig. 4.8 The microgrid components in case 3 by adding storage system

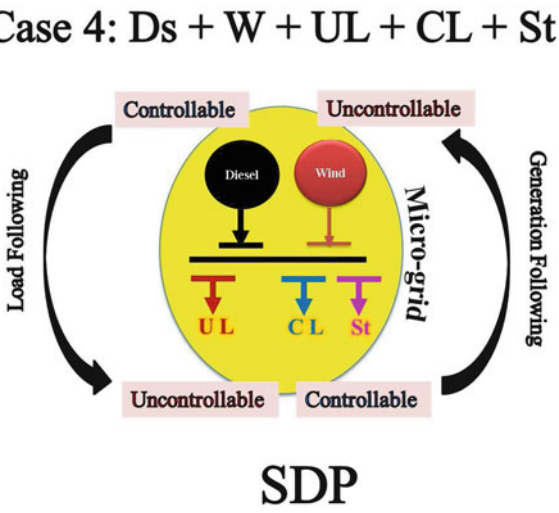


Fig. 4.9 The microgrid components in case 4 by adding wind generation

4.8 Summary and Conclusion

In this research we solved the microgrid operation by stochastic dynamic programming. Microgrid operation was implemented by adding the microgrid components step by step. Simulation results showed that the total cost of the microgrid operation reduces by adding controllable loads, storage, and wind power generation one by one. So demand dispatch and wind power integration enable grid operator to reduce the operating cost of the system.

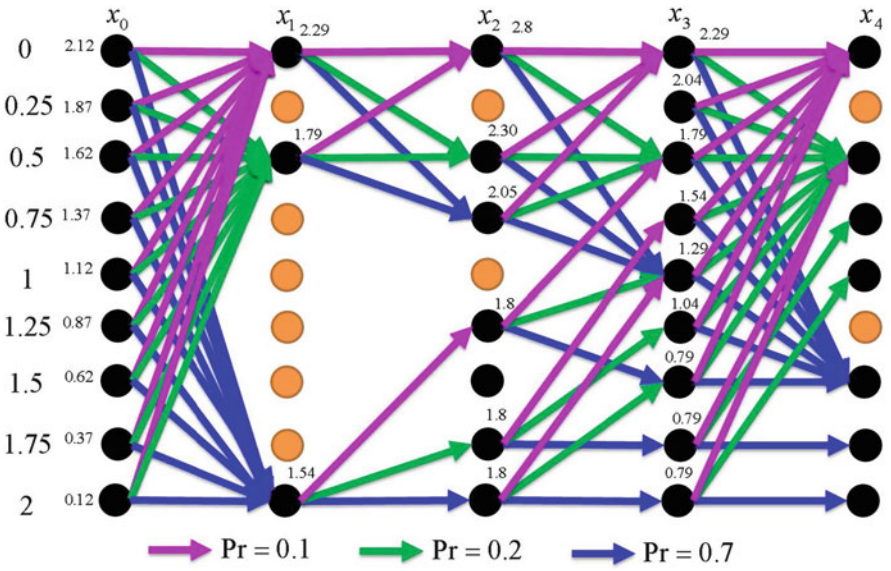


Fig. 4.10 The optimal control of the first stage

References

[AiEtAl16] Aien, A.,Hajebrahimi, A., Fotuhi-Firuzabad, M.: A comprehensive review on uncertainty modeling techniques in power system studies. J. Renew. Sust. Energ. Rev. **57**, 1077–1089 (2016)

[BeEtAl11] Berardino, J., Nwankpa, C.,Miu, K.: Economic demand dispatch of controllable building electric loads for demand response. IEEE Conference on PowerTech. Trondheim (2011)

[BoEtAl13] Botterud, A., Zhou, Z., Wang, J., Miranda, V.: Demand dispatch and probabilistic wind power forecasting in unit commitment and economic dispatch: a case study of Illinois. IEEE Trans. Sust. Energ., **4**(1), 250–261 (2013)

[BrEtAl10] Brooks, A., Lu, E., Reicher, D., Spirakis, C.,Weihl, B.: Demand Dispatch. IEEE Power Energ. Mag. **8**(3), 20–29 (2010)

[DaRa13] Daburi, F.F., Rajabi, M.H.: Effects of Demand Dispatch on operation of smart hybrid energy systems. Power System conference (PSC). Tehran (2013)

[DaRa13] Daburi, F.F., Rajabi, M.H.: Wind generation following using demand dispatch via smart grid platform. Smart Grid Conference (SGC). Tehran (2013)

[DaRa15] Daburi, F.F., Rajabi, M.H.: Modeling and implementation of demand dispatch approach on a smart micro-grid. Theoretical and Computational Advances Integral Methods in Science and Engineering, pp. 129–141. Springer International Publishing (2015)

[Di87] Dimitri P.B.: Dynamic Programming-Deterministic and Stochastic Models. Prentice-Hall, Englewood Cliffs (1987)

[NETL11] The National Energy Technology Laboratory (NETL) for the U.S. Department of Energy (DOE): Demand dispatch-intelligent demand for a more efficient grid. Office of Electricity Delivery and Energy Reliability (2011)

[WuEtAl12] Wu, T., Wu, G., Bao, Z., Yang, Q., Yan, W.: Demand dispatch of smart charging for plug-in electric vehicles. IEEE International Conference on Control Engineering and Communication Technology. Bhubaneswar (2012)

Chapter 5
Spectral Boundary Element Algorithms for Multi-Length Interfacial Dynamics

P. Dimitrakopoulos

5.1 Introduction

Interfacial dynamics in Stokes flow via the solution of boundary integral equations has developed considerably in the last decades. The main benefits of this approach are the reduction of the problem dimensionality by one and the great parallel scalability. Using this methodology, the dynamics of droplets and bubbles, elastic capsules, and erythrocytes have been investigated in basic unbounded flows and in confined microfluidic channels and vascular micro-vessels. One particularly challenging area of work is related to the study of the problem of multi-length interfacial dynamics in Stokes flow, such as the droplets coalescence, droplets and cells in close proximity to microchannel walls as well as tips and necks during large interfacial deformations. For the accurate solution of these challenging three-dimensional problems, we have developed a series of efficient and highly accurate interfacial algorithms based on our Spectral Boundary Element implementation for Stokes flow. As applications for multi-length interfacial systems, we present here our investigation of large deformation of soft particles, involving pointed tips and tails in microchannels.

5.2 Mathematical Formulation

We consider a three-dimensional soft particle such as a droplet, an artificial capsule (i.e., a fluid volume enclosed by a thin elastic membrane), or an erythrocyte, flowing inside a microfluidic channel as illustrated in Figure 5.1. To facilitate our discussion

P. Dimitrakopoulos (✉)
The University of Maryland, College Park, MD, USA
e-mail: dimitrak@umd.edu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_5

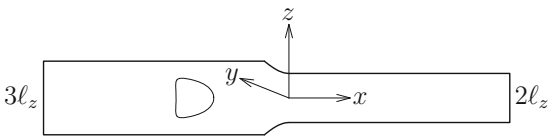


Fig. 5.1 Illustration of a capsule flowing inside a microfluidic channel

we will call as capsule all these three types of deformable multi-phase particles. The capsule's interior and exterior are Newtonian fluids, with viscosities $\lambda\mu$ and μ, and the same density. The capsule size a is specified by its volume $V = 4\pi a^3/3$ and is comparable to the micro-geometry's half-height ℓ_z. The average velocity in the channel is $\mathcal{U}$ and the time scale is $\tau_f = \ell_z/\mathcal{U}$.

Assuming low-Reynolds-number flows, the governing equations in the surrounding fluid (fluid 2) are the Stokes equations and continuity,

$$\nabla \cdot \boldsymbol{\sigma} \equiv -\nabla p + \mu \nabla^2 \boldsymbol{u} = 0 \qquad \text{and} \qquad \nabla \cdot \boldsymbol{u} = 0$$

where $\boldsymbol{\sigma}$ is the stress tensor and $\boldsymbol{u}$ the fluid velocity. Inside the capsule (fluid 1), the same equations apply with the viscosity replaced by $\lambda\mu$. It is of interest to note that in small length-scale systems, such as microfluidic channels, low-Reynolds-number flows are easily achievable. (For example, in a microfluidic channel with size $\ell_z = 100\,\mu m$, the Reynolds number remains $Re = O(10^{-3})$ even for velocities up to $\mathcal{U} = 10\,mm/s$ when we consider the density and viscosity of water.)

For the problem illustrated in Figure 5.1, the system surface S_B consists of the capsule interface S_c, the micro-device's solid surface S_s, and the fluid surface S_f of the inlets and outlets of the micro-device. At the capsule's interface, the velocity is continuous and we define the surface stress vector (or hydrostatic traction) $\Delta \boldsymbol{f}$ from the stress tensor $\boldsymbol{\sigma}$ and the surface unit normal $\boldsymbol{n}$, i.e.,

$$\boldsymbol{u}_1 = \boldsymbol{u}_2 = \boldsymbol{u} \qquad \text{and} \qquad \Delta \boldsymbol{f} \equiv \boldsymbol{n} \cdot (\boldsymbol{\sigma}_2 - \boldsymbol{\sigma}_1)$$

Here the subscripts designate quantities evaluated in fluids 1 and 2, respectively, while $\boldsymbol{n}$ is the unit normal which we choose to point into fluid 2. The boundary conditions on the rest surfaces are

$$\boldsymbol{u} = 0 \quad \text{on the solid boundary } S_s$$

$$\boldsymbol{u} = \boldsymbol{u}^{\infty} \quad \text{on the fluid boundary } S_f$$

Based on standard boundary integral formulation, the velocity at a point $\boldsymbol{x}_0$ on the system surface S_B may be expressed as a surface integral of the force vector $\boldsymbol{f} = \boldsymbol{n} \cdot \boldsymbol{\sigma}$ and the velocity $\boldsymbol{u}$ over all points $\boldsymbol{x}$ on the boundary S_B,

$$\Omega\, \boldsymbol{u}(\boldsymbol{x}_0) = -\int_{S_c} \left[\boldsymbol{S} \cdot \Delta \boldsymbol{f} - \mu(1-\lambda)\, \boldsymbol{T} \cdot \boldsymbol{u} \cdot \boldsymbol{n}\right](\boldsymbol{x})\, \mathrm{d}S$$

$$-\int_{S_s \cup S_f} (\boldsymbol{S} \cdot \boldsymbol{f} - \mu\, \boldsymbol{T} \cdot \boldsymbol{u} \cdot \boldsymbol{n})\,(\boldsymbol{x})\, \mathrm{d}S \tag{5.1}$$

where the coefficient Ω takes values $4\pi\mu(1+\lambda)$ and $4\pi\mu$ for points $\boldsymbol{x}_0$ on the surfaces S_c and $S_s \cup S_f$, respectively. The tensors $\boldsymbol{S}$ and $\boldsymbol{T}$ are the fundamental solutions for the velocity and stress for the three-dimensional Stokes equations defined by

$$S_{ij} = \frac{\delta_{ij}}{r} + \frac{\hat{x}_i \hat{x}_j}{r^3} \qquad\qquad T_{ijk} = -6\frac{\hat{x}_i \hat{x}_j \hat{x}_k}{r^5} \tag{5.2}$$

where $\hat{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{x}_0$ and $r = |\hat{\boldsymbol{x}}|$ [WaEtAl06c, Di07a].

Owing to the no-slip condition at the interface, the time evolution of the capsule surface may be determined via the kinematic condition at the interface

$$\frac{\partial \boldsymbol{x}}{\partial t} = \boldsymbol{u}$$

To produce a closed system of equations, the surface stress $\Delta\boldsymbol{f}$ on the capsule interface is determined from the interfacial properties of the specific soft particle. Thus for droplets, $\Delta\boldsymbol{f}$ is associated with the surface tension γ

$$\Delta\boldsymbol{f} = \boldsymbol{f}_2 - \boldsymbol{f}_1 = \gamma(\nabla \cdot \boldsymbol{n})\boldsymbol{n}$$

while for artificial or biological membranes the surface stress $\Delta\boldsymbol{f}$ is related to the membrane tensions which are affected by the shear and area-dilatation moduli, G_s and G_a, of the membrane [BaEtAl02, WaEtAl06c, DoEtAl09]. In particular, the surface stress is determined by the in-plane stresses which in contravariant form gives

$$\Delta\boldsymbol{f} = -\nabla_s \cdot \boldsymbol{\tau} = -(\tau^{\alpha\beta}|_\alpha\, \boldsymbol{t}_\beta + b_{\alpha\beta}\, \tau^{\alpha\beta}\, \boldsymbol{n})$$

where the Greek indices range over 1 and 2, while Einstein notation is employed for (every two) repeated indices. In this equation, the $\tau^{\alpha\beta}|_\alpha$ notation denotes covariant differentiation, $\boldsymbol{t}_\beta = \partial \boldsymbol{x}/\partial\theta^\beta$ are the tangent vectors on the capsule surface described with arbitrary curvilinear coordinates θ^β, and $b_{\alpha\beta}$ is the surface curvature tensor [Po03, DoEtAl09]. The in-plane stress tensor $\boldsymbol{\tau}$ is described by constitutive laws that depend on the material composition of the membrane. For example, the strain-hardening Skalak *et al.* law [SkEtAl73] relates $\boldsymbol{\tau}$'s eigenvalues (or principal elastic tensions $\tau^P_\beta, \beta = 1, 2$) with the principal stretch ratios λ_β by

$$\tau^P_1 = \frac{G_s \lambda_1}{\lambda_2}\{\lambda_1^2 - 1 + C\lambda_2^2[(\lambda_1\lambda_2)^2 - 1]\}$$

while to calculate τ_2^P reverse the λ_β subscripts. Note that the reference shape of the elastic tensions is the quiescent capsule shape while the membrane hardness C represents the dimensionless area-dilatation modulus, $G_a/G_s = 1 + 2C$ [SkEtAl73, Po03].

The interfacial problem depends on several physical dimensionless parameters, including the capillary number $\mathrm{C}a = \mu\mathcal{U}/\gamma$ for droplets or $\mathrm{C}a = \mu\mathcal{U}/G_s$ for membranes, the viscosity ratio λ and the ratio of the membrane moduli G_a/G_s (or C), as well as on geometric parameters such as the ratio of the capsule size to the channel size a/ℓ_z.

5.3 Interfacial Spectral Boundary Element Algorithms

To solve the interfacial problem via the boundary integral formulation we have developed a series of highly accurate interfacial algorithms based on our Spectral Boundary Element method. In particular, first we developed an interfacial spectral boundary element method for droplets and bubbles in unbounded and confined microchannel flows [WaEtAl06c] which was later expanded to membrane interfaces [DoEtAl08, DoEtAl09]. Both algorithms utilize explicit time integration to advance the interface in time, and thus they require small time steps for stability, $\Delta t < O(\mathrm{C}a\,\Delta x)$. For stiff interfacial problems, there is a need to make the employed time step Δt independent of grid density or small physical length scales Δx and the capillary number $\mathrm{C}a$. To achieve this goal, we developed an efficient fully implicit Interfacial Spectral Boundary Element algorithm for droplets and bubbles [Di07a] by combining different implicit schemes with our Jacobian-free Newton iteration. To facilitate the computational study of erythrocytes whose complicated biological membrane represents a stiff description owing to the highly inextensible lipid bilayer [SkEtAl73, SkEtAl89], we also developed a non-stiff cytoskeleton-based continuum erythrocyte modeling [DoEtAl10]. Our computational results for the deformation and tank-treading motion of erythrocytes in shear flows are in exceptional agreement with experimental findings from ektacytometry and rheoscopy systems and reveal the correct shear modulus of the erythrocyte membrane [DoEtAl10, DoEtAl11, Di02, HeEtAl99].

For any soft particle of interest (i.e., droplet, artificial capsule, or erythrocyte), the numerical solution of the interfacial problem is achieved through our spectral boundary element method [WaEtAl06c, Di07a, DoEtAl09]. Briefly, each boundary is divided into a moderate number N_E of curvilinear quadrilateral elements (as seen in Figure 5.2) which are parameterized by two variables ξ and η on the square interval $[-1, 1]^2$ [WaEtAl06c, Di07a]. The geometry and physical variables are discretized using Lagrangian interpolation in terms of these parametric variables. The N_B basis points (ξ_i, η_i) for the interpolation are chosen as the zeros of orthogonal polynomials of Gauss-type. This is equivalent to an orthogonal polynomial expansion and yields the spectral convergence associated with such expansions.

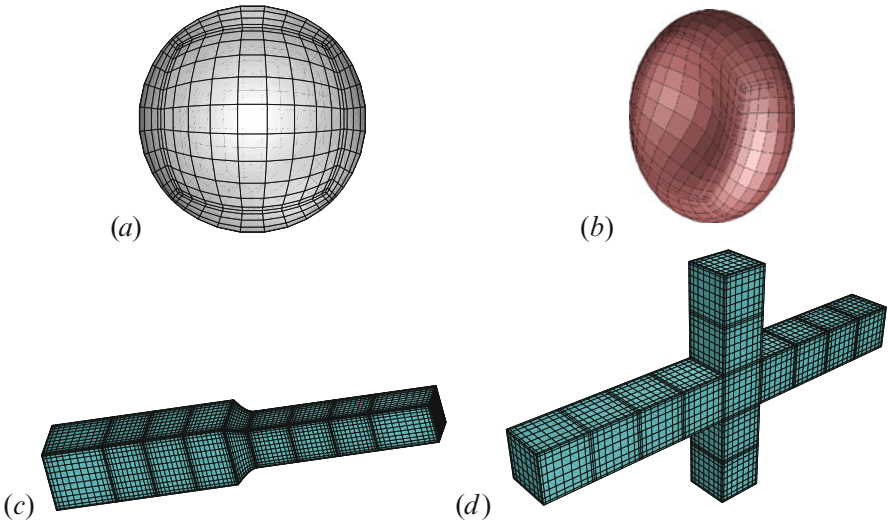


Fig. 5.2 Spectral boundary element discretization of (a) a spherical droplet or capsule, (b) an erythrocyte, and (c, d) microfluidic geometries

The boundary integral equation (5.1) admits two different types of points. The collocation points $\boldsymbol{x}_0$ where the equation is required to hold and the basis points $\boldsymbol{x}$ where the physical variables $\boldsymbol{u}$ and $\boldsymbol{f}$ are specified or determined. Our spectral boundary element method employs collocation points $\boldsymbol{x}_0$ of Legendre–Gauss quadrature, i.e., in the interior of the elements. As a result the boundary integral equation holds even for singular elements, i.e., the elements which contain the corners of the channel geometry. (Similar approach has been utilized in our earlier papers for droplets attached to solid surfaces, and vascular endothelial cells or leukocytes adhering to the surface of blood vessels, e.g., [WaEtAl06a, WaEtAl06b, Di07b].) In addition, we use basis points $\boldsymbol{x}$ of Legendre–Gauss–Lobatto quadrature and thus the physical variables are determined in the interior and on the edges of the spectral elements. For the time integration, we employ a Runge-Kutta scheme with a typical time step $\Delta t/\tau_{\mathrm{f}} \leq 10^{-3}$ for our explicit methods and a much higher Δt for our droplet fully implicit algorithm. Further details on our spectral boundary element algorithms are given in our earlier publications [WaEtAl06c, Di07a, DoEtAl09, KuEtAl11].

The main benefits of our interfacial spectral algorithms are the exponential convergence in determining the transient and steady-state interfacial shape and the ability to handle complicated solid geometries owing to the boundary element nature. As seen in Figure 5.3, by employing $N = N_E N_B^2 = 2000$ spectral points on the capsule interface, we determine the interfacial curvature with an error of 10^{-8} and the interfacial deformation with a much smaller error. Even for $N = 1000$ spectral points, the interfacial accuracy is still several significant digits.

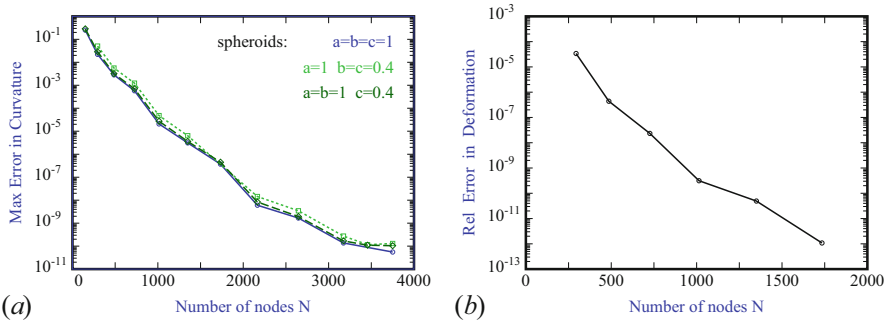


Fig. 5.3 Exponential convergence owing to the spectral accuracy in determining (*a*) the interfacial curvature of different spheroids (based on the exact solution) and (*b*) the transient deformation of droplets or capsules (based on the numerical solution of the denser grid employed)

Table 5.1 Efficiency versus the number of processors (or cores) for the calculation ("Integration") and the solution ("Solution") of the system matrix, on Linux Clusters. Owing to the fast solution (with respect to the CPU time needed for Integration), the overall parallelization for one step (i.e., combined Integration and Solution) is practically identical to that for Integration

Linux Cluster		
*N*cores	Integration	Solution
1	100.0%	100.0 %
5	99.2%	95.8 %
10	98.1%	92.8 %
15	96.9%	88.4 %
20	94.9%	85.4 %

Our spectral boundary element algorithms have the ability to exploit possible symmetry planes in the interfacial problems we study. Exploiting m symmetry levels (where usually $m = 1, 2, 3$ for a given problem) reduces the memory requirements by a factor of 4^m, the computational time for determining the system matrices by a factor of 2^m, and the solution time via direct system solvers by a factor of 8^m. Thus, the overall computational cost of our algorithm is dictated by the determination of the system matrices, and thus our algorithm achieves an overall 95% parallel efficiency on 20 cores as shown in Table 5.1.

5.4 Multi-Length Interfacial Dynamics Problems

In this section, we present several multi-length interfacial problems we have investigated in our recent publications. Figure 5.4 shows the supercritical evolution of a droplet in a planar extensional flow where the interfacial shape elongates significantly and a neck is created in the droplet middle. To account for this large

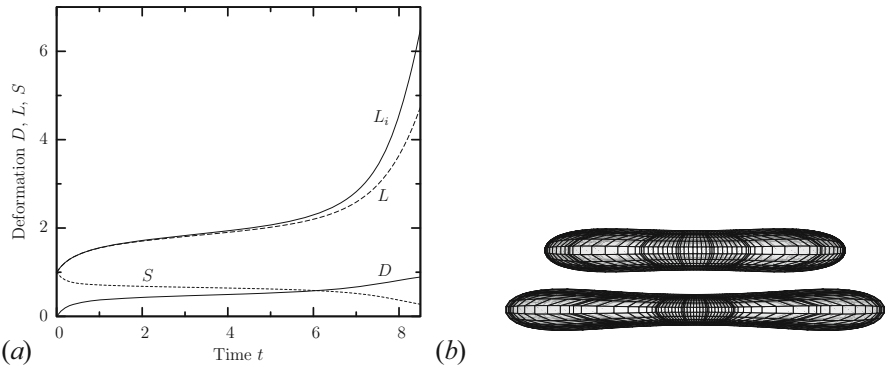


Fig. 5.4 Supercritical evolution of a droplet with $\lambda = 0.209$ in a planar extensional flow for $Ca = 0.163$ [Di07a]

deformation, we start with a droplet discretization employing 6 spectral elements and adaptively increase the number of the spectral elements in the droplet middle as seen in Figure 5.4(*b*). For this case, we utilize our fully implicit algorithm with a large time step of $\Delta t/\tau_f = 0.1$ while our results are in excellent agreement with theoretical and experimental findings [Di07a].

A sequence of steady-state shapes of a Skalak capsule in a planar extensional flow is shown in Figure 5.5. Beyond the large interfacial deformation, the edge curvature increases significantly while above a critical capillary number, cusped edges appear with negative curvature owing to a transition of the edge tensions from positive to negative (or compressive) [DoEtAl08]. Thus, our high-order spectrally accurate computational methodology predicts stable equilibrium shapes whose edges become rounded, spindled, and finally cusped with increasing flow rate, in agreement with experimental findings [Ba91]. This multi-length interfacial problem shows that the local edge length scale $\sim \frac{1}{\text{curvature}} \ll 1$, i.e., 2 to 3 orders smaller than the capsule size.

Typical interfacial shapes of strain-hardening Skalak capsules in square and rectangular microchannels are shown in Figure 5.6. Our membrane spectral boundary element algorithm determines accurately the interfacial shape (to at least 3 significant digits) utilizing rather coarse grids. The multi-length interfacial problem results from the narrow lubrication gaps between the capsule membrane and the solid walls and the creation of dimples with negative curvature at the capsule's rear. For large capsule sizes, we also developed a scaling analysis for the steady-state capsule properties utilizing a Landau-Levich-Derjaguin-Bretherton lubrication analysis extended to membranes [KuEtAl11]. In a rectangular channel, the capsule extends mainly along the less-confined lateral direction of the channel cross section (i.e., the channel width), obtaining a pebble-like shape owning to tension development on the capsule membrane required for interfacial stability [KuEtAl13].

A typical capsule deformation in a microfluidic constriction is shown in Figure 5.7. Our work highlights the effects of two different mechanisms for non-tank-

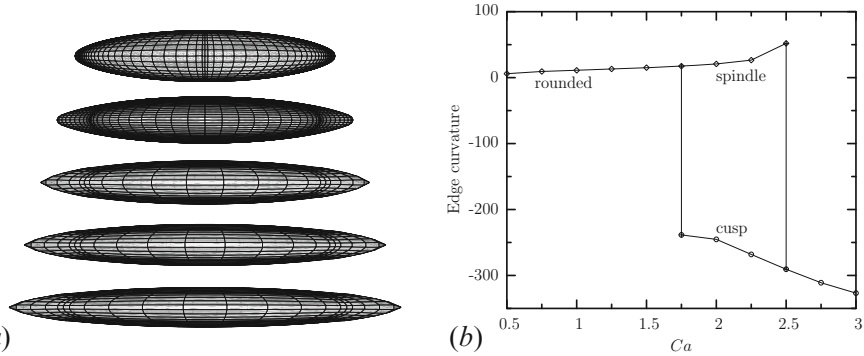


Fig. 5.5 (*a*) Steady-state capsule shapes with increasing flow rates for a Skalak capsule with $C = 1$, $\lambda = 1$ and $\mathrm{C}a = 1, 1.5, 2, 2.5, 3$, starting from a quiescent spherical shape. (*b*) Bifurcation in the edge curvature with creation of spindled and cusped edges [DoEtAl08]

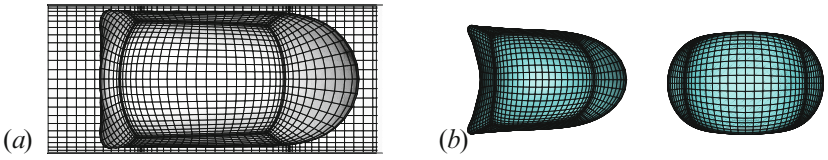


Fig. 5.6 (*a*) Steady-state capsule shape in a square microfluidic channel for $\mathrm{C}a = 0.1$, $\lambda = 1$, $C = 1$, and $a/\ell_z = 1.3$ [KuEtAl11]. (*b*) Channel and side view of the steady-state capsule shape in a rectangular microchannel with an aspect ratio of 2, for $\mathrm{C}a = 0.2$, $\lambda = 1$, $C = 1$, and $a/\ell_z = 1.1$ [KuEtAl13]

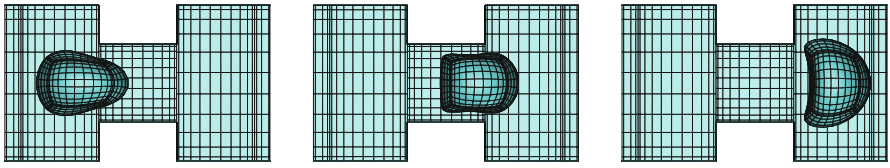


Fig. 5.7 The shape of a Skalak capsule with $C = 1$, $a/\ell_z = 1$, $\lambda = 1$, and $\mathrm{C}a = 0.1$ moving inside a microfluidic constriction [PaEtAl13]

treading transient capsule dynamics, i.e., the effects of normal stresses and shear stresses on the capsule membrane [PaEtAl13]. The capsule deformation results from the combined effects of the surrounding and inner fluids' normal stresses on the soft particle's interface, and thus when the capsule viscosity increases, its transient deformation decreases, as for droplets. However, the capsule deformation is not able to create a strong enough inner circulation (owing to restrictions imposed by the material membrane), and thus the viscosity ratio does not affect much the capsule velocity and the additional pressure difference.

As a final example of multi-length interfacial dynamics we present the transient evolution of a low-viscosity droplet (i.e., a bubble) inside a microfluidic

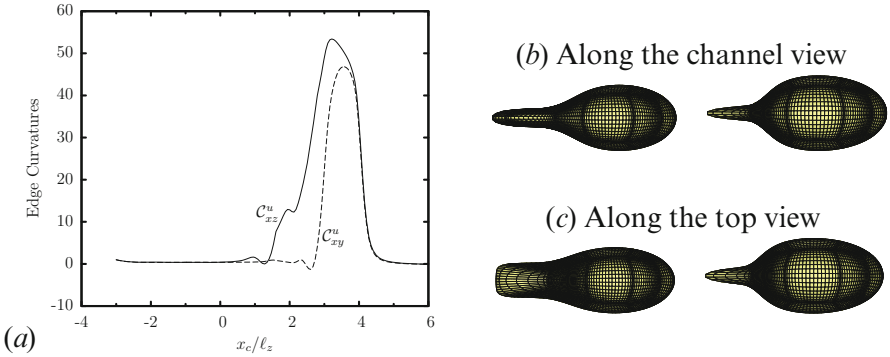


Fig. 5.8 (*a*) Interfacial tail curvature along the channel view C^d_{xz} and the top view C^d_{xy} as a function of the bubble centroid x_c. The bubble shape, along (*b*) the channel view and (*c*) the top view, at two positions after passing the cross-junction. Parameters: $Ca = 0.3$, $\lambda = 0.01$, $a/\ell_z = 0.7$, and relative lateral flow rate $Q_v = 0.75$ [BoEtAl15]

cross-junction made by two perpendicular square channels (see Figure 5.2*d*). To investigate this problem we utilized our fully implicit interfacial spectral boundary algorithm with a relatively large time step $\Delta t/\tau_f = 0.01$ independent of the space grid Δx and the interfacial deformation even during the creation of the long pointed tails [BoEtAl15].

As seen in Figure 5.8, along the channel view, bubbles develop very pointed tails for interfacial stability. At sufficiently large lateral flow rates, the multi-length nature of the interfacial problem is clearly revealed at the creating of sharp tail curvatures where the local length scale is $O(50)$ smaller than the bubble length. For such pointed tails, our implicit algorithm divides the spectral element at the droplet's tail into five smaller elements as the pointed tail is formed, to produce sufficient spatial discretization needed for the accurate determination of the interfacial shape. Element division is the adaptive mesh reconstruction technique of our spectral element algorithms so that they are able to produce a reasonable spectral element discretization, needed especially for local interfacial deformations such as tails and necks, as described in our earlier papers [WaEtAl06c, Di07a].

References

[Ba91] Barthès-Biesel, D.: Role of interfacial properties on the motion and deformation of capsules in shear flow. Phys. A **172**, 103–124 (1991)

[BaEtAl02] Barthès-Biesel, D., Diaz, A., Dhenin, E.: Effect of constitutive laws for two-dimensional membranes on flow-induced capsule deformation. J. Fluid Mech. **460**, 211–222 (2002)

[BoEtAl15] Boruah, N., Dimitrakopoulos, P.: Motion and deformation of a droplet in a microfluidic cross-junction. J. Colloid Interface Sci. **453** 216–225 (2015)

[Di02] Dimitrakopoulos, P.: Analysis of the variation in the determination of the shear modulus of the erythrocyte membrane: Effects of the constitutive law and membrane modeling. Phys. Rev. E. **85**, 041917 (2002)

[Di07a] Dimitrakopoulos, P.: Interfacial dynamics in Stokes flow via a three-dimensional fully-implicit interfacial spectral boundary element algorithm. J. Comput. Phys. **225**, 408–426 (2007)

[Di07b] Dimitrakopoulos, P.: Deformation of a droplet adhering to a solid surface in shear flow: onset of interfacial sliding. J. Fluid Mech. **580**, 451–466 (2007)

[Di14] Dimitrakopoulos, P.: Effects of membrane hardness and scaling analysis for capsules in planar extensional flows. J. Fluid Mech. **745**, 487–508 (2014)

[DoEtAl08] Dodson, W.R. III, Dimitrakopoulos, P.: Spindles, cusps and bifurcation for capsules in Stokes flow. Phys. Rev. Lett. **101**, 208102 (2008)

[DoEtAl09] Dodson, W.R. III, Dimitrakopoulos, P.: Dynamics of strain-hardening and strain-softening capsules in strong planar extensional flows via an interfacial spectral boundary element algorithm for elastic membranes. J. Fluid Mech. **641**, 263–296 (2009)

[DoEtAl10] Dodson, W.R. III, Dimitrakopoulos, P.: Tank-treading of erythrocytes in strong shear flows via a non-stiff cytoskeleton-based continuum computational modeling. Biophys. J. **99**, 2906–2916 (2010)

[DoEtAl11] Dodson, W.R. III, Dimitrakopoulos, P.: Oscillatory tank-treading motion of erythrocytes in shear flows. Phys. Rev. E., **84** 011913 (2011)

[HeEtAl99] Hénon, S., Lenormand, G., Richert, A., Gallet, F., A new determination of the shear modulus of the human erythrocyte membrane using optical tweezers. Biophys. J. **76**, 1145–1151 (1999)

[KuEtAl11] Kuriakose, S., Dimitrakopoulos, P.: Motion of an elastic capsule in a square microfluidic channel. Phys. Rev. E. **84**, 011906 (2011)

[KuEtAl13] Kuriakose, S., Dimitrakopoulos, P.: Deformation of an elastic capsule in a rectangular microfluidic channel. Soft Matter **9**, 4284–4296 (2013)

[PaEtAl13] Park, S.-Y., Dimitrakopoulos, P.: Transient dynamics of an elastic capsule in a microfluidic constriction. Soft Matter **9**, 8844–8855 (2013)

[Po03] Pozrikidis, C. (ed.): Modeling and Simulation of Capsules and Biological Cells. Chapman and Hall, London (2003)

[SkEtAl73] Skalak, R., Tozeren, A, Zarda, R.P., Chien, S.: Strain energy function of red blood cell membranes. Biophys. J. **13**, 245–264 (1973)

[SkEtAl89] Skalak, R., Özkaya, N., Skalak, T.C.: Biofluid mechanics. Ann. Rev. Fluid Mech. **21**, 167–204 (1989)

[WaEtAl06a] Wang, Y., Dimitrakopoulos, P.: Normal force exerted on vascular endothelial cells. Phys. Rev. Lett. **96**(1–4), 028106 (2006)

[WaEtAl06b] Wang, Y., Dimitrakopoulos, P.: Nature of the hemodynamic forces exerted on vascular endothelial cells or leukocytes adhering to the surface of blood vessels. Phys. Fluids **18**(1–14), 087107 (2006)

[WaEtAl06c] Wang, Y., Dimitrakopoulos, P.: A three-dimensional spectral boundary element algorithm for interfacial dynamics in Stokes flow. Phys. Fluids **18**(1–16), 082106 (2006)

Chapter 6
Kinect Depth Recovery Based on Local Filters and Plane Primitives

M.A. Esfahani and H. Pourreza

6.1 Introduction

These days RGB-D cameras, especially Kinect (introduced by Microsoft in 2010), is providing depth map besides the color image of the capturing point of view by triangulating specific infrared patterns [FrEtAl13]. This new feature is beneficial for wide number of problems in the area of Computer Vision, especially for mobile robots to understand the scene and improve their knowledge about its geometry. To create an accurate road map from the input RGB-D data collected by the mobile robots, a significant constrain is to have an accurate depth map which helps to have a better understanding of the desired scene. Having an accurate depth map as input is also an important point in wide number of other problems [ZhEtAl17, ChEtAl16].

Captured depth map using Kinect sensor suffers from both holes and invalid measurements called noise. Holes are the pixels that depth sensor was unable to compute any depth value for them; because of the lighting conditions or being a glass or mirror in front of the IR camera. Invalid measurements which are mostly called as noise in the literature are also involved in the captured depth map due to the lightning condition, the way that the IR pattern is reflecting to the camera, the properties of the object surface that IR pattern is facing with, and finally lacking in calibration and measurement of disparities. It is also important to notify that the value of noise increases according to the distance exponentially (Figure 6.1).

Overall, the problem of depth recovery breaks down into two parts of depth hole filling and fixing invalid measurements or briefly called denoising. To visualize the problem and get familiar with this issue, Figure 6.2 exemplifies holes in a depth map which captured by a Kinect sensor. In the presented depth map, brown points are holes and no value is measured for them. There exist also invalid measurements

M.A. Esfahani • H. Pourreza (✉)
Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: Mahdi.Abolfazli@stu-mail.um.ac.ir; hpourreza@um.ac.ir

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_6

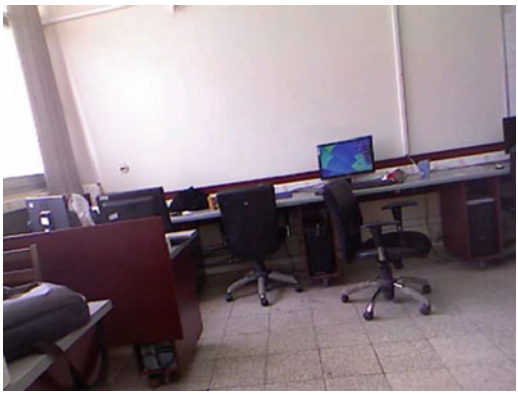


Fig. 6.1 RGB image captured with Kinect sensor

Fig. 6.2 Depth image captured with Kinect sensor

in the illustrated depth map. Figure 6.1 is the correspondence RGB image for the captured depth map and it is obvious that its resolution is more than the output of the depth sensor. The RGB camera of the Kinect sensor has the resolution of 640x480 while its depth resolution is 320x240.

Mentioned properties of the Kinect sensor and also its characteristic make it critical to remove both the invalid measurements and also filling the holes without estimated depth value. To that, recent works [YaEtAl07, DoEtAl10] applied Bilateral Filters (BF) [ToMa98] on the depth map to reduce the noise of inaccurate measurement. Sudden changes on the border of objects help BF to determine a subjective area and be able to estimate a reliable depth value for that. While it focuses on the border of objects, it is not recommended for denoising depth maps that contain high number of holes, since holes also describe a type of border to it.

To overcome with this issue, using the correspondence color image is recommended. Most of the recent works [ChEtAl12, RiEtAl12, CaSa12] used Joint Bilateral Filter (JBF) [KoEtAl07] to add properties of color image into their computations as a guidance. Despite fixing the problem of existence holes and its high performance, it works worst in the areas where the foreground and background have same color attributes. Chen et al. [ChEtAl15, ChEtAl13] formulated the

problem as an energy minimization function that merges behavior of BF, JBF, and also Joint Trilateral Filter (JTF) [LiEtAl10]. While their method performs well, they have not included the rich information of the scene structure in their minimization function which is the focus of this chapter and helps to have a better understanding about the effect of pixels on each other. Each of the mentioned filters describes and analyzes in the next steps in detail.

To focus on such filters and see how they work envisage that image I and its correspondence depth map Z exists. Hence, the recovered depth value for each pixel in the depth map describes as

$$Z' = \sum_{j \in \Omega_i} \alpha_{ij} Z(j).$$

where Ω_i is set of points with valid depth which are neighbor of pixel i and α_{ij} is the normalized weight that shows the effect of pixel j on pixel i and defines as

$$\alpha_{ij} = \frac{\beta_{ij}}{\sum_{j \in \Omega_i} \beta_{ij}}$$

where the weight β_{ij} is

$$\beta_{ij} = \begin{cases} G_S(i,j)G_Z(i,j), for\ BF \\ G_S(i,j)G_I(i,j), for\ JBF \\ G_S(i,j)G_I(i,j)G_Z(i,j), for\ JTF \end{cases}$$

and defines according to the type of the filter that system is using. In this equation, G_S, G_I, and G_Z are probability density functions and mostly define as Gaussian probability density function in spatial, color, and depth domain, respectively. Each of these shows the pairwise effect of each pair of pixels in each of the spaces. Since all of the introduced filters are local, their value for the center pixel, called β_{ij}, is equal to 1. According to relation of distance and noise subject to the Kinects' characteristic, this value is too large for the center pixel and effects worst. Adaptive methods also introduced to handle this issue, but they have not achieved grateful results [ChEtAl15, ChEtAl13].

Using each of the BF, JBF, and JTF filters benefits us to understand the scene in a specific manner. To use the pros of all of the introduced filters and reduce effect of their cons and limitations, a minimization framework that merges all these introduces. In this framework, effective features of different filters come together and combine with the structure of planes that model the scene. Using structure of the scene helps to have an initial guess for the holes and also reduces the measurement noise, since points in the 3D coordinate are standing near each other in a meaningful way and planes describe that well. Rest of the chapter is going to discuss about the way planes of the scene extracts and models the scene, the energy minimization function based on the structure of the scene, and comparing its results with the result of basic filters.

6.2 Proposed Method

In order to benefit from the structure of the scene in depth recovery and formulate that, this part goes towards modeling the structure of the scene using primitives. Efficient Ransac [ScEtAl07] is a method that helps in this process. In this chapter, an efficient modified version of that uses: Parallel RANSAC [CoEtAl15], which extracts planes of the desired scene using normal vector map of the input point cloud. Since it uses normal vectors to extract independent planes, it is possible to run this method parallel and benefit from the high speed of Graphical Processing Unit (GPU), and get a higher probability of best plane extraction by increasing number of iterations.

RANdom SAmple Consensus, briefly called RANSAC, classifies input data into two classes of inliers and outliers iteratively. In each iteration, it selects subset of data points and fits a model, e.g., line or plane, on them. The final result of the iterative RANSAC is the model that fits high number of inliers. Since RANSAC works iteratively on a subset of data points, its probability of fitting an accurate model improves by increasing its number of iterations. For instance, to fit a plane model, three points in the space are required. Hence, if the probability of extracting primary plane and selecting sampling point on that plane be ρ and u, respectively, the minimum number of iterations that require to fit the model calculates using Equation (6.1). For this reason, having more number of iterations the probability of fitting exact model improves.

$$N = \frac{log(1-\rho)}{log(1-u^3)}. \tag{6.1}$$

Alehdaghi et al. [FiBo81] introduced Parallel RANSAC based on GPU to extract planes, and showed that its computation is linear in order. To make RANSAC parallel, it is essential to determine independent parts that fitted model would not have any overlap with the other parts or segments. To make independent parts and extract plane model from them, it is conceivable to extract normal vectors, segment the image based on them, break down the global problem into small parts, and run RANSAC locally on each of the segmented boundaries. Segmenting according to the normal vector extracts the parts with high potential of having the same plane, since points of a plane are going to have same normal vector.

After estimating correspondent plane to each point of the desired scene, it is suitable to model the scene and determine an initial guess for all the points. Using this initial guess, the structure of the scene used as a part of the depth hole filling process. In the next step, the difference of the normal vector of each point with its neighbors included as a part of the minimization function to reduce the existence measurement noise besides the guidance of local filters. Next steps go toward formulating it using Kinects' characteristics.

There are some characteristics that neighbor pixels have in any input depth map. For instance, there exists less depth difference in the smooth areas or large

error exists in the border of objects. Combining all these characteristics together a minimization energy function which consists of a fidelity and data term could be signified. This minimization consists of two terms to combine the two characteristics mentioned above. Hence, the minimization energy function defines as

$$\min_{Z'(i)} E_r(Z'((i)) + \lambda E_d(Z'(i))$$

where E_r and E_d are the regularization term and data term, respectively, and Z' presents the recovered depth map. λ is a trade-off factor between data and regularization term. This minimization function was firstly introduced by Chen et al. [ChEtAl15, ChEtAl13]. In the next steps we are going to define the properties of the both regularization and fidelity terms and include the effect of the scene structure in computations.

Data term includes the fact that accuracy of measured depth decreases as the distance between the object and Kinect sensor increases, and also the fact that states the depth on the smooth areas of objects is reliable and is unreliable on their boundary. According to these, the data term defines as

$$E_d(Z'(i)) = \frac{1}{2} \sum_{i \in \Omega_d} w_i (Z'(i) - Z(i))^2$$

where Ω_d is the subset of points with a valid measured depth values. This equation goes toward minimizing the weighted squared difference between the recovered depth value and the original one according to their information quality. In this part, the weight w plays grate rule and defines as

$$w_i = \frac{Z_{max}^2 - Z_{avg_i}^2}{Z_{max}^2 - Z_{min}^2}$$

to have more focus on the reliable depth values which are nearer to the Kinect sensor. In this equation, Z_{max} and Z_{min} are the max and min distance that Kinect sensor can measure and Z_{avg_i} is average depth of boundary around pixel i with reliable and valid depth values. Beside the mentioned characteristics of Kinect sensor, it is important to include the point that difference between a point and its neighbor in a smooth region is too small. Hence, the regularization term defines as

$$E_r(U(i)) = \frac{1}{2} \sum_{i \in \Omega_s} \sum_{j \in \Omega_i} w_{ij} (U(i) - U(j))^2$$

where Ω_s is the subset of points with valid neighborhood and Ω_i is each of the neighbors of pixel i in that subset. w_{ij} plays an important rule to classify similar and dissimilar pixels subject to their region; it checks that by locating boundaries with sudden changes using different types of information.

While neighboring pixels have to be similar with low difference, they have to be dissimilar in the sudden changes of depth and colors where an edge exists. The coefficient w_{ij} controls this behavior by considering color and depth images. It defines as the normalized coefficient

$$w_{ij} = \frac{\beta_{ij}}{\sum_{j \in \Omega_i} \beta_{ij}}$$

where

$$\beta_{ij} = \begin{cases} G_S(i,j)G_I(i,j) & i \notin \Omega_d, j \in \Omega_i \\ G_S(i,j)G_I(i,j)G_Z(i,j)G_N(i,j) & i \in \Omega_d, j \in \Omega_i \end{cases}$$

and Ω_d states pixels with valid depth value and Ω_i defines subset of pixels who are neighbor to pixel i. As mentioned, G_S , G_I, and G_Z are the probability density function in the spatial, color, and depth domain. Beside these parameters which focus on the depth and color images independently, G_N applies the theorem that difference between normal vectors of the points that are on a same plane has to be minimum. The mentioned probability density functions define as

$$G_S = exp(\frac{-||i-j||^2}{\sigma_S^2})$$

$$G_I = exp(\frac{-||I(i)-I(j)||^2}{\sigma_I^2})$$

$$G_Z = exp(\frac{-||Z(i)-Z(j)||^2}{\sigma_Z^2})$$

$$G_N = exp(\frac{-||N_Z(i)-N_Z(j)||^2}{\sigma_N^2})$$

with variances σ_S, σ_I , σ_Z, and σ_N. I is intensity, Z is the depth, and N_Z is the normal vector of each pixel. The variance controls the effective area of similarity in each of the spaces. In sum, B_{ij} describes the pairwise relation between pixels i and j. This weight includes the difference of normal vectors when there exists valid depth value for pixel i and helps to include structure of the scene in our computations.

6.3 Experimental Results

The presented method is implemented and tested under the linux OS using OpenCV and Point Cloud Library (PCL). To evaluate the results of the proposed method, Middlebury datasets [Mi] that simulates Kinects' characteristic is used. In the problem of denoising Kinect depth map, a ground truth of the depth map is required

Table 6.1 Comparing Mean Absolute error of the proposed method with Chen et al. [ChEtAl15] on the Middlebury datasets

Dataset	Chen et al. (JTF) [ChEtAl15]	Proposed Method
Art	0.0073	0.0050
Book	0.0110	0.0087
Doll	0.0064	0.0043
Laundry	0.0229	0.0225
Moebius	0.045	0.044
Reinder	0.0061	0.0037

Fig. 6.3 A simulated depth input from the Middlebury dataset (Black points are the holes and no value exists for them)

Fig. 6.4 Recovered depth map for Figure 6.3

to figure out accuracy of the hole filling and denoising. Table 6.1 illustrates the Mean Absolute Error (MAE) of the depth recovery on Middlebury datasets.

According to the reported results, including structure of the scene in computations using plane primitives helps to reduce the MAE and have a better understanding of the scene. This reduction is due to the characteristic of selecting a better supporting regions for pixels in depth map and giving a more realistic pairwise weights using the normal vector of supporting plane of pixels. To have a better comparison, Figure 6.3 is an input depth map with a number of holes on it and Figure 6.4 shows the result of applying the proposed method on that input depth map.

To have a better comparison, Figure 6.5 shows a part of the result of applying Chen et al. [ChEtAl15] method on Figure 6.3, and Figure 6.6 shows result of the

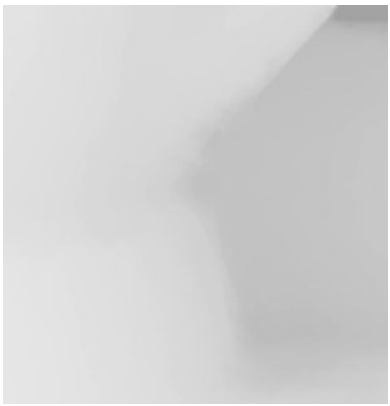


Fig. 6.5 Focusing on a part of the recovered depth map of Figure 6.3 using Chen et al. [ChEtAl15] method

Fig. 6.6 Focusing on a part of the recovered depth map of Figure 6.3 using the proposed method

proposed method applied on that part. It illustrates that using structure of the scene besides the information that extracts from local filters helps to extract edges of depth map accurately. Comparing Figure 6.7 and Figure 6.8 also shows that the proposed method is able to detect holes in the ring and fix that parts. Since the points in the hole of the ring are not in the sample plane of the ring, they will not have effected by the points that are on the ring using the proposed method.

Figure 6.9 shows another depth input and results of applying Chen et al. and proposed method are illustrated in Figure 6.10 and Figure 6.11, respectively. Comparing the outputs, it is again clear that our method performs well on edges and keeps them by looking at the scene structure, while Chen et al. [ChEtAl15] method blurs the edges.

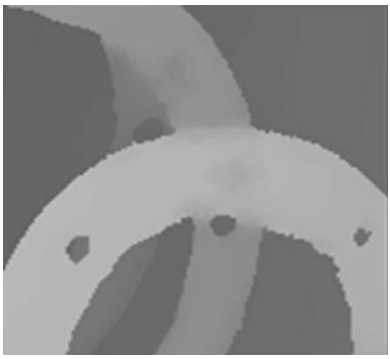


Fig. 6.7 Focusing on a part of the recovered depth map of Figure 6.3 using Chen et al. [ChEtAl15] method

Fig. 6.8 Focusing on a part of the recovered depth map of Figure 6.3 using the proposed method

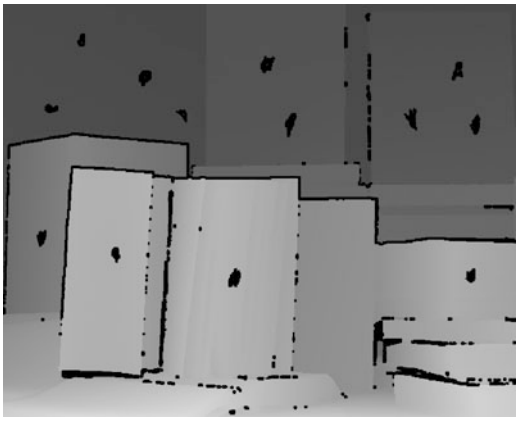


Fig. 6.9 A simulated depth input from the Middlebury dataset (black points are the holes and no value exists for them)

6.4 Conclusion

In this chapter, a novel approach for Kinect depth recovery based on both scene structure and the guidance of local filters based on color image and depth map is presented. Modeling scene structure using planes helps to get an initial guess for points with damaged or unknown depth value. Analyzing results shows that our

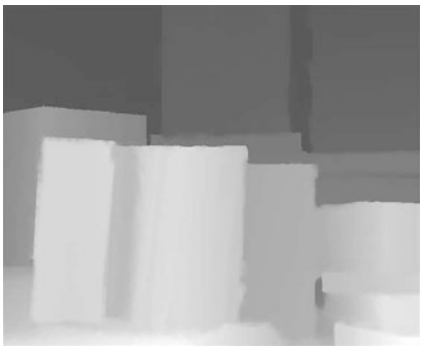


Fig. 6.10 Focusing on a part of the recovered depth map of Figure 6.7 using Chen et al. [ChEtAl15] method

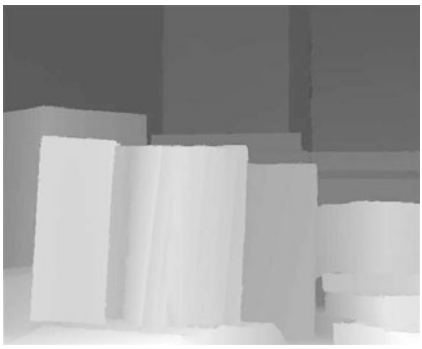


Fig. 6.11 Focusing on a part of the recovered depth map of Figure 6.7 using the proposed method

method is able to keep edges and also detects supporting regions of similar pixels perfectly. As the future work, we are going to model the scene using some other primitives like sphere and also benefit from deep Convolutional Neural Networks (CNN) to understand the model of both RGB image and the depth map.

References

[CaSa12] Camplani, M., Salgado, L.: Efficient spatio-temporal hole filling strategy for kinect depth maps. In: IST/SPIE Electronic Imaging, SPIE Proceedings, vol. 8290, pp. 82900E–82900E. International Society for Optics and Photonics (2012)

[ChEtAl12] Chen, L., Lin, H., Li, S.: Depth image enhancement for Kinect using region growing and bilateral filter. In: 21st International Conference on Pattern Recognition (ICPR), pp. 3070–3073. IEEE (2012)

[ChEtAl13] Chen, C., Cai, J., Zheng, J., Cham, T.J., Shi, G.: A color-guided, region-adaptive and depth-selective unified framework for Kinect depth recovery. In: IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 007–012. IEEE (2013)

[ChEtAl15] Chen, C., Cai, J., Zheng, J., Cham, T.J., Shi, G.: Kinect depth recovery using a color-guided, region-adaptive, and depth-selective framework. ACM Trans. Intell. Syst. Technol. (TIST) **6**(2), 12 (2015)

[CoEtAl15] Alehdaghi, M., Esfahani, M.A., Harati, A.: Parallel RANSAC: speeding up plane extraction in RGBD image sequences using GPU. In: 5th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 295–300. IEEE (2015)

[ChEtAl16] Chen, X., Kristian H., Yinhai, W.: Kinect–based pedestrian detection for crowded scenes. Comput. Aided Civ. Inf. Eng. **31**(3), 229–240 (2016)

[DoEtAl10] Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1141–1148. IEEE (2010). ISO 690

[FiBo81] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)

[FrEtAl13] Freedman, B., Shpunt, A., Machline, M., Arieli, Y.: U.S. Patent No. 8,493,496. U.S. Patent and Trademark Office, Washington, DC (2013)

[KoEtAl07] Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. ACM Trans. Graphics (TOG) **26**(3), 96 (2007)

[LiEtAl10] Liu, S., Lai, P., Tian, D., Gomila, C., Chen, C.W.: Joint trilateral filtering for depth map compression. In: Visual Communications and Image Processing, SPIE Proceedings, vol. 7744, pp. 77440F–77440F. International Society for Optics and Photonics (2010). ISO 690

[Mi] Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007). Minneapolis (2007)

[RiEtAl12] Richardt, C., Stoll, C., Dodgson, N.A., Seidel, H.P., Theobalt, C.: Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. Comput. Graph. Forum **31**(2pt1), 247–256 (2012)

[ScEtAl07] Schnabel, R., Wahl, R., Klein, R.: Efficient RANSAC for point cloud shape detection. Comput. Graphics Forum **26**(2), 214–226 (2007)

[ToMa98] Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Sixth International Conference on Computer Vision, pp. 839–846. IEEE (1998)

[YaEtAl07] Yang, Q., Yang, R., Davis, J., Nistér, D. Spatial-depth super resolution for range images. In: Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)

[ZhEtAl17] Zhu, M., Canjun, Y., Wei, Y., Qian, B.: A Kinect-based motion capture method for assessment of lower extremity exoskeleton. In: Wearable Sensors and Robots, pp. 481–494. Springer, Singapore (2017)

Chapter 7
On the Neutron Point Kinetic Equation with Reactivity Decomposition Based on Two Time Scales

C.E. Espinosa, B.E.J. Bodmann, and M.T. Vilhena

7.1 Introduction

Neutron point kinetic models are used to simulate transient behaviour of nuclear reactors, relevant for reactor control [OkEtAl13]. Typically, transients are considered for short-time intervals only, up to $10^1\ s$ [Ry03]. The present discussion is an extension to these type of models, where reactivity is decomposed in a short- and a long-term contribution. The first one represents operational reactor control, whereas the second one is due to the change of the chemical composition of the nuclear fuel as a consequence of burnup [Se07]. As a first step into a new direction we consider only the effects of the principal neutron poisons on neutron kinetics, i.e., Xe-135 and Sm-149. Note that the initial condition for Xe-135 and Sm-149 implicates whether only new reactor fuel or a fuel composition with reused elements is considered. The proposed model consists in a system of coupled nonlinear equations for the neutron density, the delayed neutron precursors and the neutron poison decay chains. The principal question we address in this work is, what is the influence of the short-time scale characteristics on the long-term behaviour? The equation system is solved using a decomposition method [Ad88], which expands the nonlinear terms in an infinite series, obtaining a recursive system, where the recursion initialization is a homogeneous linear equation and the subsequent recursion steps consider the nonlinear contributions as source terms that are constructed from the solutions of the previous recursion steps.

C.E. Espinosa (✉) • B.E.J. Bodmann • M.T. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: eduardo.espinosa@ufrgs.br; bardo.bodmann@ufrgs.br; mtmbvilhena@gmail.com

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_7

7.2 Neutron Poisons

The principal neutron absorbers that are produced from decays of fission products are Xenon-135 and Samarium-149 [Re08]. Xenon-135 has a cross section for neutron absorption of 2.6×10^6 barns and is produced predominantly (95%) in the Tellurium-135 decay chain from Iodine-135 decay. One way consider a simple model with bulk Iodine yield γ_I only, so that the simplified decay chain is then given by

$$\frac{dC_I(t)}{dt} = \gamma_I \Sigma_f \bar{\nu} n(t) - \lambda_I C_I(t) \ ,$$
$$\frac{dC_{Xe}(t)}{dt} = \gamma_{Xe} \Sigma_f n(t) + \lambda_I C_I(t) - \lambda_{Xe} C_{Xe}(t) - \sigma_{Xe} C_{Xe}(t) \bar{\nu} n(t) \ ,$$

where C_I and C_{Xe} are the Iodine-135 and Xenon-135 concentrations, respectively, γ_I and γ_{Xe} are the production yields by fission, Σ_f is the macroscopic fission cross section, n is the neutron density, λ_I and λ_{Xe} are the decay constants of Iodine-135 and Xenon-135 and σ_{Xe} is the microscopic absorption cross section of Xenon-135.

Samarium-149 has an absorption cross section of 4.1×10^4 barns and is produced in the decay chain of Neodymium-149 which decays and generates Promethium-149, and finally Samarium-149. In a simplified fashion, one may model Samarium-149 production assuming a bulk yield of Promethium by fission so that the Samarium production is given by the equations

$$\frac{dC_{Pm}(t)}{dt} = \gamma_{Pm} \Sigma_f \bar{\nu} n(t) - \lambda_{Pm} C_{Pm}(t) \ ,$$
$$\frac{dC_{Sm}(t)}{dt} = \lambda_{Pm} C_{Pm}(t) - \sigma_{Sm} C_{Sm}(t) \bar{\nu} n(t) \ ,$$

where C_{Pm} and C_{Sm} are the concentration of Promethium-149 and Samarium-149, respectively, γ_{Pm} is the production by fission of Promethium-149, λ_{Pm} is the decay constant of Promethium-149 and σ_{Sm} is the microscopic absorption cross section of Samarium-149.

7.3 Point Kinetics with Poisons

The extended point kinetics model contains besides the usual coupled neutron density and delayed neutron precursor equations also the afore-presented equations that represent neutron poison effects on the neutron population.

$$\frac{d}{dt} n(t) = \frac{\rho_s(t) - \bar{\beta}}{\Lambda} n(t) + \lambda C(t) - \sigma_{Xe} \bar{\nu} n(t) C_{Xe}(t) - \sigma_{Sm} \bar{\nu} n(t) C_{Sm}(t)$$

$$
\begin{aligned}
\frac{d}{dt}C(t) &= \frac{\bar{\beta}}{\Lambda}n(t) - \lambda C(t) \\
\frac{d}{dt}C_I(t) &= \gamma_I \Sigma_f \bar{v} n(t) - \lambda_I C_I(t) \\
\frac{d}{dt}C_{Xe}(t) &= \gamma_{Xe} \Sigma_f \bar{v} n(t) + \lambda_I C_I(t) - \lambda_{Xe} C_{Xe}(t) - \sigma_{Xe} C_{Xe}(t) \bar{v} n(t) \\
\frac{d}{dt}C_{Pm}(t) &= \gamma_{Pm} \Sigma_f \bar{v} n(t) - \lambda_{Pm} C_{Pm}(t) \\
\frac{d}{dt}C_{Sm}(t) &= \lambda_{Pm} C_{Pm}(t) - \sigma_{Sm} C_{Sm}(t) \bar{v} n(t)
\end{aligned}
\tag{7.1}
$$

7.4 Solution by Decomposition

The extended point kinetics model system (7.1) may be casted in matrix form where for convenience we separated linear from nonlinear contributions

$$
\frac{d}{dt}\mathbf{Y} = \mathbf{A}\,\mathbf{Y} + \mathbf{N}\,\mathbf{Y}\,,
$$

with

$$
\mathbf{Y} = (n(t), C(t), C_I(t), C_{Xe}(t), C_{Pm}(t), C_{Sm}(t))^T \ ,
$$

$$
\mathbf{A} = \begin{pmatrix}
\frac{(\rho_s - \bar{\beta})}{\Lambda} & \lambda & 0 & 0 & 0 & 0 \\
\frac{\bar{\beta}}{\Lambda} & -\lambda & 0 & 0 & 0 & 0 \\
\gamma_I \Sigma_f \bar{v} & 0 & -\lambda_I & 0 & 0 & 0 \\
\gamma_{Xe} \Sigma_f \bar{v} & 0 & \lambda_I & -\lambda_{Xe} & 0 & 0 \\
\gamma_{Pm} \Sigma_f \bar{v} & 0 & 0 & 0 & -\lambda_{Pm} & 0 \\
0 & 0 & 0 & 0 & \lambda_{Pm} & 0
\end{pmatrix}
$$

and

$$
\mathbf{N} = \begin{pmatrix}
0 & 0 & 0 & -\sigma_{Xe}\bar{v}n(t) & 0 & -\sigma_{Sm}\bar{v}n(t) \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -\sigma_{Xe}\bar{v}n(t) & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -\sigma_{Sm}\bar{v}n(t)
\end{pmatrix} .
$$

Note that nonlinearity is driven by the neutron density only.

In order to solve the equation system, for convenience we expand the solution into an infinite series $\mathbf{Y} = \sum_{i=0}^{\infty} \mathbf{Y}_i$. These new degrees of freedom allow one to cast the original problem into a recursive scheme, where the recursion initialization is defined by the linear part of the equation system with known solution that obeys the initial conditions of the original problem.

$$\frac{d}{dt}\mathbf{Y}_0 = \mathbf{A}\,\mathbf{Y}_0$$
$$\mathbf{Y}_0(0) = \mathbf{Y}_I$$

Here $\mathbf{Y}_I$ is the vector of non-homogeneous initial conditions. This system has a solution given by

$$\mathbf{Y}_0(t) = \exp(\mathbf{A}t)\mathbf{Y}_I \ .$$

All subsequent recursion steps are then set-up by a linear differential equation system, where the nonlinearity is present as a source term, which is composed from the solutions of the preceding recursion steps

$$\mathbf{Y}_k(t) = \int_0^t \exp(\mathbf{A}(t-\tau))\,\mathbf{F}_k(\mathbf{Y}_0(\tau), \mathbf{Y}_1(\tau), \ldots, \mathbf{Y}_{k-1}(\tau))\,d\tau \ ,$$

with

$$\mathbf{F}_k(\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_{k-1}) = -\left(\mathbf{N}_{k-1}\sum_{j=0}^{k-1}\mathbf{Y}_j + \left(\sum_{j=0}^{k-2}\mathbf{N}_j\right)\mathbf{Y}_{k-1}\right) \ .$$

Once the series by $\mathbf{Y}_k$ is convergent one may truncate the expansion at a finite k such that the solution is within a prescribed precision.

7.5 Numerical Results

In the following we present results for the proposed model and its solution using as initial conditions new fuel elements. Thus, the initial conditions are

$$\begin{aligned}
n(0) &= 1 \quad [cm^{-3}] \ , \\
C(0) &= 100 \quad [cm^{-3}] \ , \\
C_I(0) &= 0 \quad [cm^{-3}] \ , \\
C_{Xe}(0) &= 0 \quad [cm^{-3}] \ , \\
C_{Pm}(0) &= 0 \quad [cm^{-3}] \ , \\
C_{Sm}(0) &= 0 \quad [cm^{-3}] \ .
\end{aligned}$$

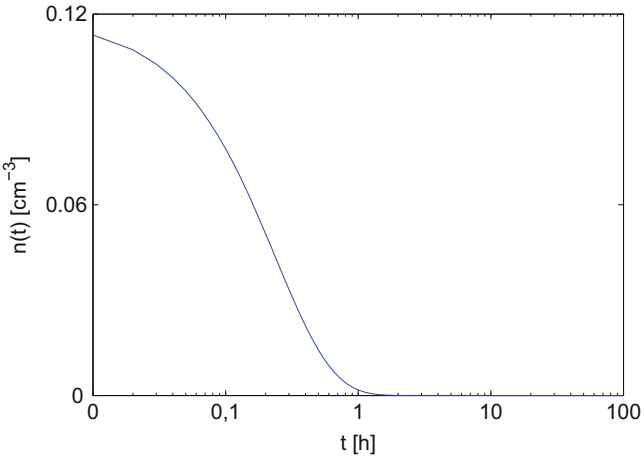


Fig. 7.1 Time dependence of the neutron density $n(t)$ with short-term reactivity $\rho_s = -0.0001t$

We present two cases, one with linear and negative reactivity and a second case with oscillating reactivity, i.e.,

$$\rho_s(t) = \rho_0 t \ ,$$
$$\rho_s(t) = \rho_0 \sin(at) \ ,$$

where ρ_0 and a are constants. We show the solutions for $n(t)$, $C(t)$, $C_{Xe}(t)$ and $C_{Sm}(t)$ for a time interval of 10^3 hours, where in Figures 7.1, 7.2, 7.3, and 7.4 the case for $\rho_s = -0.0001t$ is shown, and Figures 7.5, 7.6, 7.7, and 7.8 show the results for oscillatory reactivity $\rho_s = 0.0001 \sin\left(\frac{2\pi}{12}t\right)$.

For the linear case the neutron density follows the expected shape. Due to an increasing negative reactivity, the neutron density also decreases, as a consequence the fission rate decreases and so does the delayed neutron precursor concentration. In fact, the time evolution of the neutron density and precursor concentration is similar. Because of the time scales that characterize the decay chains that lead to Xenon-135 and Samarium-149 the time evolution of Xenon-135 has an increase and because of its half-life of $t_{1/2}^{(Xe)} = 9.2h$, after a maximum follows a decay curve. Differently for Samarium-149, which is a stable nuclide, the concentration curve increases until an asymptotic limit. Because of a vanishing neutron density there is no mechanism to reduce the C_{Sm} other than neutron absorption.

In the case with oscillatory reactivity, the neutron density as well as the precursor concentration follows the imposed time signature. Since the precursor concentration is produced in decay chains of the fission products, there is a phase difference between the neutron density and the precursor concentration variation. Due to long time scales the Xenon concentration increases until attaining an asymptotic value whereas the Samarium concentration increases without saturation, at least in the considered time interval. The obtained results for both cases are consistent with the physics of the considered problem.

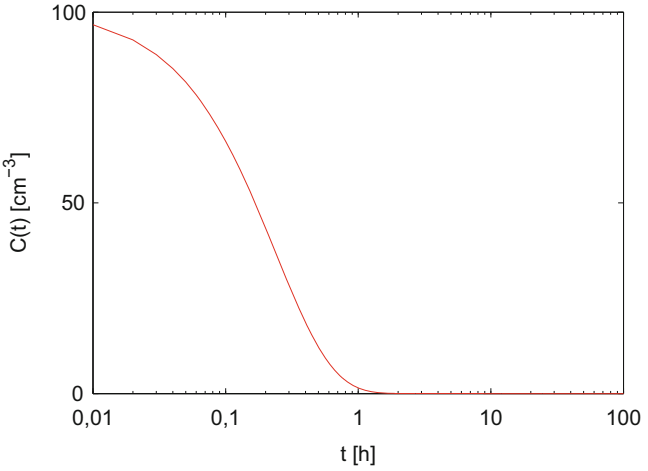


Fig. 7.2 Time dependence of the precursor concentration $C(t)$ with short-term reactivity $\rho_s = -0.0001t$

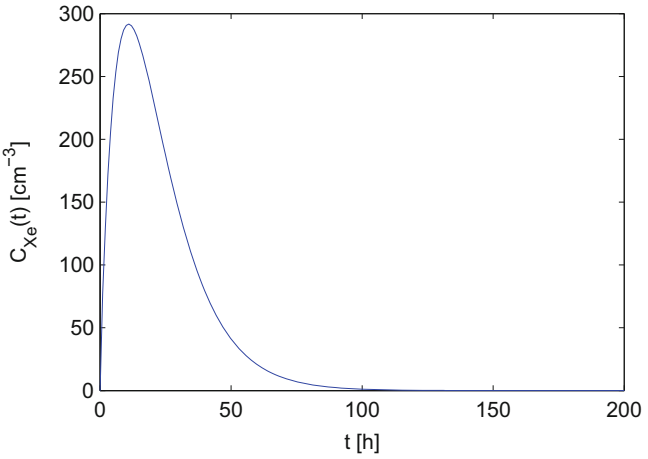


Fig. 7.3 Time dependence of the Xenon concentration $C_{Xe}(t)$ with short-term reactivity $\rho_s = -0.0001t$

7.6 Algorithm Stability

In order to analyse the stability of the truncated solutions of the equation system we use as criterion the ℓ_∞ norm for each quantity (see Table 7.1). To this end, we determine the difference of successive approximations $Y_i - Y_{i-1}$ for each recursion $i = 1$ to $i = 9$. Table 7.2 of successive approximations shows the ℓ_∞ norms for each term of the concentration $n(t)$, $C(t)$, $C_I(t)$, $C_{Xe}(t)$, $C_{Pm}(t)$ and $C_{Sm}(t)$. By inspection one observes that with increasing i the contributions are monotonically decreasing.

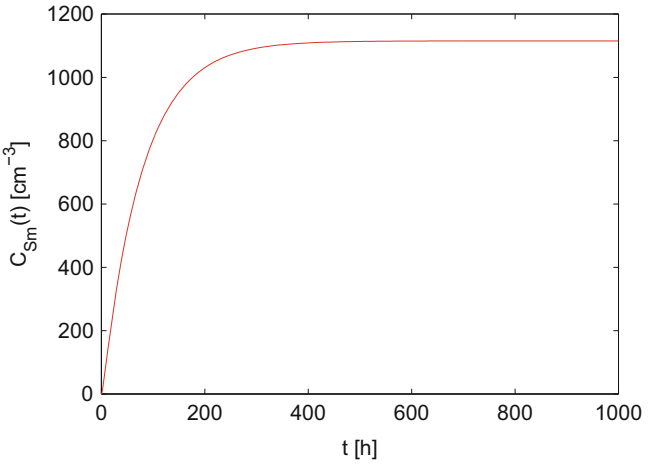


Fig. 7.4 Time dependence of the Samarium concentration $C_{Sm}(t)$ with short-term reactivity $\rho_s = -0.0001t$

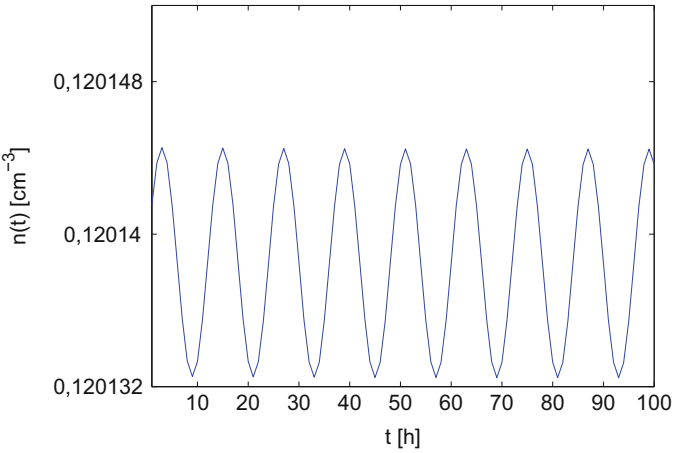


Fig. 7.5 Time dependence of neutron density $n(t)$ with short-term reactivity $\rho_s = 0.0001 \sin\left(\frac{2\pi}{12}t\right)$

Moreover, the norm of the differences of subsequent terms is also decreasing, which makes it plausible, that the solution is convergent. This is also supported by an analysis of the underlying physics of the considered phenomenon.

It is noteworthy that the present method does not impose restrictions on time intervals since the decomposition method determines the solution for each time independently of previous times and thus should work for short-time intervals the same way as for large time intervals independent of any typical time scales of the problem such as half lives or time constants related to short-term reactivity scenarios.

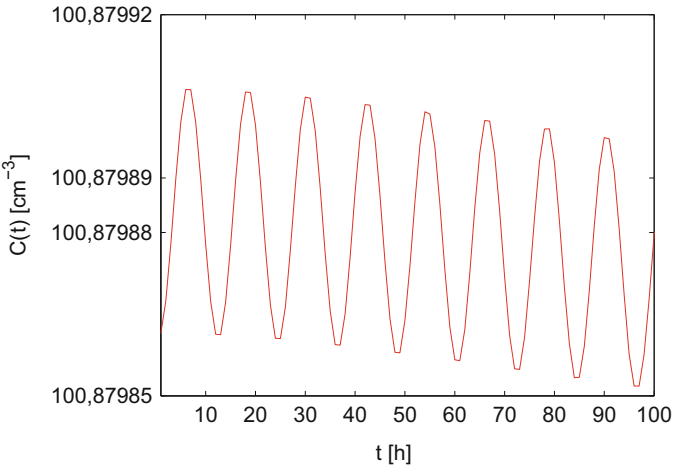


Fig. 7.6 Time dependence of the precursors concentration $C_{Xe}(t)$ with short-term reactivity $\rho_s = 0.0001 \sin\left(\frac{2\pi}{12}t\right)$

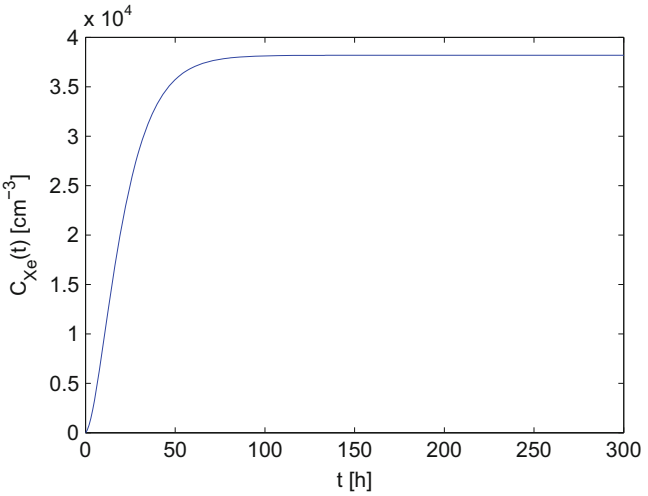


Fig. 7.7 Time dependence of the Xenon concentration $C_{Xe}(t)$ with short-term reactivity $\rho_s = 0.0001 \sin\left(\frac{2\pi}{12}t\right)$

7.7 Conclusions

In the present contribution we proposed a new model for nuclear reactor point kinetics, where besides the usual short-term reactivity changes ($\sim 10^1$ s), that stand for reactor operation effects, also long-term effects by neutron poisons ($\sim 10^4$ s) due to burnup were taken into account. It is noteworthy, that the traditional point

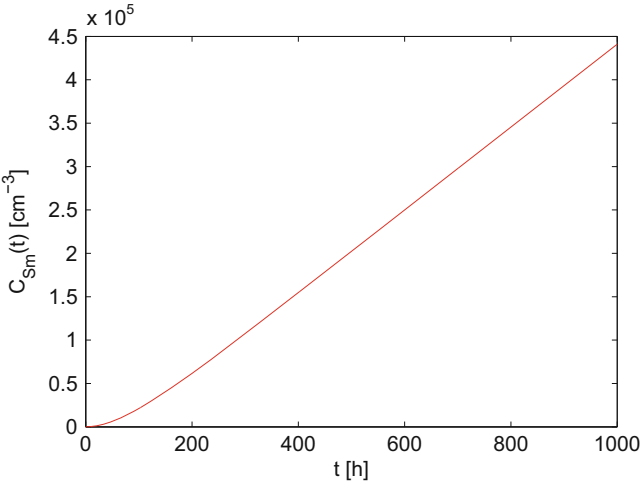


Fig. 7.8 Time dependence of Samarium concentration $C_{Sm}(t)$ with short-term reactivity $\rho_s = 0.0001 \sin\left(\frac{2\pi}{12}t\right)$

Table 7.1 ℓ_∞ norms of $n(t)$, $C(t)$, $C_I(t)$, $C_{Xe}(t)$, $C_{Pm}(t)$ and $C_{Sm}(t)$

i	$\|n_i\|_\infty$	$\|C_i\|_\infty$	$\|C_I^i\|_\infty$	$\|C_{Xe}^i\|_\infty$	$\|C_{Pm}^i\|_\infty$	$\|C_{Sm}^i\|_\infty$
1	0.06572480134	29.968904285	0.0587495689	0.08306582432	0.0714369800	0.2906340211
2	0.02767084870	9.4378906360	0.0182631570	0.02535374161	0.0204477069	0.0570586733
3	0.01031835731	2.8555089831	0.0054609233	0.00745664358	0.0057110088	0.0123106734
4	0.00357952975	0.8396462338	0.0015880676	0.00213511764	0.0015654359	0.0027716243
5	0.00118282910	0.2417877365	0.0004524942	0.00059947512	0.0004231887	0.0006388421
6	0.00037731518	0.0685258514	0.0001269417	0.00016581478	0.0001132054	0.0001494252
7	0.00011716885	0.0191787117	3.517827e-05	4.5327577e-05	3.003590e-05	3.530001e-05
8	3.5619862e-05	0.0053133036	9.652444e-06	1.2273585e-05	7.917164e-06	8.399154e-06
9	1.0643279e-05	0.0014596553	2.626865e-06	3.2974061e-06	2.075763e-06	2.009266e-06

kinetics model is linear, whereas the inclusion of neutron poisons depend on the neutron population and thus turn the model a nonlinear one. The coupled equation system was solved in analytical representation using a decomposition method in the spirit of reference [Ad88], [Ad94]. After 9 recursion steps a solution was obtained that provided accurate results for the neutron density, the precursor and the neutron poison concentrations. The obtained solution allows to calculate transient behaviour of nuclear reactor point kinetics for new reactor fuel as well as fuel compositions with reused fuel elements. We illustrated the model with its solution by two case studies, a negative linearly decreasing reactivity and an oscillatory case, and showed that the found results are in an agreement with physical expectation. In the oscillatory case an unexpected behaviour occurs at small times with a Samarium concentration that exceeds the Xenon concentration, which is considered counter intuitive when evaluated by the involved half lives. Moreover, this finding clearly

Table 7.2 ℓ_∞ of successive approximations norms of $n(t)$, $C(t)$, $C_I(t)$, $C_{Xe}(t)$, $C_{Pm}(t)$ and $C_{Sm}(t)$

i	$\|n_i - n_{i-1}\|_\infty$	$\|C_i - C_{i-1}\|_\infty$	$\|C_I^i - C_I^{i-1}\|_\infty$	$\|C_{Pm}^i - C_{Pm}^{i-1}\|_\infty$	$\|C_{Xe}^i - C_{Xe}^{i-1}\|_\infty$	$\|C_{Sm}^i - C_{Sm}^{i-1}\|_\infty$
1	1.0000000000	130.84876519	0.2600947733	0.3512482772	0.3751268108	3.6634873421
2	0.0933956500	39.406794921	0.0770127259	0.0918846870	0.1084195659	0.3476926945
3	0.0379892060	12.293399619	0.0237240804	0.0261587157	0.0328103852	0.0693693468
4	0.0138978871	3.6951552170	0.0070489910	0.0150822978	0.0095917612	0.0150822978
5	0.0047623589	1.0814339704	0.0020405618	0.0019886247	0.0027345927	0.0034104665
6	0.0015601443	0.3103135880	0.0005794360	0.0005363942	0.0007652899	0.0007882674
7	0.0004944840	0.0877045632	0.0001621200	0.0001432413	0.0002111423	0.0001847253
8	0.0001527887	0.0244920153	4.483072e-05	3.795307e-05	5.760116e-05	4.369917e-05
9	4.626314e-05	0.0067729589	1.227931e-05	9.992928e-06	1.557099e-05	1.040842e-05

shows that there is an influence of small time scales on large time scales and vice versa. In a future work spatial degrees of freedom shall be considered, where the present work may serve as a reference for a hierarchical construction of the model.

References

[Ad88] Adomian, G.: A review of the decomposition method in applied mathematics. J. Math. Anal. Appl. **135**, 501–544 (1988)

[Ad94] Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method, 1st edn. Kluwer Academic Publishers, Athens (1994)

[OkEtAl13] Oka, Y., Suzuki, K.: Nuclear Reactor Kinetics and Plant Control. Springer, Tokyo (2013)

[Re08] Reuss, P.: Neutron Physics. EDP Sciences (2008)

[Ry03] Rydin, R.A.: Nuclear Reactor Theory and Design, 3rd edn. PBS Series in Reactor Physics (2003)

[Se07] Sekimoto, H.: Nuclear Reactor Theory. Institute of Technology, COE-INES, Tokyo (2007)

Chapter 8
Iterated Kantorovich vs Kulkarni Method for Fredholm Integral Equations

R. Fernandes and F.D. d'Almeida

8.1 Introduction

When solving a Fredholm integral equation of the second kind

$$T\varphi - z\varphi = f, \tag{8.1}$$

where $T : X \to X$ is a linear compact integral operator defined by

$$(Tx)(s) := \int_a^b g(|s-t|)x(t)dt, s \in [a,b], \tag{8.2}$$

X being a Banach space and $0 \neq z \in \mathrm{re}(T)$, the resolvent set of T, among the projection methods available, Iterated Kantorovich and Kulkarni's discretization are comparable in terms of convergence rate.

For any $f \in X$, Equation (8.1) has a unique solution $\varphi \in X$, $\varphi = R(z)f$, where $R(z) := (T - zI)^{-1}$.

Projection methods solve the approximate problem

$$(T_n - zI)\varphi_n = f(\text{or } \pi_n f), \tag{8.3}$$

where $T_n : X \to X$ is a bounded linear operator and we take here approximations such that $(T_n)_{n\in N}$ is $\nu-$convergent to T, and so z is in the resolvent set of (T_n), for n

R. Fernandes (✉)
CMat and DMA, Universidade do Minho, Braga, Portugal
e-mail: rosario@math.uminho.pt

F.D. d'Almeida
CMUP and FEUP, Universidade do Porto, Porto, Portugal
e-mail: falmeida@fe.up.pt

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_8

large enough, and there is uniqueness of the solution of the approximate equation

$$\varphi_n = R_n(z)f \text{ (or } \pi_n f), \text{ where } R_n(z) := (T_n - zI)^{-1}$$

(see, for instance, [AhEtAl01]).

To define T_n we use a projection operator π_n onto a finite dimensional subspace X_n such that $\pi_n \xrightarrow{p} I$.

In the following, we recall the definition of classical methods in terms of this projection, defined with the help of a general basis in X_n.

Let $en = (e_i)_{i=1}^n$ contain n linearly independent functions on X, $(e_i^*)_{i=1}^n$ an adjoint basis on X^* defined by

$$\langle e_j, e_i^* \rangle = \delta_{i,j}; \text{ for } i, j = 1, \ldots, n,$$

and $X_n := Span\{e_i : i \in \{1, \ldots, n\}\}$.

The projection π_n is then defined by $\pi_n x := \sum_{j=1}^{n} \langle x, e_j^* \rangle e_j$, for $x \in L^1([a, b])$.

Classical projection methods:

Classical Galerkin approximation: $T_n = T_n^G := \pi_n T \pi_n$

$$T_n^G x = \sum_{j=1}^{n} \sum_{k=1}^{n} < x, e_k^* >< Te_k, e_j^* > e_j;$$

Kantorovich approximation: $T_n = T_n^K := \pi_n T$

$$T_n^K x = \sum_{j=1}^{n} < Tx, e_j^* > e_j;$$

Sloan approximation: $T_n = T_n^S = T\pi_n$

$$T_n^S x = \sum_{j=1}^{n} < x, e_j^* > Te_j.$$

The solution of Equation (8.3) for the methods defined above will be denoted by superscript G, K, S, respectively, for Galerkin, Kantorovich, and Sloan.

The idea of the Kulkarni (Rekha Kulkarni) method is to include in the operator T_n^{RK} the information available in both the operators T_n^K and T_n^S. So its approximation will be defined by (see [Ku03a], [Ku03b], [Ku04])

$$T_n = T_n^{RK} := \pi_n T + T\pi_n - \pi_n T \pi_n = T_n^K + T_n^S - T_n^G$$

and its solution denoted by φ_n^{RK}.

The Iterated Kantorovich [Sl84] is an improvement of the Kantorovich approximate solution using Equation (8.1) thus yielding $\varphi_n^{IK} = \frac{1}{z}(T\varphi_n^K - f)$.

In this work we compare Kulkarni and Iterated Kantorovich methods, focusing on the implementation details and the computational cost of building the matrices needed in the linear systems involved, when applied to a Fredholm integral equation of the second kind with weakly singular kernel (see Section 8.2).

8.2 Details of Implementation in the Case of Weakly Singular Kernels

We will address the case of integral operators where the weakly singular kernel of (8.2) is defined by $g :]0, +\infty[\to \mathbb{R}$ such that

$$g(0^+) = +\infty,$$
$$g \in L^1([0, +\infty[).$$

As an example of such kernel we often take a simplified model of radiative transfer in stellar atmospheres, [Ru04] where g is a multiple of the first exponential integral function E_1 [Ab60],

$$g(\tau) := \frac{\varpi}{2} E_1(\tau) = \frac{\varpi}{2} \int_0^1 \frac{\exp(-\tau/\mu)}{\mu} d\mu, \ \tau > 0,$$
$$a = 0, \quad b = \tau^*, \quad \tau \in [0, \tau^*], \quad \tau^* \in]0, +\infty[.$$

Let the basis $en = (e_j)_{j=1}^n$ for X_n be made of the piecewise constant canonical functions when X is the space of Lebesgue integrable functions,

$$e_j(s) := \begin{cases} 1 & \text{for } s \in [\tau_{j-1}, \tau_j], \\ 0 & \text{otherwise} \end{cases}$$

based on the grid $\mathcal{G}_n := (\tau_j)_{j=0}^n$ such that $\tau_0 := a$, $\tau_n := b$, and $h_j := \tau_j - \tau_{j-1} > 0$.

Its dual basis en^* is made of local mean functionals e_j^* defined by

$$\langle x, e_j^* \rangle := \frac{1}{h_j} \int_{\tau_{j-1}}^{\tau_j} x(t) dt,$$

and $\langle en, en^* \rangle := I_n$, the identity matrix of order n.

The classical methods recalled in the Introduction need the solution of a linear system with coefficient matrix $A_n(i, j) := \langle Te_j, e_i^* \rangle$, $i, j = 1, \ldots, n$, and its vector solution is afterwards used to obtain the solution in the space L^1, by a different formula for each method.

For the Kulkarni method the Equation (8.3), $(T_n^{RK} - zI)\varphi_n^{RK} = f$, can be decomposed into its projection onto X_n and onto $(I - \pi_n)X$:

$$\begin{cases} (\pi_n T - z\pi_n)\varphi_n^{RK} = \pi_n f, \\ (I - \pi_n)(T\pi_n - zI)\varphi_n^{RK} = (I - \pi_n)f. \end{cases}$$

We can decompose correspondingly $\varphi_n^{RK} = \varphi_{n,1}^{RK} + \varphi_{n,2}^{RK}$ and $f = f_1 + f_2$, and obtain

$$\begin{cases} \left(\pi_n T + \frac{1}{z}\pi_n TT - \frac{1}{z}\pi_n T\pi_n T\right)\varphi_{n,1}^{RK} - z\varphi_{n,1}^{RK} = f_1 + \frac{1}{z}\pi_n T f_2, \\ \varphi_n^{RK} = \varphi_{n,1}^{RK} + \frac{1}{z}\left(T\varphi_{n,1}^{RK} - \pi_n T\varphi_{n,1}^{RK} - f_2\right). \end{cases}$$

The first equation can be written as

$$\begin{aligned} &\sum_{k=1}^{n}\sum_{j=1}^{n}\langle\varphi_n^{RK}, e_j^*\rangle\langle Te_j, e_k^*\rangle e_k + \tfrac{1}{z}\sum_{k=1}^{n}\sum_{j=1}^{n}\langle\varphi_n^{RK}, e_j^*\rangle\langle TTe_j, e_k^*\rangle e_k - \\ &-\tfrac{1}{z}\sum_{t=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\langle\varphi_n^{RK}, e_j^*\rangle\langle Te_j, e_k^*\rangle\langle Te_k, e_t^*\rangle e_t - z\sum_{j=1}^{n}\langle\varphi_n^{RK}, e_j^*\rangle e_j = \\ &\sum_{j=1}^{n}\langle f, e_j^*\rangle e_j + \tfrac{1}{z}\sum_{j=1}^{n}\langle Tf_2, e_j^*\rangle e_j. \end{aligned}$$

Applying e_i^*, for $i = 1, \ldots, n$, we get

$$(A_n + \frac{1}{z}B_n - \frac{1}{z}A_nA_n - zI_n)x_n^{RK} = f_n + \frac{1}{z}b_{n,2}, \tag{8.4}$$

where

$$x_n^{RK}(i) := \langle\varphi_n^{RK}, e_i^*\rangle, \quad B_n(i,j) := \langle TTe_j, e_i^*\rangle, \quad f_n(i) := \langle f_1, e_i^*\rangle, \quad b_{n,2}(i) := \langle Tf_2, e_i^*\rangle,$$

for $i, j = 1, \ldots, n$.

We will denote by C_n the coefficient matrix of System (8.4).

After solving System (8.4) the solution φ_n^{RK} is given by

$$\varphi_n^{RK} = \sum_{j=1}^{n} x_n^{RK}(j)e_j + \frac{1}{z}\left(\sum_{j=1}^{n} x_n^{RK}(j)Te_j - \sum_{j=1}^{n}(A_n x_n^{RK})(j)e_j - f_2\right).$$

So, to obtain the Kulkarni approximation the solution of a linear system is required and afterwards the application of T to $\sum_{j=1}^{n} x_n^{RK}(j)e_j$, which can be represented through a pre-multiplication by a matrix representing T on a subspace of much greater dimension.

As for the Iterated Kantorovich the approximation is obtained by using the Equation (8.1) to set a fixed point iteration and perform one step of this, starting with the approximation of Kantorovich, thus yielding

$$\varphi_n^{IK} := \frac{1}{z}(T\varphi_n^K - f) = \frac{1}{z}(T(\sum_{j=1}^{n} x_n^K(j)e_j - \frac{1}{z}f_2) - f). \tag{8.5}$$

The solution of a linear system with A_n as coefficient matrix is needed to obtain x_n^K and afterwards the application of T to $\sum_{j=1}^{n} x_n^K(j)e_j$ is required in Equation (8.5). This can be done through a pre-multiplication by a matrix representing T on a subspace of much greater dimension.

The coefficients of the matrix A_n are given by

$$\begin{aligned} A_n(i,j) &= \langle Te_j, e_i^* \rangle \\ &= \frac{1}{h_i}\int_{\tau_{i-1}}^{\tau_i}\int_{\tau_{j-1}}^{\tau_j} g(|t-\tau|)\,d\tau\,dt, \\ &= \frac{\varpi}{2h_i}\int_{\tau_{i-1}}^{\tau_i}\int_{\tau_{j-1}}^{\tau_j} E_1(|t-\tau|)\,d\tau\;dt, \end{aligned}$$

while the coefficients of B_n are given by

$$\begin{aligned} B_n(i,j) &= \langle TTe_j, e_i^* \rangle \\ &= \frac{1}{h_i}\int_{\tau_{i-1}}^{\tau_i}\int_{0}^{\tau^*} g(|t-\tau|)(Te_j)(\tau)\,d\tau\,dt \\ &= \frac{1}{h_i}\int_{\tau_{i-1}}^{\tau_i}\int_{0}^{\tau^*}\int_{\tau_{j-1}}^{\tau_j} g(|t-\tau|)g(|\tau-s|)\,ds\,d\tau\,dt \\ &= \frac{\varpi^2}{4h_i}\int_{0}^{\tau^*}\int_{\tau_{i-1}}^{\tau_i}\int_{\tau_{j-1}}^{\tau_j} E_1(|t-\tau|)E_1(|\tau-s|)\,ds\,dt\,d\tau, \end{aligned}$$

for $i,j = 1,\ldots,n$.

Error bounds for the relative error of the solution set in terms of $\|(I-\pi_n)T\|$ and $\dfrac{\|(I-\pi_n)f\|}{\|\varphi\|}$ can be seen in [AhEtAl01], and relations of these bounds with the effective basis used in different cases are given in [AhEtAl10, AhEtAl09], and [AlEtAl13], for the classical methods.

In [AlFe16] we proved error bounds for the methods here addressed: for n large enough and $z \neq 0$ in the resolvent set of T,

$$\begin{aligned}\frac{\|\varphi_n^{IK} - \varphi\|}{\|\varphi\|} &\leq 2C\|(I-\pi_n)T\|\,\|(I-\pi_n)T^*\|,\\ \frac{\|\varphi_n^{RK} - \varphi\|}{\|\varphi\|} &\leq 2C\|(I-\pi_n)T\|^2 + \frac{\kappa}{|z|}\|(I-\pi_n)T(I-\pi_n)\|,\end{aligned}$$

where $C := \|(T-zI)^{-1}\|/|z|$ depends on the norm of the resolvent operator, and $\kappa := \|T-zI\|\,\|(T-zI)^{-1}\|$ is the condition number of $T-zI$ relative to inversion.

These bounds show that Iterated Kantorovich and Kulkarni's discretization are comparable in terms of convergence rate.

8.3 Numerical Results

When comparing Iterated Kantorovich and Kulkarni methods we see that they rely on the solution of a linear system of the same dimension n. This can be large to achieve the desired error, but in this illustration we will only deal with small values of n due to the time required by the computation of matrix B_n.

We will use $n = 100$, and in this case we will take $\tau^\star = 50$, or $n = 500$ with $\tau^\star = 100$. The other values of the constants used are $z = 1$ and $\varpi = 0.75$, and the right-hand side function is defined by

$$f(\tau) := \begin{cases} -1 & \text{if } \ 0 \leq \tau \leq \tau^*/2,\\ 0 & \text{if } \ \tau^*/2 < \tau \leq \tau^*.\end{cases}$$

The complexity of the formula of B_n in the case of weakly singular kernels led us to use the software Mathematica (see [Math]) to compute it approximately, in this case Mathematica is not able to compute this triple integral symbolically due to singularities.

In order to compare the CPU times we built A_n with Mathematica too, although in previous works we had computed the integrals (double in this case) analytically and the functions E_n in MatLab (see [MatLab]).

Table 8.1 shows the CPU time for the computation of matrices A_n, B_n, and C_n, as defined in Section 8.2. Remark that the computing time of C_n includes the time of B_n. We show both to stress that most of the time of C_n comes from building B_n.

Table 8.1 CPU time in seconds for the computation of matrices A_n, B_n, and C_n

n	$\mathbf{A_n}$	$\mathbf{B_n}$	$\mathbf{C_n}$
100	127	2165	2292
500	3019	31857	34876

Table 8.2 CPU times in seconds for the three methods

n	**Kantorovich**	**Iterated Kantorovich**	**Kulkarni**
100	0.047	0.091	0.068
500	0.094	0.114	0.182

Table 8.3 CPU time in seconds for the computation of matrices A_n, B_n, and C_n

n	A_n	B_n	C_n
500	3286	13236	16522

Table 8.4 CPU times in seconds for the three methods

n	**Kantorovich**	**Iterated Kantorovich**	**Kulkarni**
500	0.121	0.134	0.190

Table 8.2 reports the CPU times for the Kantorovich, Iterated Kantorovich, and Kulkarni excluding the time to build the coefficient matrix of the linear system A_n or C_n.

The details of implementation described in Section 8.2 can easily be adapted to other kernels. As another example of a weakly singular kernel we will take, for instance,

$$g(\tau) = -\ln(\tau/2),\ \tau \in]0, 2].$$

We will consider $z = 4$,

$$f(\tau) := \begin{cases} -1 & \text{if } 0 \le \tau \le 1, \\ 0 & \text{if } 1 < \tau \le 2, \end{cases}$$

and $n = 500$.

Table 8.3 shows the CPU time, in seconds, required to compute the matrices A_n, B_n, and C_n of this example. They were computed approximately, by the software Mathematica, with formulæ similar to the ones given in Section 8.2 with this kernel replacing $\frac{\varpi}{2}E_1$.

Table 8.4 shows the CPU time, in seconds, required by the iterations of the Kantorovich, Iterated Kantorovich, and Kulkarni methods, excluding the time required to build the matrices A_n, B_n, and C_n, for the $-\ln$ example.

8.4 Conclusion

The Kulkarni and Iterated Kantorovich are comparable in terms of accuracy and computational cost inside each iteration, but for weakly singular kernels and non-self-adjoint operators, the complexity of the formulæ to build the coefficient matrix C_n of the linear system involved in each iteration of Kulkarni method makes it much more expensive in computation time.

We showed that the computation of the matrix C_n, for typical examples of weakly singular kernel, is feasible with Mathematica in the approximate mode, for moderate values of n and therefore not very small values of the grid size, but at a considerable cost in time. MatLab cannot compute the triple integrals of matrix B_n directly. For Iterated Kantorovich method only matrix A_n is needed and this can be done both in Mathematica and Matlab (after long analytical simplifications of the double integrals in the formulæ, as in [AlEtAl13]), which takes much less time. As future work we intend to do similar analytic computations for the triple integrals involved in B_n.

Acknowledgements The first author was partially supported by CMat (UID/MAT/00013/2013) and the second author was partially supported by CMUP (UID/MAT/00144/2013), which are funded by FCT (Portugal) with national (MEC) and European structural funds (FEDER), under the partnership agreement PT2020.

References

[Ab60] Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1960)

[AhEtAl01] Ahues, M., Largillier, A., Limaye, B.V.: Spectral Computations with Bounded Operators. Chapman & Hall/CRC, Boca Raton (2001)

[AhEtAl09] Ahues, M., Amosov, A., Largillier, A.: Superconvergence of some projection approximations for weakly singular integral equations using general grids. SIAM J. Numer. Anal. **47**(1), 646–674 (2009)

[AhEtAl10] Ahues, M., d'Almeida, F.D., Fernandes, R.: Error bounds for L^1 Galerkin approximations of weakly singular integral operators. In: Constanda, C., Pérez, M.E. (eds.) Integral Methods in Science and Engineering, pp. 1–10. Birkäuser, Basel (2010)

[AlEtAl13] d'Almeida, F.D., Ahues, M., Fernandes, R.: Errors and grids for projected weakly singular integral equations. Int. J. Pure Appl. Math. **89**, 203–213 (2013)

[AlFe16] d'Almeida, F.D., Fernandes, R.: Projection methods based on grids for weakly singular integral equations. Appl. Numer. Math. **114**, 47–54 (2017)

[Ku03a] Kulkarni, R.P.: A new superconvergent projection method for approximate solutions of eigenvalues problems. Numer. Funct. Anal. Optim. **24**, 75–84 (2003)

[Ku03b] Kulkarni, R.P.: A superconvergent result for solutions of compact operators equations. Bull. Aust. Math. Soc. **68**, 517–528 (2003)

[Ku04] Kulkarni, R.P.: Approximate solution of multivariable integral equations of the second kind. J. Integr. Equ. Appl. **16**, 343–374 (2004)

[Math] Mathematica, Version 11.0, Wolfram Research, Inc., Champaign (2016)

[MatLab] MATLAB, Version 2014b, The MathWorks, Inc., Natick (2014)

[Ru04] Rutily, B.: Multiple scattering theory and integral equations. In: Constanda, C., Largillier, A., Ahues, M. (eds.) Integral Methods in Science and Engineering, pp. 211–232. Birkäuser, Basel (2004)

[Sl84] Sloan, I.H.: Four variants of the Galerkin method for integral equations of the second kind. IMA J. Numer. Anal. **4**, 9–17 (1984)

Chapter 9
Infiltration Simulation in Porous Media: A Universal Functional Solution for Unsaturated Media

I.C. Furtado, B.E.J. Bodmann, and M.T. Vilhena

9.1 Introduction

In a previous work on infiltration processes in porous media [FuEtAl15] a methodology to construct a parametrised solution for the Richards equation was discussed. The found parametrised solution was given in form of a relatively compact formula, which was validated by one soil type and its associated parameter set. Since the challenge of the problem are the nonlinearity of the Richards equation and the singular initial condition, it was not evident whether the found formula is valid for a representative selection of different soils, i.e. is "universal", which is the principal focus of the present contribution.

To this end we employ the parametrised solution, which was derived in reference [FuEtAl15], and optimise the parameter set by the method of least squares followed by the nonlinear Newton-Raphson method for application to twelve different soils and their associated hydraulic conductivity and capacity, respectively. The best-known model that relates soil parameters to hydraulic conductivity K, matrix potential ψ and soil moisture θ is based on the Van Genuchten model (1980) which can be found in ref. [Ge80]. In this work the authors used these exponential relations for soil-water parametrisation to obtain a solution of an approximate problem for the one-dimensional vertical infiltration case. To the best of our knowledge, no general solution exists in the literature that considers a general soil-water parametrisation as, for instance, the Van Genuchten relations.

I.C. Furtado (✉) • B.E.J. Bodmann • M.T. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: igorjara@gmail.com; bardo.bodmann@ufrgs.br; vilhena@mat.ufrgs.br

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_9

9.2 Modelling Infiltration by the Richards Equation

The governing equation describing infiltration in porous media is a result of a combination of the Darcy-Buckingham and the continuity equation,

$$\vec{q} = -K(\theta)\vec{\nabla}\Phi \quad \text{and} \quad \frac{\partial\theta}{\partial t} = -\vec{\nabla}\vec{q}\,, \tag{9.1}$$

where $\vec{q}$ in (m/s) is the specific flow, $K(\theta)$ in (m/s) is the hydraulic conductivity depending on soil moisture θ and Φ signifies the hydraulic potential in units of (m). Equation (9.1) exhibits, depending on the nonlinear model for the hydraulic conductivity a considerable complexity, when approximate solutions in form of analytical expressions are searched for. Moreover, this equation is established for steady-state condition or dynamical equilibrium. Most infiltration scenarios in nature are transients, and to describe such processes time dependence is introduced by the continuity equation.

For convenience one may split $\Phi = \psi + z$ into the matrix potential that contains the essential effects attributed to porosity, and the gravitational potential represented by the soil depth. Combining the Darcy-Buckingham and continuity equation yields the Richards equation.

$$C(\psi)\frac{\partial\psi}{\partial t} - \vec{\nabla}[K(\psi)\nabla\psi] - \frac{dK(\psi)}{d\psi}\vec{\nabla}\psi = 0 \qquad \psi \in \Gamma \times [0,T] \tag{9.2}$$

Here $C(\psi) = \frac{d\theta}{d\psi}$ is called hydraulic capacity (m^{-1}), $\Gamma \subset \mathbb{R}$ is the physical domain and t represents time. Equation (9.2) is subject to the boundary condition $Q\psi = J$, where Q is a boundary operator and J a known function on the boundary. This equation governs the movement of water in unsaturated soil and can be applied in the whole domain even for distinct saturated and unsaturated areas.

Equation (9.1) needs a closure, so that the system can be solved with one unique solution for ψ. A parametric description of $\theta(\psi)$ and $K(\psi)$ or $K(\theta)$ is used to estimate the otherwise open hydraulic soil-water properties. Mualen [Mu76] derived a functional model for the hydraulic conductivity using the experimental soil retention curve. The usefulness of this model may be associated with the fact that measurements of the hydraulic conductivity are an unsolved challenge. Thus, this model was used as a starting point for further parametrisations as the ones by Van Genuchten and Brooks-Corey [Ge80, BrCo64]. The model by Brooks-Corey is a power law model that introduces a finite value of air entrance, that is associated with the largest pore present, so that beyond this limit the soil is considered saturated. The classic Van Genuchten model describes a function without the value of air entrance. In [VoEtAl01] a modified model of Van Genuchten was presented that incorporates the value of air entrance, resulting in better predictions in comparison to the original model. Note that the modified Van Genuchten model is considered one of the most adequate models to predict hydraulic parametrisations.

In the model of Mualen, the authors assume that pores are connected by distances proportional to the pore radius. In each pore the Poiseuille law is assumed to be valid. Moreover, the tortuosity factor together with moisture content may be represented by a power law containing effective saturation. Thus, the relative permeability is given by

$$K(S_e) = S_e^{\tau} \left[\frac{\int_0^{S_e} \frac{1}{\psi(S)} dS}{\int_0^1 \frac{1}{\psi(S)} dS} \right]^2 , \tag{9.3}$$

where $S_e(\theta) = (\theta - \theta_r)/(\theta_s - \theta_r)$ with θ_s and θ_r the saturated and residual soil-water content. One obstacle of this model resides in the determination of the integral

$$\int_0^{S_e} \frac{1}{\psi(S)} dS = -\int_{\psi(S_e)}^{\infty} \frac{1}{\psi(S)} \frac{dS}{d\psi} d\psi \ . \tag{9.4}$$

Equation (9.3) was derived based on the Poiseuille law and results in a relative conductivity that is dominated by the larger pores. Due to the introduction of air, the model shall exhibit an increase in $dS/d\psi$ such as to regularise $1/\psi$ for $\psi \to 0$. Note that the classical model of Van Genuchten solves also analytically the Mualen model. In the literature the Van Genuchten model is frequently also cited as the Genuchten-Mualen model. The effective saturation may be described in terms of the hydraulic potential as

$$S_e = [1 + (\alpha\psi)^n]^{-m} , \tag{9.5}$$

where n, m and α are parameters to be determined by fit to data. Upon inserting the inversion of equation (9.5) into the integral (9.4) a solution is found where $m = 1 - 1/n$ is used.

$$\int_0^{S_e} \frac{1}{\psi(S)} dS = 1 - (1 - S_e^{1/m})^m$$

$$K(S_e) = S_e^{\tau}[1 - (1 - S_e^{1/m})^m]^2$$

Note that this model does not include explicitly air entrance

$$\frac{dS}{d\psi} = -\alpha mn(\alpha\psi)^{n-1}[1 + (\alpha\psi)^n]^{-(m+1)} ,$$

so that only for $n > 2, dS/d\psi$ decreases sufficiently fast as ψ tends to zero. In other words explicit effects of air entrance are essential for $n < 2$.

A model modified of Van Genuchten that includes air entrance was derived in [VoEtAl01, ScGe06]. The authors introduced a minimum capillarity ψ_s and a heuristic parameter $\theta_m > \theta_s$ without indicating an explicit physical significance.

The resulting model that describes moisture retention is given by

$$\theta(\psi) = \begin{cases} \theta_r + \frac{\theta_m - \theta_r}{(1+|\alpha\psi|^n)^m} & \psi < \psi_s \\ \theta_s & \psi \geq \psi_s \end{cases} \tag{9.6}$$

where $\theta_m = \theta_r + (\theta_s - \theta_r)(1 + |\alpha\psi_s|^n)^m$. Upon combining Equation (9.6) and Equation (9.7), one obtains a model for the modified hydraulic conductivity.

$$K(S_e) = \begin{cases} K_s S_e^{\tau} \left[\frac{1-F(S_e)}{1-F(1)}\right]^2 & \psi < \psi_s \\ K_s & \psi \geq \psi_s \end{cases} \tag{9.7}$$

where $F(S_e) = [1 - (S_e)^{1/m}]^m$ and K_s is a conductivity (m/s) scale for the saturated case. Vogel (2001) [VoEtAl01] suggested for ψ_s to set values between -1 and $-2cm$, and Schaap and co-worker (2006) [ScGe06] obtained a value by optimisation of $-4cm$, where from the author concluded that ψ_s shall be used as an additional fit parameter. For a more complete discussion of this subject see reference [IpEtAl06].

9.3 The Parametrised Solution

From comparison to experimental findings one expects the matrix potential to assume negative values in the range of $[-10m, 0m]$. The solution $\psi_0(z,t)$ of the stationary problem to (9.2) was found to have a predominant contribution [FuEtAl15] and further recursions to improve the solution turned the analytical expression more complicated but providing only spurious corrections, so that they may safely be neglected.

$$\psi_0(z,t) = a_1 \tanh(a_3 z + a_4) + a_2$$

The constants may be found using the Richards equation and minimising the remainder as shown further down for the time dependent problem. Phenomenological arguments allow to extend the stationary solution including a time dependence as follows. With increasing infiltration the surface region approaches local saturation so that the scenario characterised by the initial condition is shifted towards increasing depth. Saturation is already present in the asymptotic behaviour of the hyperbolic tangent function and because of the initial condition ($\psi(z,0) = -10, -L \leq z < 0$) the argument of the tanh function shall be singular for $t = 0$. The simplest way to introduce a shift is adding a term a_4/t to z in the argument of the hyperbolic tangent function. Last, we apply some "adjustment" to our solution by observing that there exists an asymmetry between the convex and concave parts of the profile, i.e. the edge towards the saturated region is sharper than the one at the edge where the matrix potential assumes a numerical value of approximately $-10\ m$. This may be achieved without introducing additional parameters by multiplying the

hyperbolic tangent's argument by an asymmetry factor $1 + \exp\left(a_3 z + a_4 + \frac{a_5}{t}\right)$, that makes use of the previous argument of the hyperbolic tangent function. Thus we arrive at a solution in parametrised form, that we evaluate using the original Richards equation.

$$\psi(z,t) = -a_1 \tanh\left(\left(1 + e^{a_3 z + a_4 + \frac{a_5}{t}}\right)\left(a_3 z + a_4 + \frac{a_5}{t}\right)\right) + a_2 \tag{9.8}$$

Now, the matrix potential ψ is given as a parametrised function $\psi = \psi(z,t;\{a_i\})$ with parameter a_i $(i = 1,2,3,4,5)$, where the unknown parameter set has to be determined.

To adjust the parameter set we insert the parametrised solution (ψ_P) given in expression (9.8) into the governing equation (9.2), which for convenience we write in a form where all terms are on the left-hand side and consequently the right-hand side shall be zero. Let Ω_R be the space-time differential operator that represents the Richards equation with all terms to the left, then for the true solution $\Omega_R[\psi_T] = 0$ holds. Since our solution is an approximate solution the right-hand side differs from zero by a residual term

$$||\Omega_R[\psi_P]|| = \mathcal{R}(z,t). \tag{9.9}$$

Thus, the solution presented in expression (9.8) is optimised minimising $\mathcal{R}(z,t)$ using the method of nonlinear least squares optimisation and refined by the Newton-Raphson method. Some of the constants can be determined *á priori* to optimisation. We can fix the constants a_1 and a_2 directly using the boundary conditions where

$$a_1 = (\psi_P(0,t) - \psi_P(L,t))/2 \qquad \text{and} \qquad a_2 = \psi_P(0,t) - a_1$$

The remaining parameter is determined using the aforementioned minimisation of $\mathcal{R}$.

$$||\Omega_R[\psi_P]|| \to \min.$$

Since the asymptotic behaviour of the solution was fixed using the boundary conditions we use a discrete set of points in the range that contains maximum curvature and the inflection point to optimise $\{a_3, \ldots, a_5\}$. The optimisation may then be simplified using an expansion of the hyperbolic tangent function around the inflection point (at z_0), i.e. where the argument of the function is zero $a_3 z_0 + a_4 + \frac{a_4}{t} = 0$, which allows to solve the minimisation problem in a straightforward fashion.

9.4 Comparison to Benchmark Simulations (HYDRUS) and Self-Consistency Test

The parameter sets that were used to validate the "universality" of the approximate solution formula refer to the twelve types of soils shown in Table 9.1 for situations that consider infiltration of water in a column of initially dry and homogeneous soils. We considered as depth range in soil $[0, L = 1m]$ and the initial and boundary conditions $\psi(z, 0) = -10m$, $-L \leq z \leq 0$; $\psi(0, t) = -0.75m$ and $\psi(-L, t) = -10m$ for $t > 0$. Figures 9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7, 9.8, 9.9, 9.10, 9.11, and 9.12 show the computed stationary matrix potential and the self-consistency test along the vertical coordinate for the parametrised solution. Figures 9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7, 9.8, 9.9, 9.10, 9.11, and 9.12 (left) show the matrix potential with soil depth for the considered soil type. As expected sand soils show the more intense drainage in comparison to the other soils considered. One may also observe that the matrix potential for clay soil has a rather extended region where the transition between unsaturated to saturated soil occurs. This may be attributed to irregular pore shapes of the soil grains and thus voids, where moisture may be retained. Also shown in the figures are comparisons with benchmark results using the HYDRUS software, a program package for simulating water, heat, and solute movement in two- and three-dimensional variably saturated media.

In order to analyse the quality of the found solution the reminder defined in equation (9.9) is shown in Figures 9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7, 9.8, 9.9, 9.10, 9.11, and 9.12 (right). The curves that show self-consistency of the solution with depth indicate by its numerical values that the residue lies between $10^{-2} - 10^{-3}$ and in the sand soil case even as small as 10^{-6} so that one may conclude the determined parametrised solution already reproduces with fair fidelity the exact solution. Moreover, by virtue of the model being an idealisation with its inherent model error no further refinements are of need.

Table 9.1 Soil hydraulic parameter (average values)

Soil	$\theta_r(m^3/m^3)$	$\theta_s(m^3/m^3)$	$\alpha(1/cm)$	n	$K_s(cm/s)$
(1) Sand	0.045	0.430	0.145	2.68	8.25×10^{-2}
(2) Loam sand	0.057	0.410	0.124	2.28	4.05324×10^{-3}
(3) Sand loam	0.065	0.410	0.075	1.89	1.22801×10^{-3}
(4) Loam	0.078	0.430	0.036	1.56	2.88×10^{-4}
(5) Silt	0.034	0.460	0.016	1.37	6.94444×10^{-5}
(6) Silt loam	0.067	0.450	0.020	1.41	1.25×10^{-4}
(7) Sandy clay loam	0.100	0.390	0.059	1.48	3.63889×10^{-4}
(8) Clay loam	0.095	0.410	0.019	1.31	7.22×10^{-5}
(9) Silt clay loam	0.089	0.430	0.010	1.23	1.94444×10^{-5}
(10) Sandy clay	0.100	0.380	0.027	1.23	3.33333×10^{-5}
(11) Silty clay	0.070	0.360	0.005	1.09	5.55556×10^{-6}
(12) Clay	0.068	0.380	0.008	1.09	5.55×10^{-5}
Variation	$O(10^1)$	$O(0)$	$O(10^1)$	$O(10^0)$	$O(10^3)$

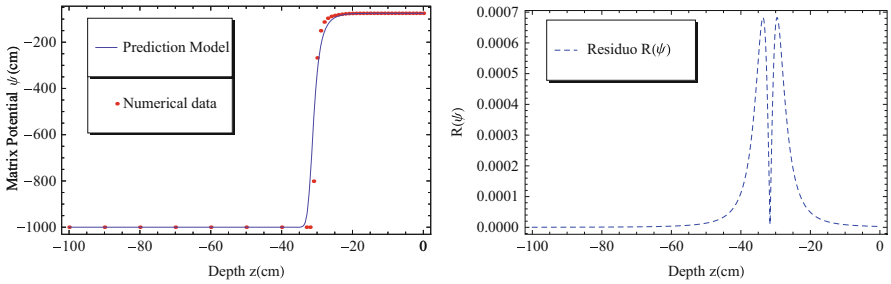


Fig. 9.1 Matrix potential profile and self-consistency with depth for a sand soil (1)

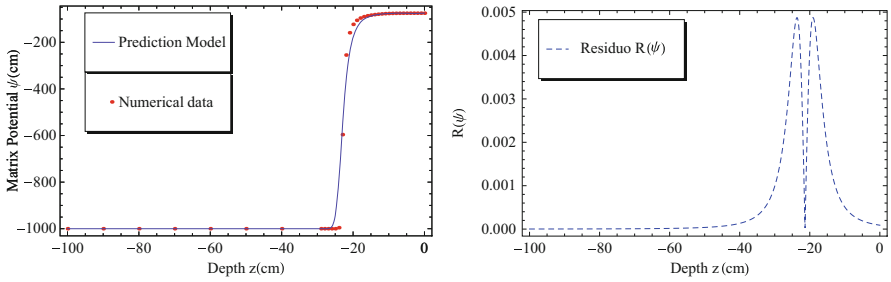


Fig. 9.2 Matrix potential profile and self-consistency with depth for a loam sand soil (2)

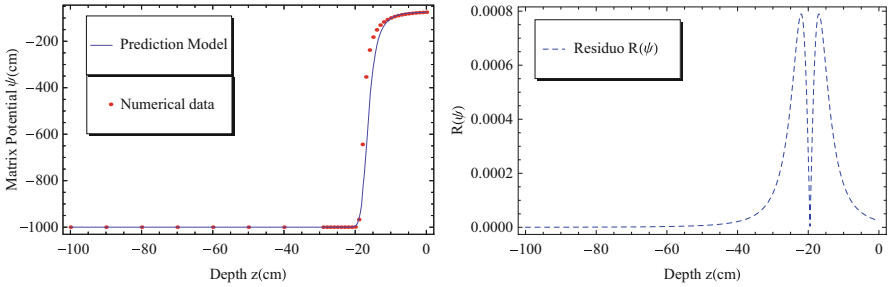


Fig. 9.3 Matrix potential profile and self-consistency with depth for a sand loam soil (3)

9.5 Conclusions

In this contribution, we analysed a problem of transient flow of water in unsaturated media, modelled by the Richards equation. We used the optimised functional solution method for the Richards equation from reference [FuEtAl15] and evaluated its accuracy using the nonlinear Richards equation and defined a self-consistency criterion. A test was performed, comparing the results for the potential matrix by our optimised formula against the profile of benchmark simulations [RaSi06] for twelve types of soil textures. It is remarkable that although the Richards equation is

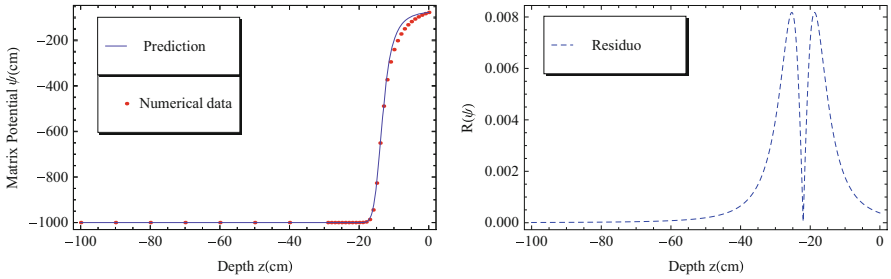


Fig. 9.4 Matrix potential profile and self-consistency with depth for a loam soil (4)

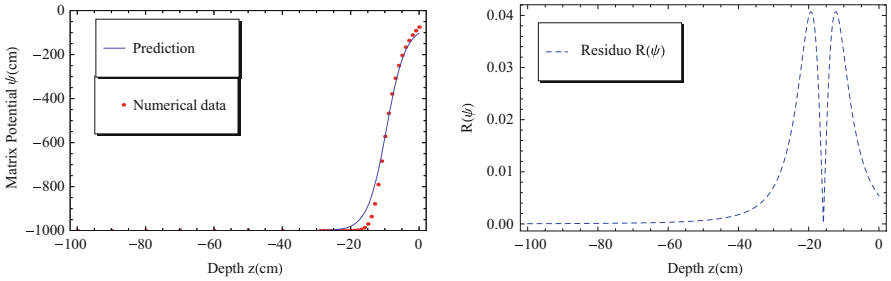


Fig. 9.5 Matrix potential profile and self-consistency with depth for a silt soil (5)

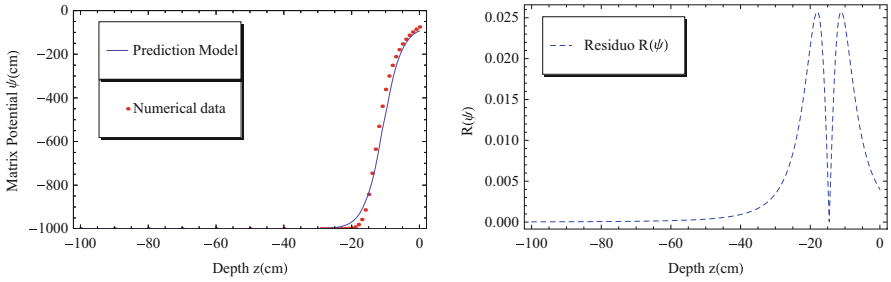


Fig. 9.6 Matrix potential profile and self-consistency with depth for a silt loam soil (6)

highly nonlinear and the initial condition singular, even though the compact solution formula provides fairly good results in all considered soil cases. Thus one may say that for physically relevant soil parameters and for practical purposes one may claim "universality" of our solution formula for the Richards equation. This is even more surprising since some of the soil parameter vary considerably from one soil type to another. Thus, the residual moisture θ_r varies over one order in magnitude, the matrix potential weight α in the effective saturation varies over one order in magnitude, the exponential coefficient n varies by a factor of three and last but not least the saturation hydraulic conductivity varies over three orders in magnitude and there seems to be no need to refine the solution formula by an additional recursion

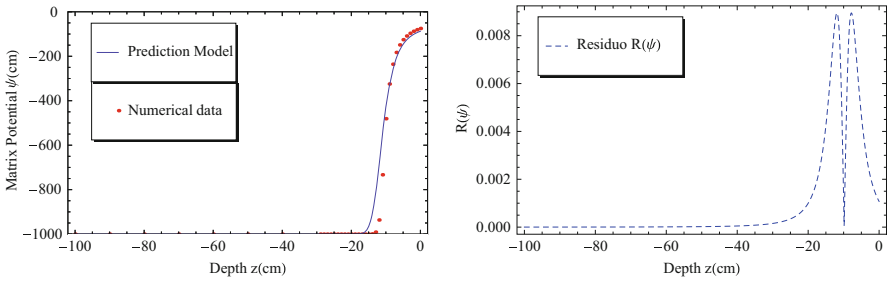


Fig. 9.7 Matrix potential profile and self-consistency with depth for a sandy clay loam soil (7)

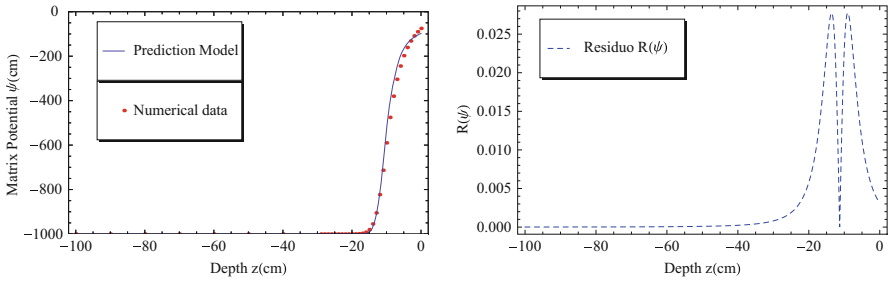


Fig. 9.8 Matrix potential profile and self-consistency with depth for a clay loam soil (8)

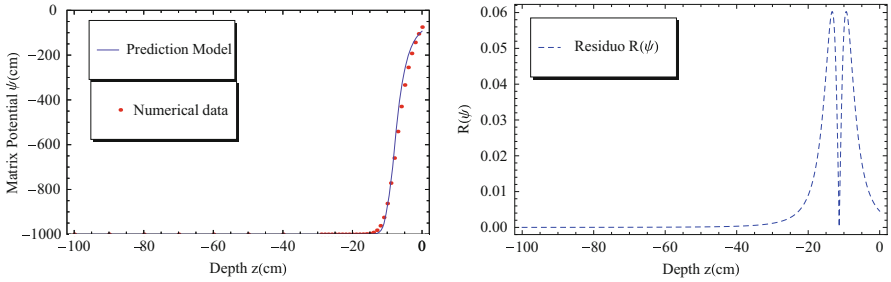


Fig. 9.9 Matrix potential profile and self-consistency with depth for a silt clay loam soil (9)

term (for details see [FuEtAl15]). Within the model error arising from idealisations that lead to the Richards equation, one may safely say our compact formula is capable of efficiently simulating one-dimensional flow of water in unsaturated and saturated porous media.

These conclusions are supported by the observations, that the parametrised solution, which was presented in equation (9.8), when optimised by the method of least squares and Newton-Raphson method, gave fairly good results for the matrix potential profile in all twelve cases as indicated by the self-consistency test which accused only small differences between the true and the parametrised solution. Moreover, even for other soil compositions and their associated parameter

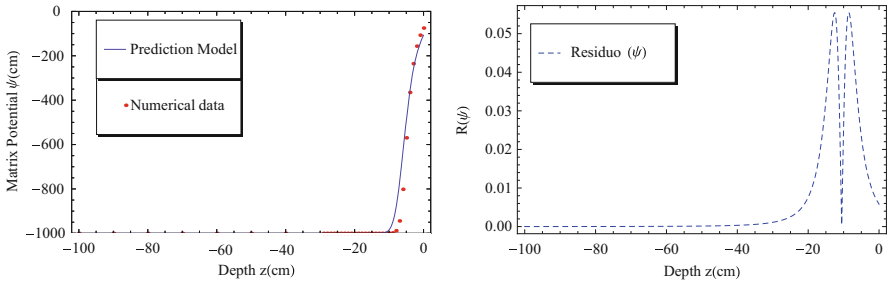


Fig. 9.10 Matrix potential profile and self-consistency with depth for a sand clay soil (10)

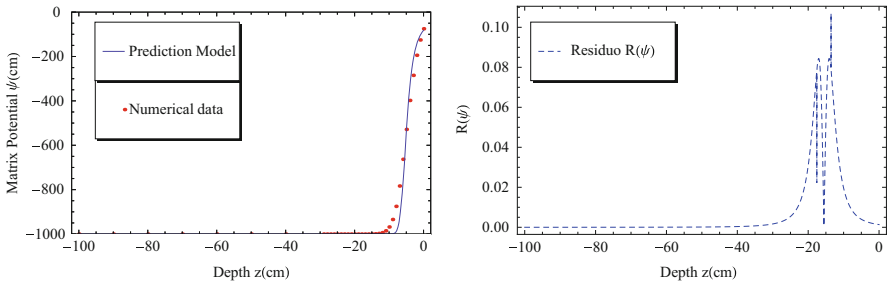


Fig. 9.11 Matrix potential profile and self-consistency with depth for a silt clay soil (11)

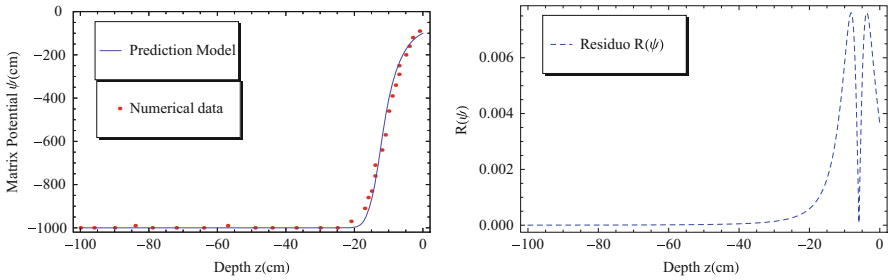


Fig. 9.12 Matrix potential profile and self-consistency with depth for a clay soil (12)

sets not shown in this contribution, the hyperbolic function formula was proven a fairly good approximation. As long as there are no new insights in the problem of infiltration problems in porous media that could alter the structure of the hydraulic conductivity and capacity functions the provided solution formula may be considered a simple and within existing uncertainties sufficiently accurate description of the phenomenon.

References

[BrCo64] Brooks, R.H., Corey, A.T.: Hydraulic Properties of Porous Media. Hydrology Papers, vol. 3 Colorado State University, Fort Collins (1964)

[FuEtAl15] Furtado, I.C., Bodmann, B.E.J., Vilhena, M.T.B.: Infiltration in porous media: on the construction of a functional solution method for the Richards equation. In: Constanda, C., Kirsch, A. (eds.) Integral Methods in Science and Engineering, pp. 235–245. Birkhäuser, New York (2015)

[Ge80] Genuchten, M.T.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. **44**, 892–898 (1980)

[IpEtAl06] Ippisch, I., Vogel, H.J., Bastian, P.: Validity limits for the van Genuchten-Mualen model and implications for parameter estimation and numerical simulation Adv. Water Resour. **29**, 1780–1789 (2006)

[Mu76] Mualen, Y.A.: A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour. Res. **12**, 513–522 (1976)

[RaSi06] Radcliffe, D., Šimůnek, J.: Soil Physics with HYDRUS: Modeling and Applications. CRC/Taylor & Francis Group, Boca Raton (2006)

[ScGe06] Schaap, M.G., Genuchten, M.T.: A modified Mualem-van Genuchten formulation for improved description of the hydraulic conductivity near saturation. Vadose Zone J. **5**(1), 27–34 (2006)

[VoEtAl01] Vogel, T., Genuchten, M.T., Cislerova, M.: Effect of the shape of the soil hydraulic functions near saturation on variably-saturated flow predictions. Adv. Water Resour. **24**, 133–144 (2001)

Chapter 10
Mathematical Models of Cell Clustering Due to Chemotaxis

P.J. Harris

10.1 Introduction

Chemotaxis is the process by which a cell, or cluster of cells, moves in response to an external chemical agent which is diffusing through the fluid medium in which the cells are immersed. The chemical agent may be emitted by the cells themselves to signal their presence to other nearby cells or clusters, or it may just be present in the surrounding medium. Receptors in the cell's outer membrane can react with the chemical and in this way the cell can detect a gradient in the chemical concentration. The cell then moves in the direction in which the concentration is increasing.

In the literature, there are two complementary methods for modeling cell migration and clustering and these can be broadly described as population based or individual based. In population-based methods it is the population densities of cell types within an environment that are considered, rather than the motion of individual cells. These models use diffusion-reaction type equations to simulate how the population density of each type of cell evolves with time (see [Ke71, La74, Ga98, Ch12] for example). Such models are often called Keller-Segel type models in the literature. The alternative is to use an individual-based model where the individual cells (or clusters of cells) are each modeled separately. Such models can range in complexity from assuming each cell can be represented by a simple geometric shape, (see [Ey08, Ki14, Th12] for example) to a mathematical model of how a single cell moves in response to it detecting a chemical signal at its outer membrane (see [El12] for example).

The paper will present a simple model for modeling the motion of cells (or clusters of cells) due to a chemical signal emitted by other nearby cells. A simple linear diffusion equation will be used to model the concentrations of the chemicals

P.J. Harris (✉)
University of Brighton, Brighton, UK
e-mail: p.j.harris@brighton.ac.uk

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_10

and the basic laws of motion will be used to determine how the cells move. In the initial simple model the effect of the fluid will be limited to a simple damping term in the differential equation governing how the cell moves. A more sophisticated model which employs the boundary integral method to model the flow of the fluid surrounding the cells will also be introduced.

10.2 Simple Model

Consider a number of biological cells that are distributed in a surrounding fluid, usually a liquid such as water. Assume that the vertical thickness of the culture is very small (typically of the same size as the dimensions of the cells) so that the cells can only move in two space dimensions and that the i^{th} cell has coordinates $(x_i(t), y_i(t))$. Further, assume that the cells have a simple geometry and can be represented as circles of radius R. If the distance between the centres of two cells is less than $2R$ the cells are assumed to be attached to each other and form a cluster, and once in a cluster the relative position of each cell in the cluster does not change.

Every cell in the culture is capable of emitting a chemical signal which will attract other nearby cells. These other cells are attracted by sensing the gradient of the chemical concentration and moving in the direction in which the concentration is increasing. The changes in the concentrations of the chemical can be modeled using the linear diffusion equation

$$\frac{\partial c_i}{\partial t} = \mu \nabla^2 c_i \tag{10.1}$$

where c_i denotes the concentration of the chemical emitted by the i^{th} cell and μ is the diffusion parameter for the chemical. If the i^{th} cell emits the chemical at time t_i and is located at $(\tilde{x}_i, \tilde{y}_i)$ at the moment the chemical is released, then it is simple to show that the solution to (10.1) can be expressed in the form

$$c_i(x, y, t) = \begin{cases} \dfrac{A_i}{\mu(t - t_i + t_\epsilon)} \exp\left(-\dfrac{(x - \tilde{x}_i)^2 + (y - \tilde{y}_i)^2}{4\mu(t - t_i + t_\epsilon)}\right) & t \geq t_i \\ 0 & t < t_i \end{cases} \tag{10.2}$$

where A_i is amount of the chemical released by the cell and t_ϵ is a small value to avoid computational problems which might arise if $t - t_i$ is zero. Note that the position of $(\tilde{x}, \tilde{y})$ is fixed as the point at which the cell emits the chemical does not change although the cell itself may subsequently move. The total concentration of the chemical in the fluid is simply the sum of the concentrations due to each cell:

$$c(x, y, t) = \sum_{i=1}^{N} c_i(x, y, t)$$

where N is the total number of cells in the culture.

Cells respond to the chemical signals by experiencing a force which is proportional to the gradient of the concentration of the chemical. That is, the force acting on the j^{th} cell is

$$\nabla c(x_j, y_j, t) = \sum_{i=1}^{N} \nabla c_i(x_j, y_j, t)$$

If the cell is part of a cluster, then the total force acting on the cluster is simply the sum of the forces acting on each cell within the cluster. If n_c denotes the number of cells in the current cluster (and $n_c = 1$ for a isolated cell), then the acceleration of each cell in the cluster is given by

$$n_c m \frac{d^2 \mathbf{x}_j}{dt^2} = \sum_{i=1}^{n_c} \nabla c(\mathbf{x}_i, t) - \lambda \frac{d\mathbf{x}_j}{dt} \tag{10.3}$$

where m is the mass of an individual cell and λ is a damping constant used to model the drag due to the fluid. By considering each cluster of cells in turn (10.3) yields a system of second order ordinary differential equations for the locations of the cells and clusters which can be solved using an adaptive fourth order Runge-Kutta scheme.

As the cells move in response to the chemical signal, they will collide with each other. In the model presented here, two cells are taken to have collided when the distance between their centres is less than twice their radii. That is they have collided when

$$|\mathbf{x}_i - \mathbf{x}_j| \leq 2R.$$

where $\mathbf{x}_i$ denotes the position vector of the centre of the i^{th} cell. When the two cells collide it is assumed that they stick together to form a cluster and if they are already part of other clusters then these are combined to form a new single cluster. The velocity of the new cluster is calculated from the velocity of the old cells and/or clusters using a simple conservation of momentum equation, where the momentum of the new cluster is equal to the sum of the momentums of the old clusters that are being combined.

10.3 Boundary Integral Model

A more sophisticated model, which is currently under development, makes use of the boundary integral method to model the motion of the fluid surrounding the cells. Assuming that the fluid is incompressible and inviscid and that the flow is irrotational, then the fluid velocity can be expressed as the gradient of a

scalar potential which, in turn, satisfies Laplace's equation. Since the geometry and velocity $\mathbf{v}$ of each cell is know the boundary condition

$$\frac{\partial \phi}{\partial \mathbf{n}} = \mathbf{v} \cdot \mathbf{n}$$

where ϕ denotes the scalar velocity potential and $\mathbf{n}$ is the unit normal to the cell's surface directed into the fluid. The standard direct boundary integral equation for the velocity potential is

$$-\frac{\phi(\mathbf{p})}{2} + \int_{\Gamma} \frac{\partial G(\mathbf{p},\mathbf{q})}{\partial \mathbf{n}_{\mathbf{q}}} \phi(\mathbf{q})\, dS_{\mathbf{q}} = \int_{\Gamma} G(\mathbf{p},\mathbf{q}) \frac{\partial \phi(\mathbf{q})}{\partial \mathbf{n}_{\mathbf{q}}}\, dS_{\mathbf{q}} \tag{10.4}$$

where Γ denotes the union of the boundaries of all the cells in the fluid and $G(\mathbf{p},\mathbf{q})$ is the free-space Greens function for the Laplace's equation

$$G(\mathbf{p},\mathbf{q}) = \frac{1}{2\pi} \ln(|\mathbf{p}-\mathbf{q}|).$$

For convenience, write (10.4) in operator notation in the form

$$A\phi = B \frac{\partial \phi(\mathbf{q})}{\partial \mathbf{n}_{\mathbf{q}}}$$

where

$$A\phi = -\frac{\phi(\mathbf{p})}{2} + \int_{\Gamma} \frac{\partial G(\mathbf{p},\mathbf{q})}{\partial \mathbf{n}_{\mathbf{q}}} \phi(\mathbf{q})\, dS_{\mathbf{q}}$$

$$B \frac{\partial \phi(\mathbf{q})}{\partial \mathbf{n}_{\mathbf{q}}} = \int_{\Gamma} G(\mathbf{p},\mathbf{q}) \frac{\partial \phi(\mathbf{q})}{\partial \mathbf{n}_{\mathbf{q}}}\, dS_{\mathbf{q}}$$

Let L be the operator which computes the components of the force acting on the cell due to the pressure on the boundary of the cell. That is,

$$\mathbf{F}_i = Lp = \int_{\Gamma_i} p\mathbf{n}\, dS \tag{10.5}$$

where $\mathbf{F}_i$ denotes the force acting on the i^{th} cell and Γ_i denotes the cells boundary. The Bernoulli equation for this problem can be expressed in the form

$$p = -\rho \frac{D\phi}{Dt} + \frac{\rho}{2}|\nabla \phi|^2 + \nabla c \cdot \mathbf{n} \tag{10.6}$$

where $\frac{D\phi}{Dt}$ is the total derivative of the potential (due to the motion of the cell boundary) and the $\nabla c \cdot \mathbf{n}$ term is the pressure on the cell boundary due to the gradient of the chemical signal. Here ρ denotes the density of the fluid, and the absolute

concentration of the chemical is assumed to be small enough to have a negligible effect on the fluid density. Substituting (10.6) into (10.5) yields

$$\mathbf{F}_i = L\left(-\rho\frac{D\phi}{Dt} + \frac{\rho}{2}|\nabla\phi|^2 + \nabla c\cdot\mathbf{n}\right) \tag{10.7}$$

If $\mathbf{a}$ denotes the rigid body acceleration of the cell, and J denotes the operator which gives the normal derivative of the acceleration on the cell boundary in terms of the cell's rigid body acceleration, then the boundary integral method can be used to rewrite (10.7) as

$$m\mathbf{a} = L\left(-\rho A^{-1}BJ\mathbf{a} + \frac{\rho}{2}|\nabla\phi|^2 + \nabla c\cdot\mathbf{n}\right) \tag{10.8}$$

which can be rearranged to make the acceleration the subject of the equation. Equation (10.8) can be integrated through time using a suitable method, such as a Runge-Kutta method, to model the motion of the cells in the culture. The concentrations of the chemical signal are still given by (10.2) in this model which does not take the motion of the fluid into account. A model for the full convection and diffusion of the chemical signal is currently being developed. When (10.8) is discretized the various operators can be replaced by their matrix equivalents.

When the cells collide in this model, rather than forming clusters of individual cells a cluster is simply represented by what is effectively a new larger cell. This is to avoid the situation where two cells are touching at a single point and points on the boundaries of different cells become very close together. This leads to the well-known problems with the boundary integral method that occur when the boundaries of two different domains are too close to each other. If two cells have radii R_i and R_j, respectively, then they will have collided if

$$|\mathbf{x}_i - \mathbf{x}_j| \le R_i + R_j$$

Assuming that all cells have the same mass density, the new cell created when two cells are combined will have the following radius, location and velocity:

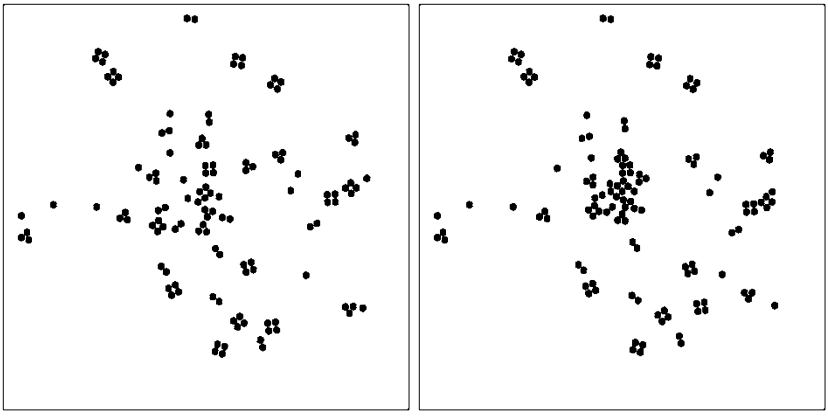
$$R_{new} = \sqrt{R_i^2 + R_j^2} \quad \text{Conservation of mass}$$

$$\mathbf{x}_{new} = \frac{R_i^2\mathbf{x}_i + R_j^2\mathbf{x}_j}{R_i^2 + R_j^2} \quad \text{Same Centre of mass}$$

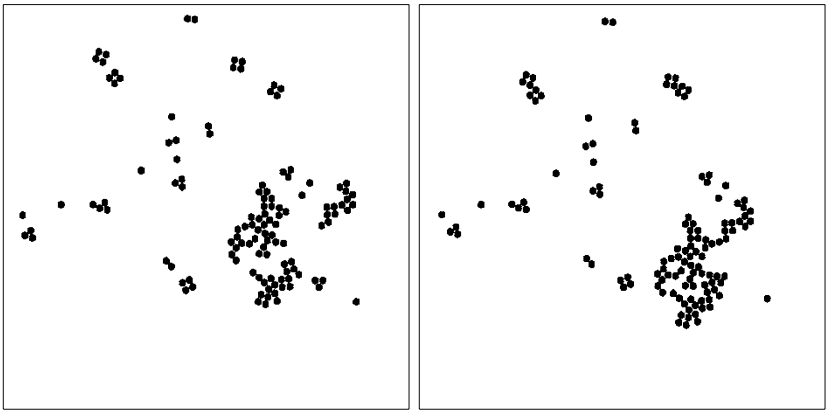
$$\mathbf{v}_{new} = \frac{R_i^2\mathbf{v}_i + R_j^2\mathbf{v}_j}{R_i^2 + R_j^2} \quad \text{Conservation of momentum.}$$

10.4 Numerical Results

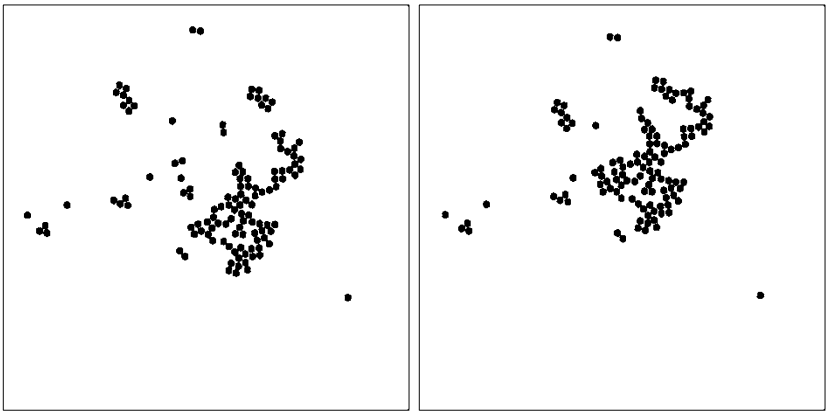
This section illustrates the mathematical models developed above for some typical examples. Figure 10.1 shows the results of using the simple model introduced in Section 10.2 to show how cells can cluster together in a culture. Here there are



(a) $t = 0$ (b) $t = 100$



(c) $t = 200$ (d) $t = 300$



(e) $t = 400$ (f) $t = 500$

Fig. 10.1 The locations of the cells at different times using the simple model

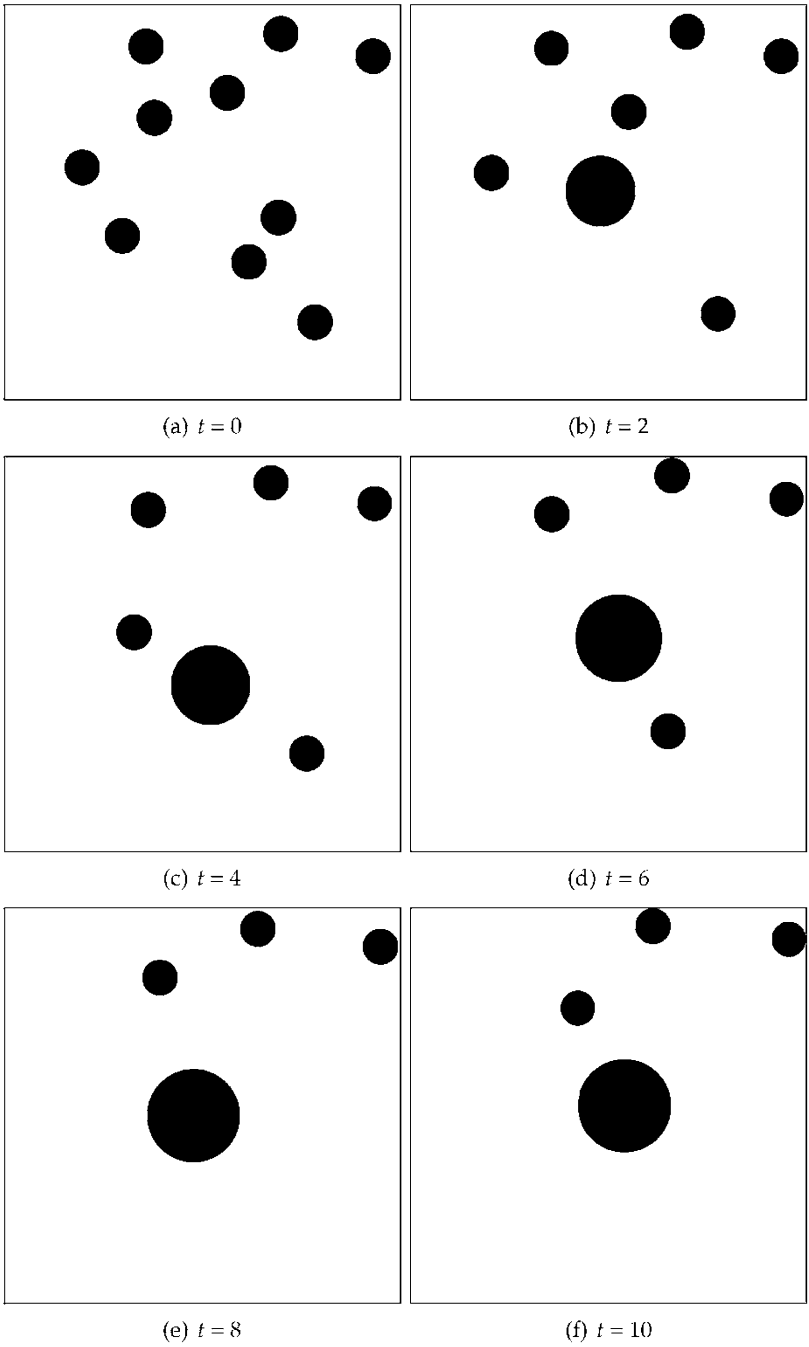


Fig. 10.2 The locations of the cells at different times using the boundary integral model

initially 123 cells in fifty clusters of 1, 2 3 or 4 cells. The location, number of cells and orientation for each of the original clusters were randomly chosen. The results show that as time progresses the cells and clusters combine to create larger clusters. The results for using the boundary integral model are shown in Figure 10.2. In this case there are only 10 cells due to the increased computational cost associated with the boundary integral method representation of the fluid flow. In addition, the cells are being attracted to a single static chemical gradient located at the centre. As with the simple model described above, the cells are collecting together to form larger clusters of cells.

10.5 Conclusions

The results presented in this paper show that it is possible to develop mathematical models of how biological cells cluster together due to chemotaxis. The simple model utilizes basic equations of motion to determine how the cells move in response to the chemical signals. The early results of using this model show that it is capable of replicating the motion of cells that has been observed in experimental work, and further research to fully demonstrate the accuracy of the model is currently being undertaken.

The more sophisticated boundary integral model can also account for the motion of the fluid surrounding the cells which is neglected by the simple model. However, the model presented here is still under development and in particular this model will need to take the convection of the chemical signals due to the motion of the fluid into account.

References

[Ch12] Chertock, A., Kurganov, A., Wang, X., Wu, Y.: On a chemotaxis model with saturated chemotactic flux. Kin. Rel. Mod. **5**, 51–95 (2012)

[El12] Elliott, C.M., Stinner, B., Venkataraman, C.: Modelling cell motility and chemotaxis with evolving surface finite elements. J. R. Soc. Interface **9**, 3027–3044 (2012)

[Ey08] Eyiyurekli, M., Manley, P., Lelkes, P.I., Breen, D.E.: A computational model of chemotaxis-based cell aggregation. BioSysytems, **93**, 226–239 (2008)

[Ga98] Gajewski, H., Zacharias, K.: Global behavior of a reaction - diffusion system modelling chemotaxis. Math. Nachr **195**, 77–114 (1998)

[Ke71] Keller, E.F., Segel, L.A.: Model for chemotaxis. J. Theor. Biol. **30**, 225–234 (1971)

[Ki14] Kim, M., Reed, D., Rejniak, K.A.: The formation of tight tumor clusters affects the efficacy of cell cycle inhibitors: a hybrid model study. J. Theor. Biol. **352**, 31–50 (2014)

[La74] Lapidus, I.R., Schiller, R.: A mathematical model for bacterial chemotaxis. Biophys. J. **14**, 825–834 (1974)

[Th12] Thompson, R.N., Yates, C.A., Baker, R.E.: Modelling cell migration and adhesion during development. Bull. Math. Biol. **74**, 2793–2809 (2012)

Chapter 11
An Acceleration Approach for Fracture Problems in the Extended Boundary Element Method (XBEM) Framework

G. Hattori, S.H. Kettle, L. Campos, J. Trevelyan, and E.L. Albuquerque

11.1 Introduction

The boundary element method (BEM) is a numerical method especially accurate and stable for fracture problems, acoustic, re-entry corners and stress intensity problems. A strong mathematical formulation allows the BEM to model an arbitrary domain through the discretization of the boundaries only. This is particularly advantageous when modelling infinite and half-space domains. BEM models produce reduced meshes and linear system of equations to be solved, compared to domain discretization methods such as the finite element method (FEM). However, the linear system of equations of a BEM model results in fully populated unsymmetric matrices, while FEM linear system yields in large matrices but they are sparse and symmetric. This difference makes BEM unattractive when dealing with large problems, for instance, 3D fracture problems with multiple cracks.

Some authors have investigated different techniques to overcome this limitation of BEM models. Rokhlin [Rok85] has developed the so-called fast multipole method (FMM), which can reduce the complexity of solving the linear system of equations from $O(n^3)$ to $O(n)$. A good review of the method applied to BEM can be found in [Liu09]. Some authors have explored the use of FMM in BEM for fracture problems. In [NYK99], the FMM is combined to a boundary integral formulation

G. Hattori (✉) • S.H. Kettle • J. Trevelyan
School of Engineering and Computing Sciences, Durham University, Durham, UK
e-mail: gabriel.hattori@durham.ac.uk; s.h.kettle@outlook.com; jon.trevelyan@durham.ac.uk

L. Campos
Federal University of Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, 29075-910, Brazil
e-mail: luca.s.campos@ufes.br

E.L. Albuquerque
Department of Mechanical Engineering, University of Brasilia, Brasilia, Brazil
e-mail: eder@unb.br

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_11

of the Laplace equation for 3D crack problems. Yoshida et al. [YNK01] have used a symmetric Galerkin formulation with FMM, analysing problems with 512 penny shaped cracks.

Although FMM is very efficient, it depends on a multipole expansion that has to be considered by the fundamental solution of a BEM code. For this reason, it is not straightforward to implement if a BEM code is available. However, the adaptive cross approximation (ACA) depends only on geometrical features of the problem, but generates an approximation of the solution instead. The idea of the method is to use the smoothness of the operator to approximate the so-called admissible blocks, thus accelerating the evaluation of the matrix-vector product that lies within each iteration of an iterative solver. Several authors have used ACA for multiple crack problems [GG10], time-domain BEM elasticity [MS10], anisotropic materials [BMA09, BA10] and time-domain BEM for anisotropic materials [MBA12]. However, there are still no works on how ACA behaves when coupled with enriched formulations, such as the extended boundary element method (XBEM).

An enriched formulation of the BEM has been proposed by [ST11] for the first time, where partition of unity has been applied in a similar way as in the extended finite element method (XFEM). The asymptotic behaviour at the crack tip is described more accurately, at the expense of increasing the conditioning of the linear system of equations. Later, Alatawi and Trevelyan [AT15] used an implicit enrichment scheme, where the additional degrees of freedom correspond to the stress intensity factors (SIF). In this case, the number of elements enriched does not affect the number of degrees of freedom, an issue with XFEM formulations and the formulation employed by [ST11]. Additionally, there is no need for post-processing (such as the J-integral) to obtain the SIF, since they are calculated as part of the displacement solution.

In this work we investigate the use of ACA in an XBEM formulation for anisotropic 2D materials, using the formulation obtained by the authors in [HAT16] for anisotropic materials. We detail how ACA can be implemented in a BEM framework, and we present some examples that demonstrate how this technique can be useful to overcome the limitation of solving large linear systems with unsymmetric and fully populated matrices found in BEM.

11.2 Extended Boundary Element Method

A dual BEM formulation is modified with the enrichment in the same way as in [HAT16]. Two boundary integral equations (BIE) are necessary to avoid the mathematical degeneration which arises from the coincidence of the crack faces. The displacement boundary integral equation (DBIE) and the traction boundary integral equation (TBIE) are given by [HAT16]

$$c_{ij}(\boldsymbol{\xi})u_j(\boldsymbol{\xi}) + \int_{\Gamma} p^*_{ij}(\mathbf{x},\boldsymbol{\xi})u_j(\mathbf{x})d\Gamma(\mathbf{x}) + \int_{\Gamma_c} p^*_{ij}(\mathbf{x},\boldsymbol{\xi})\tilde{K}_{lj}F_{lj}(\boldsymbol{\xi})d\Gamma(\mathbf{x}) = \int_{\Gamma} u^*_{ij}(\mathbf{x},\boldsymbol{\xi})p_j(\mathbf{x})d\Gamma(\mathbf{x}) \tag{11.1}$$

$$c_{ij}(\boldsymbol{\xi})p_j(\boldsymbol{\xi}) + N_r\int_{\Gamma} s^*_{rij}(\mathbf{x},\boldsymbol{\xi})u_j(\mathbf{x})d\Gamma(\mathbf{x}) + N_r\int_{\Gamma_c} s^*_{rij}(\mathbf{x},\boldsymbol{\xi})\tilde{K}_{lj}F_{lj}(\boldsymbol{\xi})d\Gamma(\mathbf{x}) = N_r\int_{\Gamma} d^*_{rij}(\mathbf{x},\boldsymbol{\xi})p_j(\mathbf{x})d\Gamma(\mathbf{x}) \tag{11.2}$$

where $\Gamma_c = \Gamma_+ \cup \Gamma_-$ stands for the crack surfaces Γ_+ and Γ_-, N_r is the normal at the observation point $\mathbf{x}$, c_{ij} is the free term coming from the integration of the singular kernels, F_{lj} is the enrichment function and $\tilde{K}_{lj}$ is the additional degree of freedom which stands for the SIF. Let us recall that strongly singular and hypersingular terms arise from the integration of the p^*_{ij}, d^*_{rij} and s^*_{rij} kernels and they have to regularized before numerical integration is possible. More details about the regularization procedure and the XBEM formulation can be found in [AT15].

The enrichment function used in Equations (11.1) and (11.2) are the same as defined in [HRDS+12] for anisotropic materials using the extended finite element method (XFEM) and are given by

$$F_{lj}(r,\theta) = \sqrt{\frac{2r}{\pi}}\begin{pmatrix} A_{11}B_{11}^{-1}\beta_1 + A_{12}B_{21}^{-1}\beta_2 & A_{11}B_{12}^{-1}\beta_1 + A_{12}B_{22}^{-1}\beta_2 \\ A_{21}B_{11}^{-1}\beta_1 + A_{22}B_{21}^{-1}\beta_2 & A_{21}B_{12}^{-1}\beta_1 + A_{22}B_{22}^{-1}\beta_2 \end{pmatrix}$$

where $\beta_i = \sqrt{\cos\theta + \mu_i \sin\theta}$, r is the distance between the crack tip and an arbitrary position, θ is the orientation measured from a coordinate system centred at the crack tip; $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\mu}$ come from the Stroh formalism and depend only on the material properties.

11.3 Adaptive Cross Approximation

The adaptive cross approximation is a technique which combines the concept of hierarchical matrices with low-rank approximation. Hierarchical matrices were first introduced by Hackbusch [Hac99], where the matrix is sub-divided into blocks through a geometrical criteria. A hierarchical tree is constructed using the following algorithm:

1. Find the centre of the current cluster/block;
2. Obtain the covariance matrix of the cluster;
3. Take the eigenvector associated with the largest eigenvalue of the covariance matrix;

4. The cluster is divided into 2 new blocks using the eigenvector;
5. Repeat step 1 for each cluster until a minimum block size is achieved.

These blocks are further classified into admissible and non-admissible. If admissible, the cluster is sufficiently smooth to be approximated, which indicates that a low-rank approximation can be used. This procedure will be detailed later in this section. If the block is not admissible, no approximation can be done, and the elements of the matrix have to be obtained using Equations (11.1) and (11.2).

The minimum block size is a parameter that ultimately defines the number of blocks in the hierarchical tree. If the minimum block size is too large and the operator on the matrix is not reasonably smooth, fewer blocks will be classified as admissible. If the minimum block size is too small, many admissible small blocks will be formed, and the approximation will not be accurate.

In BEM, the smoothness of the matrix will depend whether the field and source nodes are well separated geometrically. The admissibility parameter is defined as

$$\min\left(\text{diam}(Cl_x), \text{diam}(Cl_y)\right) \leq \eta \text{dist}(Cl_x, Cl_y)$$

where $0 < \eta < 1$ is the admissibility parameter; Cl_x and Cl_y are two arbitrary clusters; diam represents the size of the cluster and dist stands for the distance between the clusters. These parameters are given by

$$\begin{aligned}
\text{diam}(Cl_x) &= 2 \max_k |X - x_k| \\
\text{diam}(Cl_y) &= 2 \max_k |Y - y_k| \\
\text{dist}(Cl_x, Cl_y) &= |X - Y| - 0.5\left(\text{diam}(Cl_x) + \text{diam}(Cl_y)\right)
\end{aligned}$$

where X, Y are the average of the cluster and x_k, y_k is an element of cluster x, y, respectively.

Finally, the admissible blocks are approximated using the same criteria as in [BR03]. The main idea is that admissible blocks are approximated by low-rank approximants formed as a series of outer products of row and column vectors. While the FMM deals with the analytical decomposition of the integral kernels, ACA will provide an almost optimal approximation of the original matrix. The approximation of matrix $\mathbf{A} \in C^{t \times S}$ is given by

$$\mathbf{A} \approx \mathbf{S}_k = \mathbf{U}\mathbf{V}^t, \text{ where } \mathbf{U} \in C^{t \times k} \text{ and } \mathbf{V} \in C^{s \times k}$$

where k is a low-rank compared to t and s. It is important to remark that the low-rank representation can only be found when the generating kernel function in the computational domain of $\mathbf{A}$ is asymptotically smooth. It has been shown in [Beb08] that elliptic operators with constant coefficients have this property. A detailed explanation about ACA applied to BEM can be found in [RS07].

The low-rank approximation is obtained by splitting the matrix $\mathbf{A} \in C^{t \times s}$ into $\mathbf{A} = \mathbf{S}_k + \mathbf{R}_k$, where $\mathbf{S}_k$ is the rank k approximation and $\mathbf{R}_k$ is the residuum which has to be minimized, through the following algorithm:

1. Define $k = 0$ where $\mathbf{S}_0 = \mathbf{0}$ and $\mathbf{R}_0 = \mathbf{A}$ and the first scalar pivot to be found is $\gamma_1 = (\mathbf{R}_0)_{ij}^{-1}$, and i, j are the row and column indices of the actual approximation step;
2. For each step υ, obtain

$$\begin{aligned}\mathbf{v}_{\upsilon+1} &= \gamma_{\upsilon+1}(\mathbf{R}_\upsilon)_i \\ \mathbf{u}_{\upsilon+1} &= (\mathbf{R}_\upsilon)_j \\ \mathbf{R}_{\upsilon+1} &= \mathbf{R}_\upsilon - \mathbf{u}_{\upsilon+1}\mathbf{v}_{\upsilon+1}^t \\ \mathbf{S}_{\upsilon+1} &= \mathbf{S}_\upsilon + \mathbf{u}_{\upsilon+1}\mathbf{v}_{\upsilon+1}^t\end{aligned}$$

   where the operators $()_i$ and $()_j$ indicate the i-th row and the j-th column vectors, respectively;
3. The next pivot $\gamma_{\upsilon+1}$ is chosen to be the largest entry in modulus of the row $(\mathbf{R}_\upsilon)_i$ or the column $(\mathbf{R}_\upsilon)_j$;
4. The approximation stops when the following criterion holds:

$$\|\mathbf{u}_{\upsilon+1}\|_F \|\mathbf{v}_{\upsilon+1}\|_F < \varepsilon \|\mathbf{S}_{\upsilon+1}\|_F$$

where $\|.\|_F$ stands for the Frobenius norm.

In this paper the form of ACA used is fully pivoted ACA. While partially pivoted ACA allows for reductions in storage and generation of the system matrix, the subject of this paper is addressing reductions in computations in the solution. The number of operations required for generation and storage is each proportional to $O(n^2)$, n is the order of the matrix. The direct solution of the linear system requires a number of operations proportional to $O(n^3)$. Iterative solvers such as the generalized minimal residual method (GMRES) [SS86] reduce the complexity but involve an expensive matrix-vector product within each iteration.

Performing hierarchical clustering on a dual formulated BEM matrix requires separation of the boundary nodes and the crack nodes at all times [BAD08]. Discontinuous elements are used on the crack, with one surface corresponding to the DBIE and the other corresponding to the TBIE. For this reason, when constructing the hierarchical tree, the crack surfaces must be separated from the external boundaries, and both crack surfaces must be separated from each other in collocation, but can be considered as a whole when being integrated over by for a remote collocation point. Thus it is important to place nodes on opposing crack surfaces in different clusters row-wise, but unimportant column-wise in the matrix. This creates asymmetry in the structure of the hierarchical tree.

Constraints other than the admissibility criterion must be put in place to assure that these conditions are met, although the geometry of crack nodes are usually clearly distinguishable from those of the boundary, this is not always the case. Furthermore, coincident nodes on opposite crack surfaces will be clustered together unless the algorithm is instructed otherwise.

11.4 Results

Figure 11.1 depicts a BEM mesh for a two-dimensional anisotropic plate containing 12 internal cracks under uniform unitary load applied at the top and bottom of the plate. The problem is discretized with 2000 nodes for the external boundaries and 300 nodes per crack. Continuous elements are employed on the external boundaries and discontinuous elements are used on the crack faces. There are 9248 degrees of freedom (DOF) in total, of which 48 DOFs correspond to the stress intensity factors (SIFs) of each crack (mode I and II). The material properties are given by: $C_{11} = 117.97$ GPa, $C_{12} = 14.19$ GPa, $C_{16} = 35.43$ GPa, $C_{22} = 15.64$ GPa, $C_{26} = 7.49$ GPa, $C_{66} = 21.38$ GPa.

Figure 11.2 represents the matrix structure resulting from the hierarchical clustering of the problem.

Table 11.1 illustrates the savings of solving the linear system of equations using ACA. The solution error is defined as

$$\text{error} = \frac{\|\mathbf{x}_{ACA} - \mathbf{x}\|_F}{\|\mathbf{x}\|_F}$$

where $\mathbf{x}_{ACA}$ is the displacement solution when the linear system of equations was approximated with ACA, $\mathbf{x}$ is the displacement solution obtained by solving the full system using a direct solver and $\|.\|_F$ stands for the Frobenius norm. The parameter ε_C is the threshold error of the ACA approximation using the Frobenius norm.

The label 'Operations' stands for the number of computations required to perform a matrix-vector product using the approximated system matrix. This is the total sum of $O(k(N + M))$ for every low-rank block combined with $O(NM)$ for every full rank block, where k represents the rank and N and M represent the rows and columns, respectively. 'Saving' represents the gain in number of computations required to perform the full rank matrix-vector product.

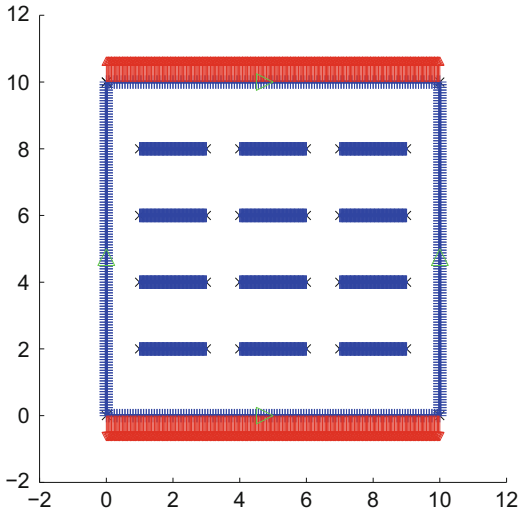


Fig. 11.1 Anisotropic plate containing multiple cracks

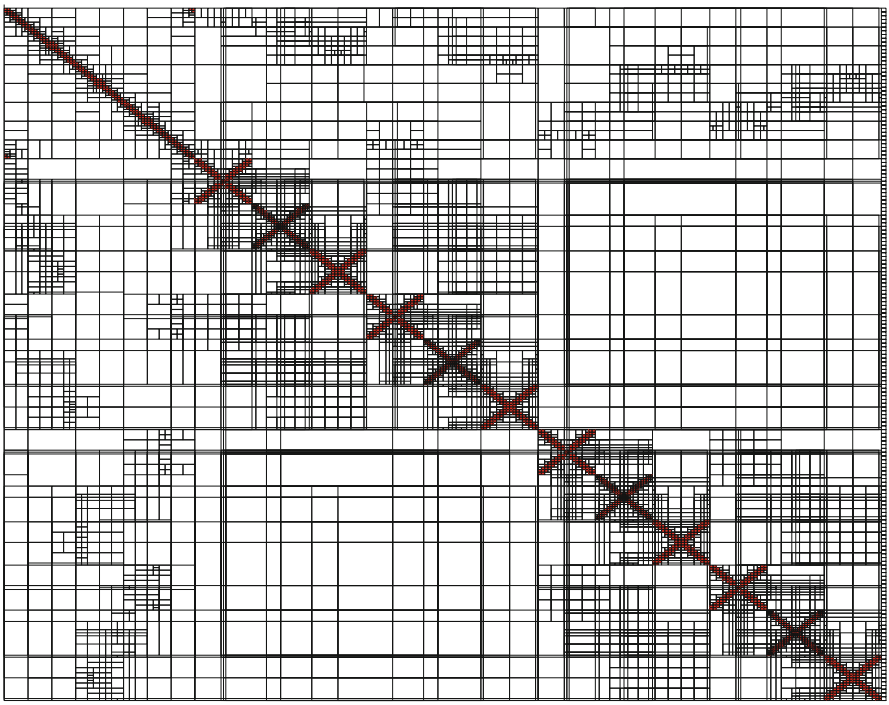


Fig. 11.2 ACA generated matrix partitioning

Table 11.1 Operational savings data

	Solution error			
ε_C	Total (%)	SIFs (%)	Operations	Saving(%)
1×10^{-4}	0.310	0.130	6.562×10^6	92.3
1×10^{-5}	0.022	0.018	7.499×10^6	91.2
1×10^{-6}	1.27×10^{-5}	1.26×10^{-5}	8.449×10^6	90.1

The results show that reductions in ε_C, as expected, improve the accuracy of the eventual solution. However, the number of operations to assemble the approximated matrix will increase, which reduces slightly the saving obtained using ACA.

ACA is successfully applied to all parts of the matrix, with exception of the last 48 rows and columns. For the latter, the columns are partitioned into 50×48 blocks and then approximated using ACA. For the former, these are extra rows necessary to balance the system of equations as shown in [AT15], and contain the so-called tying equations. These rows are sparsely populated, therefore are not considered for ACA. Applying these configurations produces an accurate solution with computational savings in excess of 90% per iteration and low errors for both displacement field and SIFs.

Table 11.2 Operational savings data for different levels of error

ε_C	ε_{SIF}	Solution error		Operations	Saving(%)
		Total (%)	SIFs (%)		
1×10^{-3}	1×10^{-4}	2.995	1.407	5.515×10^6	93.55
	1×10^{-8}	2.995	1.404	5.627×10^6	92.42
1×10^{-4}	1×10^{-4}	0.306	0.127	6.507×10^6	92.39
	1×10^{-8}	0.306	0.125	6.618×10^6	92.26
1×10^{-5}	1×10^{-4}	0.023	0.017	7.471×10^6	91.26
	1×10^{-8}	0.022	0.018	7.583×10^6	91.13

In Table 11.1 the error for the Frobenius norm is the same for boundary displacements and the SIFs in the solution vector. However, it is known that the terms of the linear system of equations associated with the enriched terms can be of very different order of magnitude from the other terms. Moreover, it might be speculated that the accuracy of these terms is strongly influential over the accuracy of the computed SIFs. In this case, we investigate the effect of a lower error tolerance for the sub-blocks containing terms related to the SIF, allowing a higher error when approximating the other blocks. Table 11.2 analyses this issue considering two errors for the Frobenius norm, ε_C and ε_{SIF}. One can verify that the error in SIFs is strongly governed by ε_C, and less so by ε_{SIF} for the SIFs reduce slightly even for high accuracy, nevertheless the number of operations increases.

11.5 Conclusions

In this work we applied ACA with the XBEM for solving an anisotropic fracture problem. ACA has been used to accelerate the solution times of the problem, using only 10% of the time required to solve the system using regular Gauss elimination. Future work includes the use of partial pivoting, in order to save memory, and extending the formulation for 3D problems.

Acknowledgements The first author acknowledges the Faculty of Science, Durham University, for his Postdoctoral Research Associate funding.

References

[AT15] Alatawi, I.A., Trevelyan, J.: A direct evaluation of stress intensity factors using the extended dual boundary element method. Eng. Anal. Bound. Elem. **52**, 56–63 (2015)

[BA10] Benedetti, I., Aliabadi, M.H.: A fast hierarchical dual boundary element method for three-dimensional elastodynamic crack problems. Int. J. Numer. Methods Eng. **84**(9), 1038–1067 (2010)

[BAD08] Benedetti, I., Aliabadi, M.H., Davi, G.: A fast 3D dual boundary element method based on hierarchical matrices. Int. J. Solids Struct. **45**(7), 2355–2376 (2008)

[Beb08] Bebendorf, M.: Hierarchical Matrices. Springer, New York (2008)

[BMA09] Benedetti, I., Milazzo, A., Aliabadi, M.H.: A fast dual boundary element method for 3D anisotropic crack problems. Int. J. Numer. Methods Eng., **80**(10), 1356–1378 (2009).

[BR03] Bebendorf, M., Rjasanow, S.: Adaptive low-rank approximation of collocation matrices. Computing **70**(1), 1–24 (2003)

[GG10] Grytsenko, T., Galybin, A.N.: Numerical analysis of multi-crack large-scale plane problems with adaptive cross approximation and hierarchical matrices. Eng. Anal. Bound. Elem. **34**(5), 501–510 (2010)

[Hac99] Hackbusch, W.: A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: Introduction to $\mathcal{H}$-matrices. Computing **62**(2), 89–108 (1999)

[HAT16] Hattori, G., Alatawi, I.A., Trevelyan, J.: An extended boundary element method formulation for the direct calculation of the stress intensity factors in fully anisotropic materials. Int. J. Numer. Methods Eng. **109**(7), 965–981 (2017)

[HRDS+12] Hattori, G., Rojas-Díaz, R., Sáez, A., Sukumar, N., García-Sánchez, F.: New anisotropic crack-tip enrichment functions for the extended finite element method. Comput. Mech. **50**(5), 591–601 (2012)

[Liu09] Liu, Y.J.: Fast Multipole Boundary Element Method: Theory and Applications in Engineering. Cambridge University Press, Cambridge (2009)

[MBA12] Milazzo, A., Benedetti, I., Aliabadi, M.H.: Hierarchical fast BEM for anisotropic time-harmonic 3-D elastodynamics. Comput. Struct. **96**, 9–24 (2012)

[MS10] Messner M., Schanz, M.: An accelerated symmetric time-domain boundary element formulation for elasticity. Eng. Anal. Bound. Elem. **34**(11), 944–955 (2010)

[NYK99] Nishimura, N., Yoshida, K.-I., Kobayashi, S.: A fast multipole boundary integral equation method for crack problems in 3D. Eng. Anal. Bound. Elem. **23**(1), 97–105 (1999)

[Rok85] Rokhlin, V.: Rapid solution of integral equations of classical potential theory. J. Comput. Phys. **60**(2), 187–207 (1985)

[RS07] Rjasanow, S., Steinbach, O.: The Fast Solution of Boundary Integral Equations. Springer Science & Business Media, New York (2007)

[SS86] Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)

[ST11] Simpson, R., Trevelyan, J.: A partition of unity enriched dual boundary element method for accurate computations in fracture mechanics. Comput. Meth. Appl. Mech. Eng., **200**(1), 1–10 (2011)

[YNK01] Yoshida, K.-I., Nishimura, N., Kobayashi, S.: Application of fast multipole Galerkin boundary integral equation method to elastostatic crack problems in 3D. Int. J. Numer. Methods Eng. **50**(3), 525–547 (2001)

Chapter 12
Flux Characterization in Heterogeneous Transport Problems by the Boundary Integral Method

R.D. Hazlett

12.1 Introduction

Consider solutions to the following partial differential equation in Cartesian coordinates subject to potentially different initial and boundary conditions. Here, P is pressure, k_x, k_y, and k_z are the directional transport coefficients, ϕ is porosity, μ is viscosity, C_t is compressibility, and q is flow rate.

$$k_x \frac{\partial^2 P}{\partial x^2} + k_y \frac{\partial^2 P}{\partial y^2} + k_z \frac{\partial^2 P}{\partial z^2} = \phi \mu C_t \frac{\partial P}{\partial t} - q\mu \cdot \delta(X - x_o)\delta(y - y_o)\delta(z - z_o) \qquad (12.1)$$

In seeking a Green's function solution for Neumann boundary conditions, N_o, we note that for a domain of volume V for the Poisson equation with a point source at $\vec{r}_o$, the following must be satisfied:

$$\nabla \cdot (\frac{k}{\mu} \nabla N_o) = C_f \cdot q_o(\delta_o(\vec{r}, \vec{r}_o) - \frac{1}{V})$$

where C_f is a constant and

$$(\frac{k_\eta}{\mu}) \cdot \frac{\partial N_o}{\partial \eta} = 0$$

on the boundary S. Note there is a mathematical singularity with the source term in a defined location (x_o, y_o, z_o) that could be integrated in space to yield a line source. The singularities in Equation (12.1) are in space, not time, making the proper

R.D. Hazlett (✉)
McDougall School of Petroleum Engineering, The University of Tulsa, Tulsa, OK, USA
e-mail: randy-hazlett@utulsa.edu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_12

handling of the spatial singularities the most important aspect of getting a proper transient response for arbitrary observation point. The solution can be considered as the product of three one-dimensional solutions [Ne36].

$$P_D(x,y,z;x_1,y_1,z_1,\alpha,\beta,\gamma,L;t) = \frac{t}{\phi\mu C_t} + \frac{1}{L}\cdot\int_0^L \sum_{l,m,n\neq 0}^{\infty} \frac{C_{lmn}}{\pi^2}\cdot\frac{1-e^{\frac{-\pi^2 D_{lmn}^2 t}{\phi\mu C_t}}}{D_{lmn}^2}\cdot$$
$$cos(\frac{\pi lx}{a})cos(\frac{\pi my}{b})cos(\frac{\pi nz}{h})cos(\frac{\pi lx_o}{a})cos(\frac{\pi my_o}{b})cos(\frac{\pi nz_o}{h})ds \qquad (12.2)$$

where

$$D_{lmn}^2 \equiv \frac{k_x l^2}{a^2} + \frac{k_y m^2}{b^2} + \frac{k_z n^2}{h^2}$$

Equation (12.2) is computationally challenging in this raw form but can be transformed using a number of identities [Br08, GR80] into a computationally efficient form containing only analytic constructs and rapidly converging, highly accurate, infinite series summation approximations [MHB01]. Babu and Odeh produced a semi-analytical solution for the special case of a horizontal well for transient [OB90] and semi-steady-state production [BO89]. Hazlett and Babu [HB14] gave a highly accurate and computationally efficient solution to Equation (12.1) with analytic integration in time and space for an arbitrarily oriented line source term with Neumann external boundary conditions. The solution is for the dimensionless pressure difference between any observation point and the volume averaged pressure per unit of fluid withdrawn, termed drawdown to indicate how hard the well must work to produce a barrel of fluid.

Other researchers purport that numerical spatial integration of the point source was adequate to model a well of any trajectory [Du00, EBF96, EDB91, WDA03]. These efforts used an integrand that contained a singularity and was numerically time-consuming to evaluate. The numerical method is equivalent to representing a line source by a dense number of point sources. Far field behavior may be adequately captured in this approach, but if we want to perform evaluations at a distance of one well radius from the source, there are issues related to nearby, unmitigated singularities. Attempts to numerically soften the singularity were less than satisfactory [Ma96].

12.2 Boundary Integral Method for Coupled Analytic Solutions

If the medium properties become a function of space, Equation (12.1) applies only locally. A heterogeneous transport property domain could be represented by

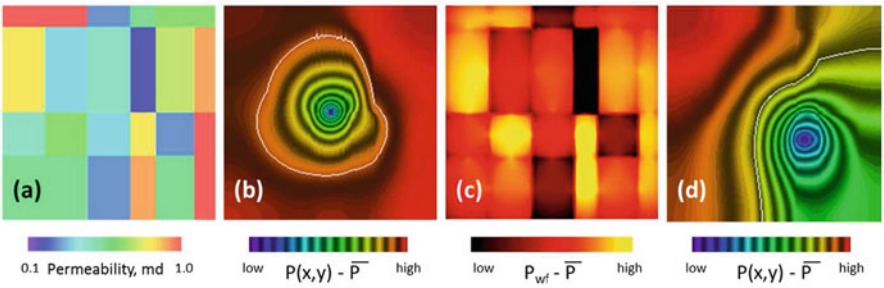


Fig. 12.1 Heterogeneous problems solved using coupled analytic solutions and a boundary integral method: (a) a patch-wise distribution of transport coefficients, (b) pressure distribution for a centrally located well, (c) pressure drawdown values for a dense set of well locations, and (d) pressure distribution for an optimally located well

a system of interacting semi-analytical solutions [HB05]. Via Green's Theorem, a closed system solution can be augmented with a boundary integral to allow material transport across the interface as represented in Equation (12.3) for the time independent portion of the solution to Equation (12.1). Such boundary element and boundary integral methods are well established [Co00, HB05, KH93, LLK98, MH88].

$$P(\vec{r},\vec{r}_o) = \overline{P} - \frac{q}{q_o}\cdot N_o(\vec{r},\vec{r}_o) - \frac{1}{q_o}\cdot\int_S N_o(\vec{r},\vec{r}_o)g(\sigma)d\sigma \tag{12.3}$$

Here, $g(\sigma)$ is the outward normal boundary flux, $\overline{P}$ is the average pressure, and q_o is a reference flow rate.

Without knowledge of the structure of the integrand, numerical methods yield an equation matrix with normal flux at predefined boundary elements and the average pressure in each domain as unknowns. Figure 12.1 illustrates results possible as solutions to coupled analytic solutions with patch-wise continuous transport properties as shown in Figure 12.1a. Figure 12.1b gives the pressure distribution for a centrally placed well with the location of the average value highlighted in white. Figure 12.1c shows the value of well drawdown for every possible location of the well, indicating a global minimum as the brightest color. Figure 12.1d shows the spatial distribution of pressure if we placed the well in this optimum location.

12.3 Numerical Boundary Integral Evaluation

Figure 12.2 shows error evaluation in a simple, 4-cell homogeneous problem containing a point source that was solved as a multi-region problem using Gaussian quadrature [St71] with variable node density. Equation (12.4) indicates the integration approximation as a weighted sum of a set of location specific integrand function

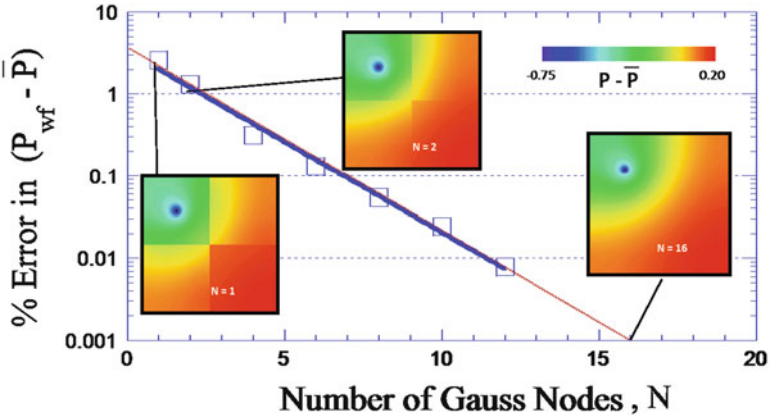


Fig. 12.2 Error in drawdown computation for a homogeneous transport problem on a square with a point source centered in quadrant two as a function boundary nodes density. While the error in pressure computation in the neighborhood of the singularity is tolerable with few nodes, the inset figures of the pressure field show considerable error elsewhere

evaluations. While error magnitude in drawdown computation for the well is small, the inset figures with pressure computed on a dense grid of observation points show that the solution with a small number of Gauss nodes is globally unsatisfactory. Figure 12.3a and 12.3b indicates the character of the Neumann function solution, its normal derivative, and the lumped boundary integrand for a point and line source, respectively, as a function of the relative distance between the source and an interface. Integral splitting the boundary integral at the cusp location into two separate integrals is warranted.

$$\int_{x_1}^{x_2} f(x)dx = \sum_{k=1}^{n} w_k \cdot f(x_k) \tag{12.4}$$

A problem similar to that using numerical integration to approximate line sources was encountered when introducing points sources of unknown flux magnitude as boundary elements. This is illustrated in Figure 12.4 where the pressure is evaluated on the boundary between two regions. The anomalous pressure behavior near a boundary node due to the introduction of a point source is quite obvious.

Two-dimensional Gaussian quadrature application yielded non-diagonal dominance in matrices, indicating an improper weighting scheme. Analytical integration eliminated this issue. For full boundary integration of either top or bottom interfaces ($z_o = 0$ or h) and an observation point away from the boundary, we get Equation (12.5). If the observation point is also moved to the same boundary ($z = z_o$), as often the case in the construction of pressure matching conditions at the boundary nodes, then the integral collapses to $h^2/(3k_z)$.

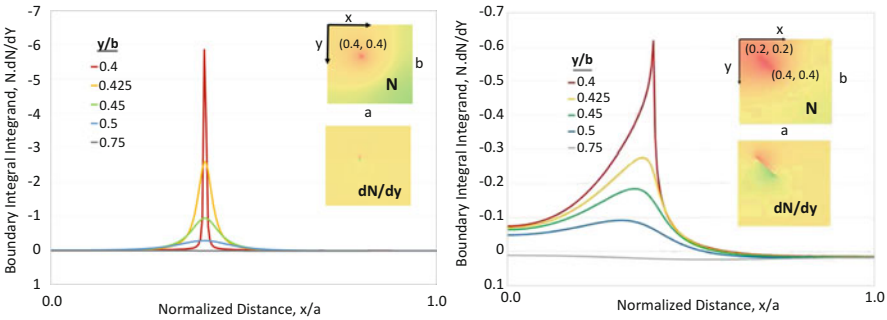


Fig. 12.3 The nature of the boundary integrand as a function of location of the fictitious boundary in a homogeneous problem and a depiction of the Neumann function and its normal derivative: (*left*) point source and (*right*) line source. Both suggest subdividing the integral

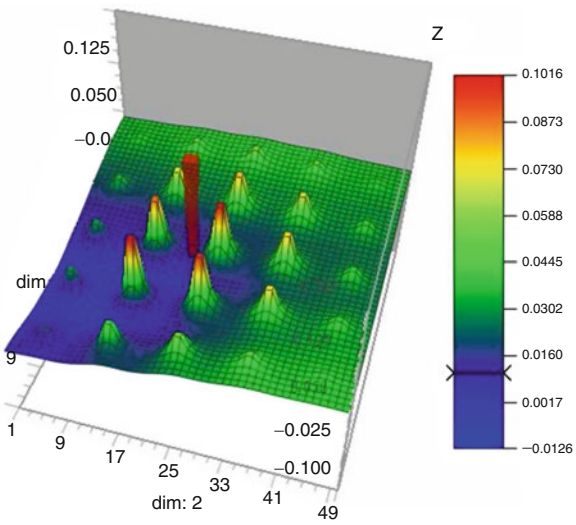


Fig. 12.4 Anomalous behavior of the pressure at a boundary where the flux is represented by point sources of determined magnitude at sparse fixed boundary node locations. The location of the marker coincides with the numerical value of pressure signified on the colorscale

$$\frac{1}{ab}\int_0^b \int_0^a N_o(x, y, z; x_o, y_o, z_o)dx_o\, dy_o$$

$$= \frac{h^2}{k_z}\cdot\left[\frac{1}{3} - \frac{\max(z, z_o)}{h} + \left(\frac{z^2 + z_o^2}{2h^2}\right)\right] \qquad (12.5)$$

A number of other full analytic integrals were listed by Hazlett and Babu [HB09] for special case triangles. Unfortunately, flux patterns were not seen to conform to uniform flux except for case where the interface is remote from a source term. This has strong implications for strictly numerical routines that introduce only one value of flux per interface. With a single value, it can only represent the average and must be interpreted as representative of the entire interface.

12.4 Piecewise Continuous Solutions

In an attempt to further exploit the benefits of analytic boundary integration, the integral was decomposed into piecewise uniform flux patches over which the average flux is locally representative. Partial boundary integration with local average normal flux (see Appendix) eliminated numerical integration artifacts, alleviated dependency upon node spacing and weights, and effectively removed spatial singularities association with BIM application. Still, concerns over computational speed and large matrices cast doubt on practical application to large systems [Do11]. However, each cell in a strictly numerical solution to Equation (12.1) contains no information below the cell size; whereas in the method described, each cell has a complete analytic solution and access to as much detail as required. Still, the patch density required to access an analytic solution everywhere without knowledge of the integrand structure severely limits the application.

12.5 Parametric Methods

Hazlett and Babu [HB13] proposed a parametrized functional form for the flux consisting of a linear combination of separate uniform flux and uniform pressure boundary problems. This hybrid boundary condition is illustrated in Figure 12.5 and is seen to be exact for equal-cell-size problems [Zh15]. The uniform flux solution was already given as Equation (12.5). The Dirichlet boundary condition solution produces a zero value of pressure on the boundary. Thus, boundary integration for this contribution can be avoided entirely in favor of evaluation of the source term influence on the pressure on the boundary, since the sum of this and the boundary integral must be everywhere zero on the boundary. Thus, the parametric form for boundary flux posed by Hazlett and Babu [HB13] consists of readily evaluated parts.

12.6 Prolongation

Parametric representation of boundary flux in transport problems was investigated more thoroughly [Zh15, ZH16]. For cells of different size, shown also to correspond to heterogeneous transport property systems, these authors showed through prolongation that the unknown flux is indeed composed of uniform flux and uniform pressure contributions, but an additional term is required that corresponds to circulation, as illustrated in Figure 12.6. If we look at the origin of the circulation term, we find that we can relate the unknown flux to that of yet another prolonged problem by expanding the solution domain to create another equal-cell-size problem. This process can be repeated indefinitely (Figure 12.7).

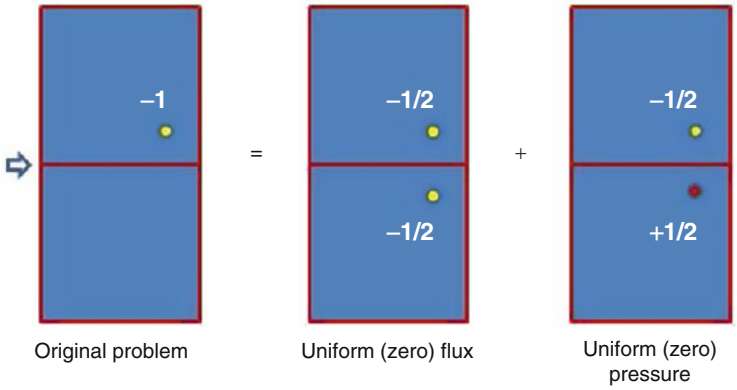


Fig. 12.5 Exact replacement of an equal-cell-size problem of unknown solution with a sum of easily evaluated steady-state uniform (zero) flux and uniform pressure boundary problems

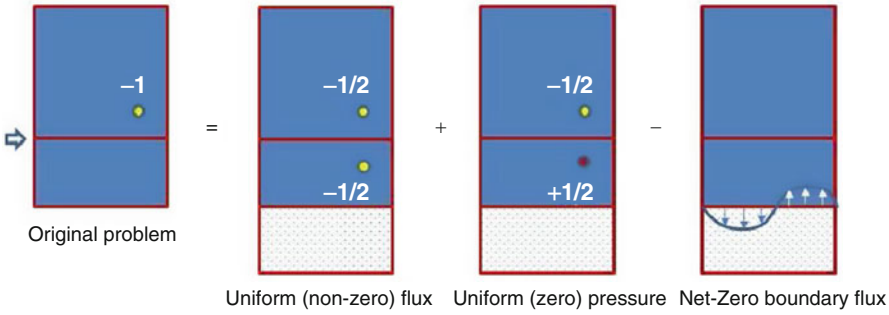


Fig. 12.6 Prolongation as a route to a more easily solved problem as envisioned by [Zh15]

12.7 Conclusions

In contrast to traditional BIM numerical methods, coupled analytic solutions for heterogeneous domains can alternatively be posed and solved numerically as boundary integral problems with analytically integrated patch-wise uniform flux. An ad-hoc parametric representation of the boundary flux as a combination of uniform flux and uniform pressure easily evaluated constituents can greatly reduce problem bandwidth. Extending the parametric method, heterogeneous problems (those with either unequal cell size or permeability contrast) can be linked to a homogeneous, equal-cell-size problem via prolongation. The correction term to be supplied for the original problem is linked to a set of cascading prolongation problems to elucidate the boundary flux at ever-increasing distance from the original interface of interest. The solution is then linked to a set of uniform flux and uniform pressure problems whose boundary integrals are easily evaluated analytically. The

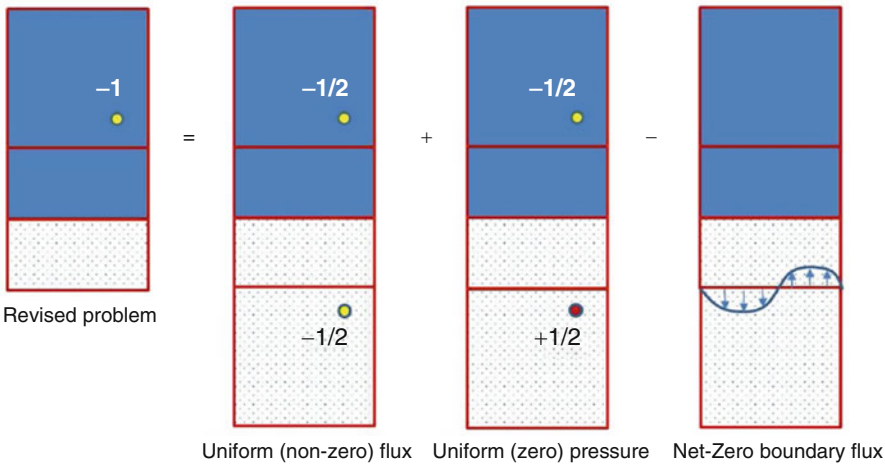


Fig. 12.7 Origin of the rotational contribution and the redirection of interest to a more remote interface. The process is repeated indefinitely with successive prolongations at ever-increasing distance from the original interface of interest

repeated prolongation may be truncated after just a few prolongation cycles, since the pressure at the original problem interface is not sensitive to flux distribution at significant distance.

Acknowledgements I gratefully acknowledge the immense contributions of my longtime collaborator in this field of investigation, Dr. D. Krishna Babu.

Appendix: Partial Integration of the Neumann Function on a Boundary

In patch-wise uniform flux boundary elements, assuming $\frac{a^2}{k_x} = \max(\frac{a^2}{k_x}, \frac{b^2}{k_y}, \frac{h^2}{k_z})$, we examine partial integration of the point source Neumann function on the boundary as $t \to \infty$. Focusing on the triple infinite series term requiring computational advantage, we obtain the following:

$$\int_{z_1}^{z_2}\int_{x_1}^{x_2}\sum_{l,m,n=1}^{\infty}\frac{C_{lmn}}{\pi^2}\cdot\frac{1}{D_{lmn}^2}$$

$$\cdot cos(\frac{\pi lx}{a})cos(\frac{\pi my}{b})cos(\frac{\pi nz}{h})cos(\frac{\pi lx_o}{a})cos(\frac{\pi my_o}{b})cos(\frac{\pi mz_o}{b})dx_o dz_o =$$

$$\frac{ah \cdot H(x-x_1)H(x_2-x)}{4\pi} \cdot \sum_{j=1}^{4} \sum_{m,n=1}^{\infty} \frac{sin(\frac{\pi n Z_j}{h})cos(\frac{\pi m y}{b})cos(\frac{\pi m y_o}{b})}{nD_{mn}^2}$$

$$- \frac{ah}{8\pi} \cdot sign(X_i) \cdot \sum_{m,n=1}^{\infty} \frac{sin(\frac{\pi n Z_j}{h})cos(\frac{\pi m y}{b})cos(\frac{\pi m y_o}{b})}{nD_{mn}^2} \cdot \frac{sinh[\frac{\pi a}{\sqrt{k_x}} D_{mn}(1 - \frac{|X_i|}{a})}{sinh[\frac{\pi a}{\sqrt{k_x}} D_{mn})}$$

where

$$D_{mn}^2 \equiv \frac{k_y m^2}{b^2} + \frac{k_z n^2}{h^2}$$

$$X_i \equiv [(x_2 + x), (x_2 - x), -(x + x_1), (x - x_1)]$$

$$Z_j \equiv [(z_2 + z), (z_2 - z), -(z + z_1), (z - z_1)]$$

The first term on the RHS can be reduced to a single series with exponential damping, whereas the hyperbolic functions are replaced with exponentials and reformulated in terms of a rapidly convergent double infinite series with exponential damping.

If $\frac{h^2}{k_z} \geq \frac{b^2}{k_y}$, then

$$\sum_{j=1}^{4} \sum_{m,n=1}^{\infty} \frac{sin(\frac{\pi n Z_j}{h})cos(\frac{\pi m y}{b})cos(\frac{\pi m y_o}{b})}{nD_{mn}^2} =$$

$$\frac{1}{2} \cdot \sum_{m=1}^{\infty} \frac{h^2}{k_z} cos(\frac{\pi m(y \pm y_o)}{b}) \cdot [\sum_{n=1}^{\infty} \frac{\sum_{k=1}^{4} sin(\frac{\pi n Z_k}{h})}{n(n^2 + \frac{k_y}{k_z}\frac{h^2}{b^2}m^2)}]$$

The portion in parentheses can be further reduced. Otherwise,

$$\sum_{j=1}^{4} \sum_{m,n=1}^{\infty} \frac{sin(\frac{\pi n Z_j}{h})cos(\frac{\pi m y}{b})cos(\frac{\pi m y_o}{b})}{nD_{mn}^2} =$$

$$\frac{1}{2} \cdot \sum_{n=1}^{\infty} \frac{b^2}{nk_y} \sum_{k=1}^{4} sin(\frac{\pi n Z_k}{h}) \cdot \sum_{m=1}^{\infty} \frac{cos(\frac{\pi m(y \pm y_o)}{b})}{(m^2 + \frac{k_z}{k_y}\frac{b}{h}n^2)}$$

Concerning the second term in the integration,

$$sign(X_i) \cdot \sum_{m,n=1}^{\infty} \frac{sin(\frac{\pi n Z_j}{h})cos(\frac{\pi m y}{b})cos(\frac{\pi m y_o}{b})}{nD_{mn}^2} \cdot \frac{sinh[\frac{\pi a}{\sqrt{k_x}} D_{mn}(1 - \frac{|X_i|}{a})}{sinh[\frac{\pi a}{\sqrt{k_x}} D_{mn})} =$$

$$4 \cdot \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{cos(\frac{\pi m y}{b}) cos(\frac{\pi m y_o}{b}) \sum_{k=1}^{4} sin(\frac{\pi n Z_k}{h})}{n(m^2 \frac{k_y}{b^2} + \frac{k_z}{h^2} n^2)} \cdot \frac{\sum_{k=1}^{4} sign(X_k) \cdot (e^{-\pi D_{mn} |\frac{X_k}{a}|} - e^{-\pi D_{mn}(2-|\frac{X_k}{a}|)})}{1 - e^{-2\pi D_{mn}}}$$

This last term can be efficiently coded with termination of infinite summations to desired accuracy.

References

[BO89] Babu, D., Odeh, A.: Productivity of a horizontal well. SPE Reserv. Eng. **4**(4), 417–421 (1989)

[Br08] Bromwich, T.J.I'A.: An Introduction to the Theory of Infinite Series. Macmillan and Co., London (1908)

[Co00] Constanda, C.: Direct and Indirect Boundary Integral Equation Methods. Chapman & Hall/CRC, Boca Raton (2000)

[Do11] Dogru, A.H.: Giga-cell simulation. Saudi Aramco J. Technol. **Spring**, 2–7 (2011)

[Du00] Durlofsky, L.J.: Advanced techniques for reservoir simulation and modeling of non-conventional wells. Final Report, DOE Award **DE-AC26-99BC15213**. Department of Petroleum Engineering, Stanford University (2004)

[EDB91] Economides, M.J., Deimbacher, F.X., Brand, C.W. et al.: Comprehensive simulation of horizontal-well performance. SPE Form. Eval. **6**(4), 418–426 (1991)

[EBF96] Economides, M.J., Brand, C.W., Frick, T.P.: Well configurations in anisotropic reservoirs. SPE Form. Eval. **11**(4), 257–262 (1996)

[GR80] Gradshteyn, I.S.,Ryzhik, I.M.: Table of Series Integrals and Products. Academic, New York (1980)

[HB05] Hazlett, R.D., Babu, D.K.: Optimal well placement in heterogeneous reservoirs through semianalytic modeling. SPE J. **10**(03), 286–296 (2005)

[HB09] Hazlett, R., Babu, D.: Readily computable Green's and Neumann functions for symmetry-preserving triangles. Q. Appl. Math. **67**(3), 579–592 (2009)

[HB13] Hazlett, R., Babu, D.: Influence of cell boundary flux distribution on well pressure. Proc. Comput. Sci. **18**, 2137–2146 (2013)

[HB14] Hazlett, R.D., Babu, D.K.: Discrete wellbore and fracture productivity modeling for unconventional wells and unconventional reservoirs. SPE J. **19**(01), 19–33 (2014)

[KH93] Kikani, J., Horne, R.N.: Modeling pressure-transient behavior of sectionally homogeneous reservoirs by the boundary-element method. SPE Form. Eval. **8**(2), 145–152 (1993)

[LLK98] Lough, M.F., Lee, S.H., Kamath, J.: An efficient boundary integral formulation for flow through fractured porous media. J. Comput. Phys. **143**(2), 462–483 (1998)

[Ma96] Maizeret, P.-D.: Well indices for non conventional wells. MS Thesis, Stanford University, Stanford (1996)

[MH88] Masukawa, J., Horne, R.N.: Application of the boundary integral method to immiscible displacement problems. SPE Reserv. Eng. **3**(03), 1069–1077 (1988)

[MHB01] McCann, R.C., Hazlett, R.D., Babu, D.K.: Highly accurate approximations of Green's and Neumann functions on rectangular domains. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. **457**(2008), 767–772 (2001)

[Ne36] Newman, A.B.: Heating and cooling rectangular and cylindrical solids. Ind. Eng. Chem. **28**(5), 545–548 (1936)

[OB90] Odeh, A.S., Babu, D.K.: Transient flow behavior of horizontal wells pressure drawdown and buildup analysis. SPE Form. Eval. **5**(01), 7–15 (1990)

[St71] Stroud, A.H.: Approximate Calculation of Multiple Integrals. Prentice-Hall, Englewood Cliffs (1971)

[WDA03] Wolfsteiner, C., Durlofsky, L.J., Aziz, K.: Calculation of well index for nonconventional wells on arbitrary grids. Comput. Geosci. **7**, 61–82 (2003)

[Zh15] Zhang, Y.: Parametric representation of boundary flux in heterogeneous potential flow problems. MS Thesis, University of Tulsa, Tulsa (2015)

[ZH16] Zhang, Y., Hazlett, R.D.: Parametric representation of cell boundary flux distributions in well equations. J. Comput. Appl. Math. **307**, 65–71 (2016)

Chapter 13
GPU Based Mixed Precision PWR Depletion Calculation

A. Heimlich, A.C.A. Alvim, F.C. Silva, and A.S. Martinez

13.1 Introduction

A pressurized light water nuclear reactor (PWR) refueling typically replaces about a third of the pent fuel every twelve to eighteen months, depending on fuel burnup. The nuclide concentrations within each nuclear fuel element evaluated by burnup calculations indicate whether a fuel element must be replaced or reallocated. The combinatorial optimization problem of refueling, and as such, the number of evaluated candidate solutions are a function of time and of computational resources.

Computational tools to model the reactor core are used to evaluate its reactivity, spatial neutronic behavior, power distribution, isotopic inventory, and fuel burnup. The fuel burnup can be represented by a system of first-order, ordinary, coupled differential equations (ODE), accounting for all fissionable actinides and fission fragment yields for the radioactive reaction chains under analysis.

A previous study [He16] has shown that the computational power of Graphic Processor Unit (GPU) can substantially improve the speed performance of nuclear fuel burnup calculations.

The main objective of the present study is to evaluate mixed precision with adaptive time step solver based on Adams-Moulton-Bashford to calculate fuel burnup in PWR reactors in massive multicore GPU using parallel programming techniques.

A. Heimlich (✉)
Nuclear Engineering Institute, Rio de Janeiro, RJ, Brazil
e-mail: adino.heimlich@ien.gov.br

A.C.A. Alvim • F.C. Silva • A.S. Martinez
Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
e-mail: alvim@lmp.ufrj.br; fernando@nuclear.ufrj.br; martinez@nuclear.ufrj.br

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_13

13.2 Theory

The balance which relates the production and consumption of each nuclide for each radioactive chain reaction into the core implies a huge system of first-order differential equations.

The nuclide concentrations in the fuel are dependent of decay constant, neutron flux and microscopic fission, radioactive capture, and scattering cross sections for each nuclide chain. This system is described by Bateman's equation [Ba10]. This work evaluates 17 actinides and 20 fission yields in 2 neutron energy groups.

Equation (13.1) represents the abundance variation for nuclide k.

$$\frac{dN_k^x}{dt}(t) = \overbrace{\sum_{i=1,i\neq k}^{\text{Actinide-series}}}\ N_i^x(t)\sum_{g=1}^{G}\left(\sigma_{\gamma,i}^{g,x}(t) - \sigma_{f,i}^{g,x}(t) + \sigma_{(n,2n),i}^{g,x}(t) + \right)\phi_x^g(t)$$
$$- \overbrace{\sum_{y=1}^{\text{Decay Fractions}}}\ \lambda_{i,y}N_i^x(t) + \overbrace{\sum_{z\neq i}^{\text{Production Fractions}}}\ \lambda_z N_z^x(t) \tag{13.1}$$

The abundance variation of nuclide l induced by fission of actinides is described by Equation (13.2).

$$\frac{dN_l^x}{dt}(t) = \overbrace{\sum_{i=1,i\neq l}^{\text{Fission Yields}}}\left(N_i^x(t)\sum_{g=1}^{G}\left(\Gamma_{i,l}^g\sigma_{f,i}^{g,x}(t) - \sigma_{\gamma,i}^{g,x}(t)\right)\phi_x^g(t)\right)$$
$$- \lambda_l N_l^x(t) \tag{13.2}$$

Variables x, t, and g represent the spatial position, time, and energy group, respectively. $N_k^x(t)$ represents actinide k concentration. Microscopic cross sections are represented by $\sigma_{f,i}^{g,x}(t)$, $\sigma_{\gamma,i}^{g,x}(t)$, and $\sigma_{(n,2n),i}^{g,x}(t)$. Decay constants are $\lambda_{i,y}$ with branch y. The neutron flux is $\phi_x^g(t)$ and finally $\Gamma_{i,l}^g$ represents nuclide l yield from fission reaction of actinide i.

Burnup calculi are based on the assumption that neutron distribution flux and microscopic cross sections are static throughout the reactor core in each time step. Thus, in the beginning of each step, the solution of neutron diffusion equation to compute the neutron flux distribution based on the solution of the previous time step is needed. Neutron flux distribution is calculated using a multigroup NEM solver [Fi77] and the reaction rates and thermohydraulics are computed, including Xenon and Boron effect feedback to reconstruct the new microscopic cross sections constants used in the next burnup step. Furthermore, a predictor-corrector scheme is applied to improve the solution using present and previous concentrations.

Bateman's balance equations can be written in differential equations system formulation by Equations (13.1) and (13.2) and can be represented in matrix form by Equation (13.3).

$$\frac{d\vec{N}(t)}{dt} = \mathbf{A}\cdot\vec{N}(t), \quad \vec{N}(0) = \vec{N}_0, \tag{13.3}$$

where vector $\vec{N}(0)$ represents the initial concentration and $\mathbf{A}$ is the depletion matrix.

$$\vec{N}(t) = e^{\mathbf{A}\,\Delta t} \cdot \vec{N}(0) \tag{13.4}$$

Thus, nuclide concentrations can be evaluated using the recursive procedure in Equation (13.5) where variable $\Delta t = t_n - t_{n-1}$ is the burnup step.

$$\vec{N}(t_n) = e^{\mathbf{A}\,\Delta t} \cdot \vec{N}(t_{n-1}) \tag{13.5}$$

GPU hardware architecture achieves high performances based on high occupancy of many thousand processors. This study uses a very large vector and a matrix depletion operator to increase the calculi efficiency. Linear algebra operator *direct sum* $\oplus$ is used to create a big, sparse matrix with more than one hundred thousand lines. This approach creates a large system of equations that can be represented in matrix form.

Equation (13.3) shows the ODE system written in matrix form. Operator $\mathbf{A}$ is rebuilt in each step of the burnup according to neutron flux, isotopic concentrations, and microscopic neutron cross sections of actinides and fission fragments. The operator $\mathbf{A}$ is built in GPU to maximize performance. This matrix, initially in COO format, consists of 2998 nodes evaluating 37 nuclides each. Furthermore, this matricial operator is converted to ELLPACK format to improve the efficiency of sparse matrix multiplication. Figure 13.1 shows the matrix associated with the depletion operator for one node and eight nodes.

Considering that the simulation of 2998 fuel nodes with 37 nuclides implies in a large sparse matrix operator, with over one hundred thousand lines and at least five million non-zero elements, a very large number of operations are implied.

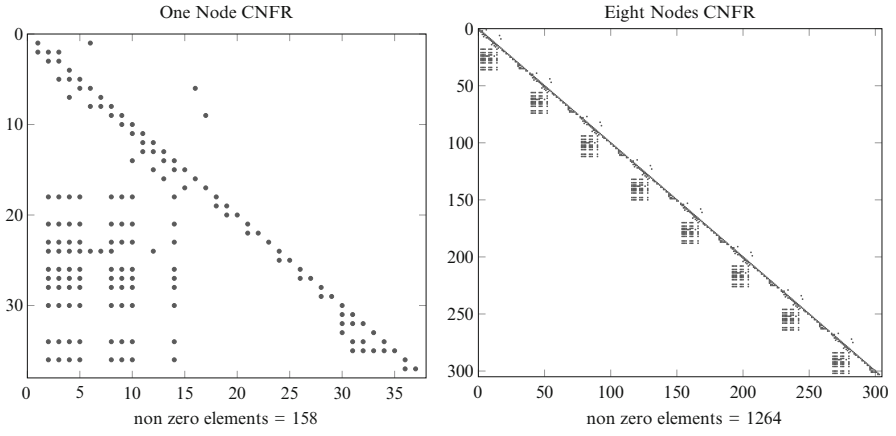


Fig. 13.1 One node and Eight nodes matrices

13.3 Exponential Matrix

The function $exp(A\,\Delta t)$ represents an exponential of transition matrix and can be solved by matrix exponential methods like Taylor or Padé expansion series. This is a well-known problem in differential equation theory and its solution can be given by matrix exponential methods in at least nineteen ways [Mo03]. Equation (13.6) shows the expansion of $e^{\mathbf{A}\,\Delta t}$ in Taylor series.

$$e^{\mathbf{A}\,\Delta t} = \left(I + (\mathbf{A}\,\Delta t) + \frac{(\mathbf{A}\,\Delta t)^2}{2!} + \frac{(\mathbf{A}\,\Delta t)^3}{3!} + \cdots + \frac{(\mathbf{A}\,\Delta t)^n}{n!} \right) \tag{13.6}$$

Multiplying the concentration vector $\vec{N}_0$ in both sides of Equation (13.6), Equation (13.7) is obtained.

$$e^{\mathbf{A}\,\Delta t} \cdot \vec{N}_0 = \left(I + (\mathbf{A}\,\Delta t) + \frac{(\mathbf{A}\,\Delta t)^2}{2!} + \frac{(\mathbf{A}\,\Delta t)^3}{3!} + \cdots + \frac{(\mathbf{A}\,\Delta t)^n}{n!} \right) \cdot \vec{N}_0 \tag{13.7}$$

The calculations are improved if we use the matrix-vector decomposition $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n\}$ shown in Equation (13.8), where $B = A\,\Delta t$.

$$e^{\mathbf{A}\,\Delta t} \cdot \vec{N}_0 = \left(\vec{N}_0 + \underbrace{\mathbf{B} \cdot \vec{N}_0}_{\vec{v}_1} + \underbrace{\frac{\mathbf{B}}{2}(\underbrace{\mathbf{B} \cdot \vec{N}_0}_{\vec{v}_1})}_{\vec{v}_2} + \underbrace{\frac{\mathbf{B}}{3}(\frac{\mathbf{B}}{2}(\underbrace{\mathbf{B} \cdot \vec{N}_0}_{\vec{v}_1}))}_{\vec{v}_3} + \cdots \right) \tag{13.8}$$

However, Taylor series expansion calculi requires extended precision provided by multi-precision arithmetic and small time step size to achieve good results. The maximum step size is restricted by fast reactions in chains under analysis and can be estimated using the norm of matrix $\mathbf{A}\,\Delta t$ calculated in Equation (13.9).

$$|\mathbf{A}\,\Delta t| = min \left\{ \frac{max}{j} \sum_i |a_{i,j}|, \frac{max}{i} \sum_j |a_{i,j}| \right\} \tag{13.9}$$

The norm of matrix $\mathbf{A}\,\Delta t$ is a limiting factor for the maximum time step used to calculate the exponential matrix $e^{\mathbf{A}\,\Delta t}$. The nuclide that has shortest half-life in the burnup chain used in this study is Xe^{135}, whose half-life is approximately 9.08 h. Thus, the balance of maximum and minimum time steps used by the numerical solutions, the precision of 64 bits floating-point representation, the desired accuracy, and speed must be considered to choose these bounds.

13.4 Runge-Kutta Methods

Runge-Kutta methods are used to find an approximation of a solution to an initial value problem of the form:

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0 \tag{13.10}$$

and to obtain an approximation of $y(x)$ using a truncated Taylor series. In this study the implementation of fourth-order Runge-Kutta method (RK4) is used to achieve four points needed by the Adams-Moulton-Bashford (AMB) algorithm and this procedure can be shown in Figure 13.2. This implementation of RK4 can use single or double precision arithmetic to produce the vector of the predictor stage of AMB.

13.4.1 Runge-Kutta-Fehlberg

The implementation of the Runge-Kutta-Fehlberg (RKF) [Sh77] method in parallel GPUs [He16] applied to calculate nuclear fuel burnup in PWR reactors shows accuracy and great speed performance. It uses an adaptive time-step correction based on local error measure. The Euclidean distance is calculated by the fifth-order and fourth-order approximations difference. Equation (13.11) shows the vectors $\vec{k}_i$ calculi and six *Sparse Matrix Vector Multiplication* (SpMv) operations.

$$\vec{k}_1 = h \cdot \mathbf{f}(t_n, \vec{y}_n)$$

$$\vec{k}_2 = h \cdot \mathbf{f}(t_n + \frac{1}{4}h, \vec{y}_n + \frac{1}{4}\vec{k}_1)$$

$$\vec{k}_3 = h \cdot \mathbf{f}(t_n + \frac{1}{8}h, \vec{y}_n + \frac{3}{32}\vec{k}_1 + \frac{9}{32}\vec{k}_2)$$

$$\vec{k}_4 = h \cdot \mathbf{f}(t_n + \frac{12}{13}h, \vec{y}_n + \frac{1932}{2197}\vec{k}_1 + \frac{7200}{2197}\vec{k}_2 + \frac{7296}{2197}\vec{k}_3)$$

1: $\vec{y}_n \leftarrow \vec{y}_o$ ▷ Read concentration
2: **for** $i \geq n$ **do** ▷ Order 4 Runge-Kutta
3: $\vec{k_1} = h \cdot \mathbf{f}(t, \vec{y})$
4: $\vec{k_2} = h \cdot \mathbf{f}(t, \vec{y})$
5: $\vec{k_4} = h \cdot \mathbf{f}(t, \vec{y})$
6: $\vec{k_4} = h \cdot \mathbf{f}(t, \vec{y})$
7: $\vec{y}_n{+}1 = \vec{y}_n + h \cdot \frac{1}{6} \cdot (\vec{k_1} + 2\vec{k_2} + \vec{k_3} + \vec{k_4})$
8: $t = t + h$

Fig. 13.2 RK4 Algorithm

while $t < End_t$ **do** ▷ Read concentration
  $\vec{y_n} \leftarrow \vec{y_0}$ ▷ *tolerance* must be near to $1e^{-9}$
  $\vec{k_1} = h \cdot \mathbf{f}(\vec{y_n})$,
  $\vec{k_2} = h \cdot \mathbf{f}(\vec{y_n} + \frac{1}{4}\vec{k_1})$
  $\vec{k_3} = h \cdot \mathbf{f}(\vec{y_n} + \frac{3}{32}\vec{k_1} + \frac{9}{32}\vec{k_2})$
  $\vec{k_4} = h \cdot \mathbf{f}(\vec{y_n} + \frac{1932}{2197}\vec{k_1} - \frac{7200}{2197}\vec{k_2} + \frac{7296}{2197}\vec{k_3})$
  $\vec{k_5} = h \cdot \mathbf{f}(\vec{y_n} + \frac{439}{216}\vec{k_1} - 8\vec{k_2} + \frac{3680}{513}\vec{k_3} - \frac{845}{4104}\vec{k_4})$
  $\vec{k_6} = h \cdot \mathbf{f}(\vec{y_n} - \frac{8}{27}\vec{k_1} + 2\vec{k_2} - \frac{3544}{2565}\vec{k_3} + \frac{1859}{4104}\vec{k_4} - \frac{11}{40}\vec{k_5})$
  $\vec{w} = \frac{1}{360}\vec{k_1} + \frac{-128}{4275}\vec{k_3} + \frac{-2197}{75240}\vec{k_4} + \frac{2}{55}\vec{k_5})$ ▷ $\approx \mathcal{O}^5$
  $\vec{y}_{n+1} = \vec{y_n} + \frac{16}{135}\vec{k_1} + \frac{6656}{12825}\vec{k_3} + \frac{28561}{56430}\vec{k_4} - \frac{9}{50}\vec{k_5} + \frac{2}{55}\vec{k_6})$ ▷ $\approx \mathcal{O}^6$
  $distance = \|\vec{w}\|$ ▷ Local distance evaluation
  if $distance < tolerance$ **then**
    $t = t + h$
    $\vec{y_0} \leftarrow \vec{y}_{n+1}$ ▷ Choice a new time-step
    $h = h \cdot (\frac{tolerance}{distance})^{\frac{1}{4}}$
  else
    $h = h \cdot 0.84 \cdot (\frac{tolerance}{distance})^{\frac{1}{5}}$

Fig. 13.3 RKF algorithm

$$\vec{k}_5 = h \cdot \mathbf{f}(t_n + h, \vec{y}_n + \frac{439}{216}\vec{k}_1 - 8\vec{k}_2 + \frac{3680}{513}\vec{k}_3 - \frac{845}{4104}\vec{k}_4)$$
$$\vec{k}_6 = h \cdot \mathbf{f}(t_n + \frac{1}{2}h, \vec{y}_n - \frac{8}{27}\vec{k}_1 + 2\vec{k}_2 - \frac{3544}{2565}\vec{k}_3 - \frac{1859}{4104}\vec{k}_4 - \frac{11}{40}\vec{k}_5) \quad (13.11)$$

The previous study result [He16] obtained by RKF and Jacobi Colocation method is used as benchmark of speed performance and accuracy, respectively (Figure 13.3).

The fourth-order approximation is given by

$$\vec{y}_{n+1} = \vec{y}_n + \frac{25}{216}\vec{k}_1 + \frac{1408}{2565}\vec{k}_3 + \frac{2197}{4104}\vec{k}_4 - \frac{1}{5}\vec{k}_5.$$

The fifth-order approximation is given by

$$\vec{y}_{n+1} = \vec{y}_n + \frac{16}{135}\vec{k}_1 + \frac{6656}{12825}\vec{k}_3 + \frac{28561}{56430}\vec{k}_4 - \frac{9}{50}\vec{k}_5 + \frac{2}{55}\vec{k}_6.$$

Absolute local error is given by

$$distance = |\vec{w}| = |\frac{1}{360}\vec{k}_1 + \frac{-128}{4275}\vec{k}_3 + \frac{-2197}{75240}\vec{k}_4 + \frac{2}{55}\vec{k}_5)|.$$

The optimized step h is calculated in Equation (13.4.1).

$$h = \begin{cases} h \cdot \lambda \left(\dfrac{tolerance}{d} \right)^{\frac{1}{5}} & \text{if } d > tolerance \\ h \cdot \left(\dfrac{tolerance}{d} \right)^{\frac{1}{4}} & \text{if } d \leq tolerance \end{cases}$$

The algorithm increases or reduces the time step h depending on how much the local error *distance* is lower than *tolerance*. In this case was used $\lambda = 0.84$.

13.5 Adams-Moulton-Bashford Method

The Adams-Moulton-Bashford [Ja91] method (AMB) is a linear multi-step, predictor-corrector algorithm, which uses information from the previous step to calculate the next value and approximates the solution of initial value problem $y' = \mathbf{f}(y, t)$. The function $\mathbf{f}(y, t) :\rightarrow y$ must be evaluated into points $y_0, y_1, \cdots, y_{1-k}$ to obtain an approximation of order k shown by Equation (13.12).

$$\vec{y} \approx \vec{y}_{i+1} = \vec{y}_i + \int_{t_i}^{t_{i+1}} \mathbf{f}(t, y(t))\, dt \tag{13.12}$$

Equation (13.13) shown the formulation of fourth-order AMB which requires four points to do the interpolation procedure. These points are obtained using a fourth-order Runge-Kutta method shown in Figure 13.2 using matrix valued function. This work proposes an implementation of adaptive time step using local error evaluation similar one used by RKF to evaluate fuel burnup of nuclear reactor. The mixed precision approach is performed using single precision evaluations of $\mathbf{f}$ to improve speed performance in predictor stage, i.e., the RK4 stage, and double precision evaluations in corrector stage.

$$\begin{aligned} y_{old} &= \mathbf{f}(y_3) \\ y_{n+1} &= y_n + h \cdot \left(\frac{55}{24}\mathbf{f}(y_{old}) - \frac{59}{24}\mathbf{f}(y_2) + \frac{37}{24}\mathbf{f}(y_1) - \frac{9}{24}\mathbf{f}(y_0) \right) \\ y_{n+1} &= y_n + h \cdot \left(\frac{9}{24}\mathbf{f}(y_{n+1}) + \frac{19}{24}\mathbf{f}(y_{old}) - \frac{5}{24}\mathbf{f}(y_2) + \frac{9}{24}\mathbf{f}(y_1) \right) \end{aligned} \tag{13.13}$$

Figure 13.4 shows the algorithm of fourth-order adaptive AMB. New and old solutions used by corrector stage, $\vec{y}_{n+1}$ and $\vec{y_{old}}$, are calculated using fourth and third-order approximation, respectively.

$\vec{y} \leftarrow \vec{x_0}$ ▷ Read concentration
$i \leftarrow 0$
while $count < Turns$ **do** ▷ Adjust Turns to best fit
  for $i \geq 4$ **do** ▷ Runge-Kutta loop, 4 stages
    $s[i] \leftarrow RK4(\vec{y})$
    $t = t + h$
while $t < t_{end}$ **do** ▷ Adams-Moulton-Bashford loop
  $i \leftarrow 0$
  for $i \geq 4$ **do**
    $\vec{y_{old}} \leftarrow \vec{y}$ ▷ Save old $\vec{y}$
    $s[3] \leftarrow \mathbf{f}(t, \vec{y})$
    $\vec{y} = \vec{y} + \frac{55h}{24} \cdot s[3] - \frac{59h}{24} \cdot s[2] + \frac{37h}{24} \cdot s[1] - \frac{9h}{24} \cdot s[0]$ ▷ Explicit
    $\vec{y} = \vec{y_{old}} + \frac{9h}{24}\mathbf{f}(t, \vec{y}) + \frac{19h}{24} \cdot s[3] - \frac{5h}{24} \cdot s[2] + \frac{h}{24} \cdot s[1]$ ▷ Implicit
    for $j \geq 3$ **do** ▷ Shift stages
      $s[j] \leftarrow s[j+1]$
  $d = \|\vec{y} - \vec{y_{old}}\|$ ▷ Local distance evaluation
  if $(d > tolerance)$ AND $(h > hmin)$ **then** ▷ Choice a new time-step
    $h = 0.87h \cdot (\frac{tolerance}{d})^{\frac{1}{5}}$ ▷ If d is big $\rightarrow \downarrow h$
  else
    $h = h \cdot (\frac{tolerance}{d})^{\frac{1}{4}}$ ▷ If d is low $\rightarrow \uparrow h$
    $t = t + 5 \cdot h$ ▷ Walking in time
    $\vec{y} \leftarrow \vec{y_{old}}$

Fig. 13.4 AMB algorithm

The new solution is evaluated implicitly by third-order approximation

$$\vec{y} = \vec{y_{old}} + \frac{9h}{24}\mathbf{f}(t, \vec{y}) + \frac{19h}{24} \cdot \vec{s}[3] - \frac{5h}{24} \cdot \vec{s}[2] + \frac{h}{24} \cdot \vec{s}[1]$$

and absolute local error is given by Equation (13.5)

$$d = |\vec{y} - \left(\frac{55h}{24} \cdot \vec{s}[3] - \frac{59h}{24} \cdot \vec{s}[2] + \frac{37h}{24} \cdot \vec{s}[1] - \frac{9h}{24} \cdot \vec{s}[0]\right)|.$$

AMB mixed method requires a copy of predictor vectors to single float precision and execution of RK4 method, subsequently, another copy of these vectors to double float must be done. The local error correspond to difference between two approximated solutions of different orders, given by Equation (13.5) results whether imply in acceptance or refuse and rebuild the procedure using smaller step h. However, minimum and maximum time step (h_{min}, h_{max}) are chosen by experimental results depending of Δt and the fastest nuclide decay constant.

13.6 Results

The reactor core evaluated in this work is similar to the Areva ANP lightweight pressurized water reactor. The simulation divides the core in 4369 nodes, of which 2988 are fuel nodes, disposed in 18 horizontal layers and consists of one burnup cycle with fresh fuel and 32 burnup steps with a 498 days duration. A representative fuel node is chosen to measure the error between each GPU implemented method and result based on third-order Jacobi collocation method.

This study uses a GeForce GTX580 to build tests and the accumulated time to calculus of 458 days of burnup using RKF is 1.69 seconds and AMB is 1.8 seconds in double precision and 1.9 seconds in mixed mode.

Root Mean Squared Error (RMSE) method was used to measure the Euclidean distance between the sequences of nuclides inventory calculated by the methods under analysis and the benchmark and this formula is shown in Equation (13.14), where the X set contains the benchmark and Y the sequence under analysis.

$$\mathbf{RMSE} = 100 \times \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}}{(Max\{y_i\} - Min\{y_i\})}, \quad x_i \in X_k \text{ and } y_i \in Y_k^m. \tag{13.14}$$

Figure 13.5 shows the RMSE error of Adams-Moulton-Bashford relative to benchmark $n = 32$ burnup steps covering one-cycle burnup in the fuel cycle of ANGRA II PWR reactor.

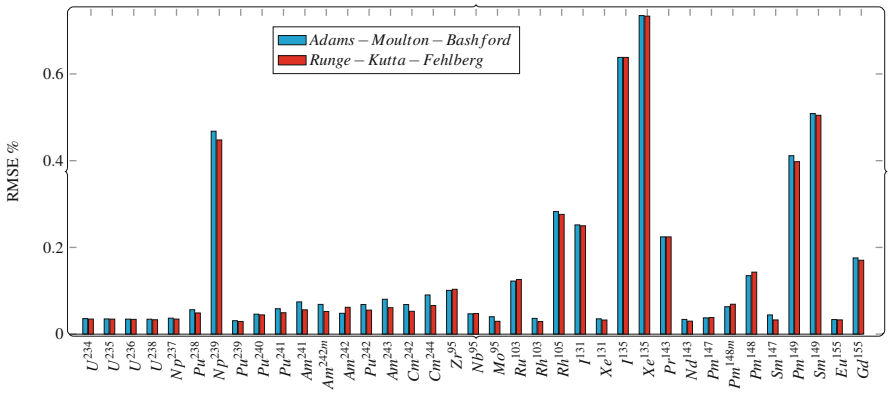


Fig. 13.5 Total RMSE (%) vs Nuclide

13.7 Conclusions and Further Developments

This work presents the implementation of mixed precision, adaptive step Adams-Moulton-Bashford method to evaluate fuel burnup and isotopic inventory using GPU parallel programming. The computational performance is compared against a burnup and spent fuel inventory solver based on Runge-Kutta-Fehlberg and accuracy of both with the benchmark produced by Jacobi colocation method. The comparative performance shows close results within and accuracy according to needs to solve one-cycle burnup in the fuel cycle of ANGRA II PWR reactor.

The mixed precision approach seemed full of promise because GPU hardware is three times more powerful in single precision than double, but we are now seeing that approach presents many problems. For example, a full copy and conversion from double to float precision of burnup operator $\mathbf{A}\,\Delta t$. Furthermore, four copies and conversion to double precision of predictor condition ($s[i]\,i = 0 \cdots 3$) are a bottle-neck to use mixed precision and low quality single floating-point arithmetic produces a bad choice of predictor and therefore the corrector stage needs more time to converge. Further work consists in implementation of quad precision or arbitrary-precision floating-point arithmetic in GPU using a variable order AMB.

Acknowledgements The authors gratefully acknowledge the support of *Instituto de Engenharia Nuclear*, *Instituto Nacional de Ciência e Tecnologia de Reatores Nucleares Inovadores*, and the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPQ).

References

[Ba10] Bateman, H.: The solution of a system of differential equations occurring in the theory of radioactive transformations. Proc. Camb. Philos. Soc **15**, 423–427 (1910)

[Fi77] Finnemann, H., Bennewitz, F., Wagner, M.R.: Interface current techniques for multidimensional reactor calculations. Atomkernenergie (ATKE) **30**, 123–128 (1977)

[He16] Heimlich, A., Silva, F.C., Martinez, A.S.: Parallel GPU implementation of PWR reactor burnup. Ann. Nucl. Energ. **91**, 135–141 (2016)

[Ja91] Jackiewicz, Z., Lo, E.: The numerical solution of neutral functional differential equations by Adams predictor-corrector methods. Appl. Numer. Math. **8**(6), 477–491 (1991)

[Mo03] Moler, C., Van Loan, C.: Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. **45**(1), 3–49 (2003)

[Sh77] Shampine, L.: Stiffness and nonstiff differential equation solvers, ii: detecting stiffness with Runge-Kutta methods. ACM Trans. Math. Softw. (TOMS) **3**(1), 44–53 (1977)

Chapter 14
2D Gauss-Hermite Quadrature Method for Jump-Diffusion PIDE Option Pricing Models

L. Jódar, M. Fakharany, and R. Company

14.1 Introduction

An option is a contract whose holder has the right to sell or buy an asset traded in the market in a future time called the maturity at a prefixed price called the strike. The classic Black–Scholes model for valuing option contracts does not depict stock price changes, market risks, heavy tails and asymmetries observed in market data [Du93]. Typical alternatives to the Black–Scholes model are based on non-lognormal assumptions for the stochastic differential equation which represents the changes of the underlying asset and/or the volatility, see, for instance, [He93, BaEtAl98]. These models use to price the option contract as the solution of a second-order parabolic partial differential equation (PDE). Alternatives to capture the reality of the market are the jump-diffusion models which incorporate jumps in the variation of the underlying asset. This fact introduces a non-local integral term in the equation of the option price related to the assumed type of the underlying asset jumps and intensities. Thus the option price is given by the solution of a partial integro-differential equation (PIDE) [CoEtAl04], having as many independent variables as underlying assets apart from the time. An excellent overview for the existing numerical methods of one asset jump-diffusion models may be found in [ClEtAl08].

L. Jódar (✉) • R. Company
Institute of Multidisciplinary Mathematics, Polytechnic University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain
e-mail: ljodar@imm.upv.es; rcompany@imm.upv.es

M. Fakharany
Faculty of Science, Mathematics Department, Tanta University, Tanta, Egypt
e-mail: fakharany@aucegypt.edu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_14

In this paper we consider finite difference methods for the two-asset Merton jump-diffusion model for put options on the minimum of two assets described by the risk-neutral dynamics of two-asset jump-diffusion model

$$\frac{dS_i(t)}{S_i(t)} = (r - q_i - \lambda\kappa_i)dt + \sigma_i dW_i + (e^{J_i} - 1)dZ(t), \quad i = 1, 2,$$

where, for $i = 1, 2$, S_i denote the underlying assets, J_i are the jump sizes, the expectation $E(e^{J_i} - 1)$ is denoted by κ_i, q_i are the dividend yields, σ_i denote the volatilities, W_i are correlated standard Brownian motions with $\rho \in (-1, 1)$ and r is the risk free interest. Z and λ are the Poisson process and its jump intensity, respectively [RaEtAl13]. Using Itô calculus and the transformation $x_i = \ln(S_i/E)$, $i = 1, 2$, $\tau = T - t$, where E is the strike price and T represents the maturity, the option price $U(x_1, x_2, \tau)$ is given by the solution of the PIDE

$$\frac{\partial U}{\partial \tau} = \frac{\sigma_1^2}{2}\frac{\partial^2 U}{\partial x_1^2} + \rho\sigma_1\sigma_2\frac{\partial^2 U}{\partial x_1 \partial x_2} + \frac{\sigma_2^2}{2}\frac{\partial^2 U}{\partial x_2^2} + \left(r - q_1 - \lambda\kappa_1 - \frac{\sigma_1^2}{2}\right)\frac{\partial U}{\partial x_1}$$

$$+ \left(r - q_2 - \lambda\kappa_2 - \frac{\sigma_2^2}{2}\right)\frac{\partial U}{\partial x_2} - (r + \lambda)U + \lambda \int_{\mathbb{R}^2} U(x_1 + \eta_1, x_2 + \eta_2) g(\eta_1, \eta_2) d\eta_1 d\eta_2, \tag{14.1}$$

where

$$g(\eta_1, \eta_2) = \frac{\exp\left[-\frac{1}{2(1-\rho_J^2)}\left(\left(\frac{\eta_1 - \mu_1}{\hat{\sigma}_1}\right)^2 - \frac{2\rho_J(\eta_1 - \mu_1)(\eta_2 - \mu_2)}{\hat{\sigma}_1\hat{\sigma}_2} + \left(\frac{\eta_2 - \mu_2}{\hat{\sigma}_2}\right)^2\right)\right]}{2\pi\hat{\sigma}_1\hat{\sigma}_2\sqrt{1 - \rho_J^2}}, \tag{14.2}$$

is the probability density function of a bivariate normal distribution, μ_i, $\hat{\sigma}_i$ are the means and standard deviations of the jump sizes J_i, $i = 1, 2$, respectively. Here $\rho_J \in (-1, 1)$ is the correlation parameter between the jump sizes [ClEtAl08, RaEtAl13]. Apart from the PIDE (14.1), the solution must satisfy the initial condition [ClEtAl08, Du06] corresponding to the payoff of the American put over the minimum of two-asset

$$f(x_1, x_2) = E\max(1 - \min(e^{x_1}, e^{x_2}), 0). \tag{14.3}$$

Suitable boundary conditions are included suggested by the 1D put option problem and the payoff as follows:

$$\begin{aligned} &\lim_{x_1 \to -\infty} U(x_1, x_2, \tau) = Ee^{-r\tau}, \quad \lim_{x_2 \to -\infty} U(x_1, x_2, \tau) = Ee^{-r\tau}, \\ &U(x_1, x_2, \tau) \approx f(x_1, x_2), \text{ as } x_1 \to \infty \text{ or } x_2 \to \infty. \end{aligned} \tag{14.4}$$

A recent numerical antecedent for two-asset jump-diffusion models using Galerkin finite element approach may be found in [RaEtAl13]. Finite difference

methods for this model are treated in [ClEtAl08], where the numerical treatment of the integral term is based on the use of the trapezoidal rule on a truncated domain and further Fast Fourier Transform.

In this paper we use a finite difference approach to solve the PIDE problem (14.1)–(14.4). However, as it has been pointed out in [ZvEtAl03], the presence of the mixed derivative term in the differential part of (14.1) may generate numerical instabilities and inaccuracy, apart from a high number of nodes in the scheme stencil what grows the computational cost. This paper is organized as follows. In Section 14.2, the elimination of the mixed derivative term of the PIDE (14.1) is developed by using the canonical transformation of a second-order PDE [Ga98]. Discretization of the transformed problem is treated in Section 14.3. Standard finite difference approximation of the differential part is considered, while the integral part is approximated by using Gauss-Hermite quadrature and further bivariate interpolation. Both discretizations of the differential and integral part are matched in an appropriate way as the zeroes of the Hermite polynomials not need to be nodes of the numerical scheme. Positivity, stability and consistency of the numerical scheme are also treated. For the sake of brevity and taking into account the limited extension of the chapter, we omit the proofs of some results. A numerical example is included in Section 14.4. Summarizing, our contribution is based on the elimination of the cross-derivative term to avoid known drawbacks reducing the computational cost and on the use of Gauss-Hermite quadrature and a bivariate interpolation for the approximation of the integral part allowing accurate solutions with a few number of terms. The material presented in this paper is work in progress, an extended version of which, inclusive of all the necessary proofs and tests of the results, will be published elsewhere in due course.

14.2 Mixed Derivative Elimination

In this section a transformation of the problem (14.1)–(14.4) is proposed in order to remove the cross-derivative term in (14.1). Let us consider the change of independent variables,

$$y_1 = \frac{\sigma_2\tilde{\rho}}{\sigma_1}x_1, \quad y_2 = x_2 - \frac{\sigma_2\rho}{\sigma_1}x_1, \quad \tilde{\rho} = \sqrt{1-\rho^2}, \tag{14.5}$$

together with the transformation of the unknown variable

$$V(y_1, y_2, \tau) = \frac{\exp((r+\lambda)\tau)}{E}U(x_1, x_2, \tau). \tag{14.6}$$

From (14.5) and (14.6), Equation (14.1) becomes

$$\begin{aligned}\frac{\partial V}{\partial \tau} &= \frac{\sigma_2^2\tilde{\rho}^2}{2}\left(\frac{\partial^2 V}{\partial y_1^2} + \frac{\partial^2 V}{\partial y_2^2}\right) + a_1\frac{\partial V}{\partial y_1} + a_2\frac{\partial V}{\partial y_2} \\ &+ \frac{\sigma_1\lambda}{\sigma_2\tilde{\rho}}\int_{\mathbb{R}^2} V(\phi_1,\phi_2,\tau)g\left(\frac{\sigma_1}{\sigma_2\tilde{\rho}}(\phi_1 - y_1), \phi_2 - y_2 + \tilde{m}(\phi_1 - y_1)\right)d\phi_1 d\phi_2,\end{aligned} \tag{14.7}$$

where the coefficients a_i and the new variables for the integral part ϕ_i are given by

$$a_1 = \frac{\tilde{\rho}\sigma_2}{\sigma_1}(r - q_1 - \lambda k_1 - \frac{\sigma_1^2}{2}),$$
$$a_2 = (1 - \frac{\rho\sigma_2}{\sigma_1})r - (q_2 - \frac{\rho\sigma_2}{\sigma_1}q_1) - \lambda\kappa_2 + \frac{\rho\sigma_2}{\sigma_1}\lambda\kappa_1 - \frac{\sigma_2^2}{2} + \frac{\rho\sigma_1\sigma_2}{2},$$
$$\phi_1 = y_1 + \frac{\sigma_2\tilde{\rho}}{\sigma_1}\eta_1,\ \phi_2 = y_2 - \frac{\sigma_2\rho}{\sigma_1}\eta_1 + \eta_2,\ \tilde{m} = \frac{\rho}{\tilde{\rho}}.$$

Transformed boundary conditions (14.4) become

$$\lim_{y_1\to-\infty} V(y_1, y_2, \tau) = e^{\lambda\tau},\ \lim_{y_2\to-\infty} V(y_1, y_2, \tau) = e^{\lambda\tau},$$
$$V(y_1, y_2, \tau) \approx f(y_1, y_2),\ \text{as } y_1 \to \infty \text{ or } y_2 \to \infty.$$

14.3 Numerical Scheme Construction and Properties

In order to discretize the transformed problem (14.7), let us choose firstly the bounded numerical domain following criteria of [KaEtAl00, EhEtAl08]. Thus, let us take a rectangular domain in x_1x_2- plane with boundaries $x_1 \in [a, b]$ and $x_2 \in [c, d]$ such that e^b and e^d are about ten times the strike E and e^a and e^c are close enough to zero. Under the transformation (14.5), the rectangular domain is converted to a rhomboid domain Ω with vertices $ABCD$ in y_1y_2-plane as shown in Figure 14.1. Let $N_1 + 1$ and $N_2 + 1$ be the numbers of mesh points in x_1 and x_2 directions, respectively, such that the spatial stepsizes are $h_{x_1} = (b - a)/N_1$ and $h_{x_2} = (d - c)/N_2$. From (14.5) the original mesh points $(N_1 + 1)(N_2 + 1)$ are mapped into the rhomboid domain with new stepsizes $h_1 = \frac{\sigma_2}{\sigma_1}\tilde{\rho}h_{x_1}$, $h_2 = h_{x_2}$. Hence the

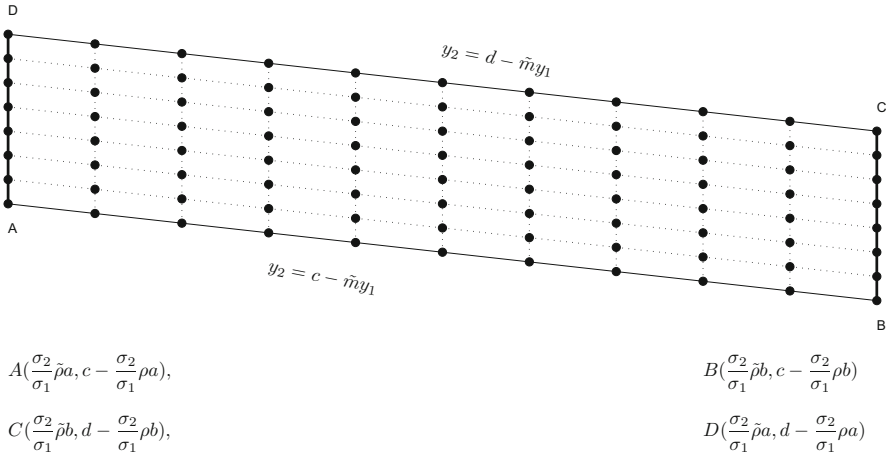


Fig. 14.1 Rhomboid numerical domain $ABCD$

new rhomboid mesh points are $(y_{1,i}, y^i_{2,j})$ where

$$y_{1,i} = y_{1,0} + ih_1, \;\; 0 \le i \le N_1, \;\; y_{1,0} = \frac{\sigma_2}{\sigma_1}\tilde{\rho}a,$$

and for each i value,

$$y^i_{2,j} = \hat{y}_{i,0} + jh_2, \;\; \hat{y}_{i,0} = c - \frac{\sigma_2}{\sigma_1}\rho(a + ih_{x_1}), \;\; 0 \le j \le N_2.$$

The time variable is discretized by $\tau^n = nk$, $0 \le n \le N_\tau$, $k = 1/N_\tau$.

As we mentioned in the introduction we are going to use Gauss-Hermite quadrature to discretize the integral part. So for the sake of convenience if we denote by $\Phi = \{(\phi_{1,\ell}, \phi_{2,m}),\ 1 \le \ell \le L, 1 \le m \le M\}$ the set of all the pairs of zeroes of Hermite polynomial of degrees L in the first argument and M in the second, respectively, we select parameters a, b, c, d identifying the rhomboid numerical domain Ω so that $\Phi \subset \Omega$ after L and M are prefixed.

Let us denote the approximation of $V(y_{1,i}, y^i_{2,j}, \tau^n)$ by $V^n_{i,j}$. Differential part of PIDE (14.7) is discretized by using central finite difference approximations for the first and second spatial derivatives and forward finite difference approximation is implemented to approximate the time partial derivative of V. Dealing with the discretization of the unbounded support bidimensional integral term of PIDE (14.7), we recall the Gauss-Hermite quadrature in 2D for a function $z(x, y)$ given by

$$\sum_{l=1}^{L}\sum_{m=1}^{M} \omega_l \omega_m z_{lm} \approx \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)} z(x, y)dxdy, \tag{14.8}$$

where $\boldsymbol{\omega}_l$, $l = 1, 2, \ldots, L$ and $\boldsymbol{\omega}_m$, $m = 1, 2, \ldots, M$ represent the corresponding weights for the roots of Hermite polynomial of degrees L and M, respectively. $z_{lm} = z(x_l, y_m)$ represents the value of the integrand function at node (x_l, y_m) for x_l, $1 \le l \le L$ and y_m, $1 \le m \le M$ being the roots of Hermite polynomials of degrees L and M [AbEtAl61].

By applying formula (14.8) to the improper double integral of (14.7), one gets

$$\hat{I}^n_{i,j} = \sum_{\ell=1}^{L}\sum_{m=1}^{M} \omega_\ell \omega_m C_{\ell m}(i, j) V^n(\phi_{1,\ell}, \phi_{2,m}), \tag{14.9}$$

where

$$C_{\ell m}(i, j) = g(\frac{\sigma_1}{\sigma_2\tilde{\rho}}(\phi_{1,\ell} - y_{1,i}), \phi_{2,m} - y^i_{2,j} + \tilde{m}(\phi_{1,\ell} - y_{1,i})) \exp[\phi^2_{1,\ell} + \phi^2_{2,m}],$$

and $V^n(\phi_{1,\ell}, \phi_{2,m})$ denotes the approximate value of V at point $(\phi_{1,\ell}, \phi_{2,m}, \tau^n)$. Note that expression (14.9) involves evaluation at points that usually are different from those of the grid $(y_{1,i}, y^i_{2,j})$, $1 \le i \le N_1$, $1 \le j \le N_2$, and it is necessary to

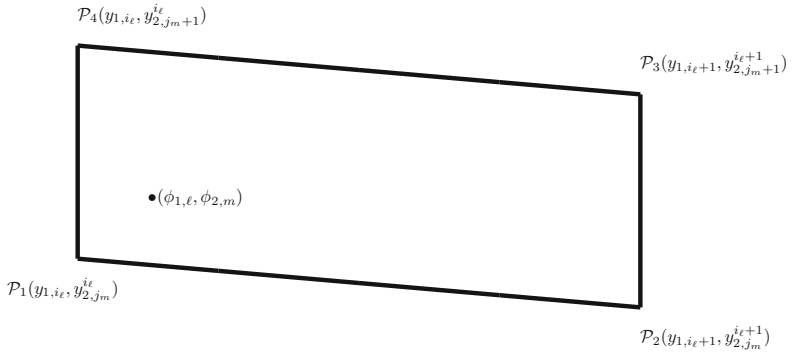


Fig. 14.2 Neighbour coordinate points to $(\phi_{1,\ell}, \phi_{2,m})$

match both discretizations of the differential and the integral part. For each value $V^n(\phi_{1,\ell}, \phi_{2,m})$ appearing in (14.9) we want to locate the pair $(\phi_{1,\ell}, \phi_{2,m})$ in one of the sub-rhomboids of the rhomboid grid. Let us denote such sub-rhomboid as $R(i_\ell, j^{i_\ell}_m)$ in such way that the point $(\phi_{1,\ell}, \phi_{2,m})$ does not lie in the right-up sides of the sub-rhomboid. Rhomboid vertices of $R(i_\ell, j^{i_\ell}_m)$ are given by $(y_{1,i_\ell}, y^{i_\ell}_{2,j_m})$, $(y_{1,i_\ell+1}, y^{i_\ell+1}_{2,j_m})$, $(y_{1,i_\ell+1}, y^{i_\ell+1}_{2,j_m+1})$ and $(y_{1,i_\ell}, y^{i_\ell}_{2,j_m+1})$. Following the idea of the bivariate interpolation four point formula [AbEtAl61, Ch. 25, p. 882], we modify such approximation for the rhomboid in Figure 14.2 as follows:

$$V^n(\phi_{1,\ell}, \phi_{2,m}) \approx \hat{\delta}_{i_\ell,2}(\delta_{j_m,2} V^n_{i_\ell,j_m} + \delta_{j_m,1} V^n_{i_\ell,j_m+1}) + \hat{\delta}_{i_\ell,1}(\delta_{j_m,3} V^n_{i_\ell+1,j_m+1} + \delta_{j_m,4} V^n_{i_\ell+1,j_m}), \quad (14.10)$$

where

$$\begin{aligned} &\hat{\delta}_{i_\ell,1} = \tfrac{\phi_{1,\ell} - y_{1,i_\ell}}{h_1};\ \hat{\delta}_{i_\ell,2} = \tfrac{y_{1,i_\ell+1} - \phi_{1,\ell}}{h_1};\ \delta_{j_m,1} = \tfrac{\phi_{2,m} - y^{i_\ell}_{2,j_m}}{h_2};\\ &\delta_{j_m,2} = \tfrac{y^{i_\ell}_{2,j_m+1} - \phi_{2,m}}{h_2};\ \delta_{j_m,3} = \tfrac{\phi_{2,m} - y^{i_\ell+1}_{2,j_m}}{h_2};\ \delta_{j_m,4} = \tfrac{y^{i_\ell+1}_{2,j_m+1} - \phi_{2,m}}{h_2}. \end{aligned} \quad (14.11)$$

Consequently, the approximation of the integral part is obtained by substituting (14.10) into (14.9) ($I^n_{i,j} \approx \hat{I}^n_{i,j}$). Hence

$$I^n_{i,j} = \sum_{\ell=1}^{L} \sum_{m=1}^{M} \beta^{(i,j)}_{i_\ell,j_m} V^n_{i_\ell,j_m} + \hat{\beta}^{(i,j)}_{i_\ell,j_m+1} V^n_{i_\ell,j_m+1} + \tilde{\beta}^{(i,j)}_{i_\ell+1,j_m} V^n_{i_\ell+1,j_m} + \breve{\beta}^{(i,j)}_{i_\ell+1,j_m+1} V^n_{i_\ell+1,j_m+1},$$

where

$$\begin{aligned} \beta^{(i,j)}_{i_\ell,j_m} &= \omega_\ell \omega_m C_{\ell m}(i,j) \hat{\delta}_{i_\ell,2} \delta_{j_m,2},\ \hat{\beta}^{(i,j)}_{i_\ell,j_m+1} = \omega_\ell \omega_m C_{\ell m}(i,j) \hat{\delta}_{i_\ell,2} \delta_{j_m,1},\\ \tilde{\beta}^{(i,j)}_{i_\ell+1,j_m} &= \omega_\ell \omega_m C_{\ell m}(i,j) \hat{\delta}_{i_\ell,1} \delta_{j_m,4},\ \breve{\beta}^{(i,j)}_{i_\ell+1,j_m+1} = \omega_\ell \omega_m C_{\ell m}(i,j) \hat{\delta}_{i_\ell,1} \delta_{j_m,3}. \end{aligned} \quad (14.12)$$

Note that given Φ we can always choose values of h_1 and h_2 such that coefficients in (14.11) are nonnegative and thus the resulting 2D interpolation formula is also nonnegative. With previous notation, the corresponding finite difference scheme becomes

$$V_{i,j}^{n+1} = \alpha_1 V_{i-1,j}^n + \alpha_2 V_{i,j-1}^n + \alpha_3 V_{i,j}^n + \alpha_4 V_{i,j+1}^n + \alpha_5 V_{i+1,j}^n + \frac{k\lambda\sigma_1}{\sigma_2\tilde{\rho}} I_{i,j}^n, \tag{14.13}$$

$$V_{i,0}^n = e^{\lambda\tau^n},\ V_{0,j}^n = e^{\lambda\tau^n},\ V_{N_1,j}^n = f_{N_1,j},\ V_{i,N_2}^n = f_{i,N_2},\ 1 \le i \le N_1,\ 1 \le j \le N_2, \tag{14.14}$$

where

$$\begin{aligned} &\alpha_1 = \tfrac{k}{2h_1}(\tfrac{\sigma_2^2\tilde{\rho}^2}{h_1} - a_1);\ \ \alpha_2 = \tfrac{k}{2h_2}(\tfrac{\sigma_2^2\tilde{\rho}^2}{h_2} - a_2);\\ &\alpha_3 = 1 - k\sigma_2^2\tilde{\rho}^2(\tfrac{1}{h_1^2} + \tfrac{1}{h_2^2});\ \ \alpha_4 = \tfrac{k}{2h_2}(\tfrac{\sigma_2^2\tilde{\rho}^2}{h_2} + a_2);\\ &\alpha_5 = \tfrac{k}{2h_1}(\tfrac{\sigma_2^2\tilde{\rho}^2}{h_1} + a_1). \end{aligned} \tag{14.15}$$

Reliable numerical solutions need to be nonnegative because they represent the price of the option. This goal is achieved by guaranteeing that coefficients (14.13) given by (14.15) become nonnegative. It can be shown that under stepsizes conditions

$$k < \frac{h_1^2 h_2^2}{\sigma_2^2\tilde{\rho}^2(h_1^2 + h_2^2)},\ \ h_1 < \frac{\sigma_2^2\tilde{\rho}^2}{|a_1|},\ \ h_2 < \frac{\sigma_2^2\tilde{\rho}^2}{|a_2|}, \tag{14.16}$$

the coefficients of (14.13) are nonnegative. Hence, starting from nonnegative initial and boundary conditions (14.14), the numerical solution $\{V_{i,j}^n\}$ is nonnegative.

As there are different concepts of stability in the literature, we state the concept of stability used previously in [FaEtAl14]. Firstly, let $\mathbf{V}^n$ denote the vector solution of the finite difference scheme (14.13)–(14.15) written in the following form:

$$\mathbf{V}^n = [\mathcal{V}_0^n\, \mathcal{V}_1^n \ldots \mathcal{V}_{N_1}^n]^T,\ \mathcal{V}_i^n = [V_{i,0}^n\, V_{i,1}^n \ldots V_{i,N_2}^n].$$

Definition 1 *Consider a numerical solution $\{V_{i,j}^n\}$ of the PIDE computed from the scheme (14.13)–(14.15) with stepsizes $h_1 = \Delta y_1$, $h_2 = \Delta y_2$ in a rhomboid computational domain and $k = \Delta\tau$. On the one hand, it is said that $\{V_{i,j}^n\}$ is strongly uniformly $\|.\|_\infty$ stable, if the vector solution $\mathbf{V}^n$ satisfies*

$$\|\mathbf{V}^n\|_\infty \le \Upsilon \|\mathbf{V}^0\|_\infty,\ \ 0 \le n \le N_\tau,$$

where $\Upsilon > 0$ does not depend on the stepsizes h_1, h_2 and k.

On the other hand, it is said to be conditionally stable when Υ is obtained under a certain condition on the stepsizes.

Following the numerical analysis techniques developed in [FaEtAl14, FaEtAl16], one can show that the scheme (14.13)–(14.15) is conditionally strongly uniformly

$\|.\|_\infty$ stable under conditions (14.16). Furthermore, the proposed scheme (14.13) is consistent with PIDE (14.7) in the sense that the exact theoretical solution approximates well the difference scheme with local truncation error tending to zero as the discretization stepsizes h and k tend to zero [Sm85], and the Hermite polynomials degrees L and M tend to infinity.

One might think that the effect of the interpolation could contaminate the approximation error. However as the next Example 1 shows, this is not the case because the effect of the interpolation error is irrelevant. Apart from the fact that the final approximation error in space of the numerical solution is of the same order 2 as the one of the interpolation error.

14.4 Numerical Example

Next example illustrates the results of the proposal numerical scheme for European put options on the minimum of two assets. The numerical example has been executed using Matlab on a Microprocessor 2.8 GHz Intel Core i5. It has been used the following definition of the root mean square relative error (**RMSRE**) of a distribution of N observations $U(x_i),\ i = 1,\dots N$, whose expected values are $\overline{U}(x_i)$, respectively,

$$\text{RMSRE} = \sqrt{\sum_{i=1}^{N}\left(\frac{1}{N}\left|\frac{U(x_i)-\overline{U}(x_i)}{\overline{U}(x_i)}\right|^2\right)}$$

Example 1 Here we compare the value of European put options obtained using scheme (14.13)–(14.15) with the corresponding values in [ClEtAl08]. Consider an European put option with parameters $T = 1$, $E = 100$, $r = 0.05$, $q_1 = q_2 = 0$, $\sigma_1 = 0.12$, $\sigma_2 = 0.15$, $\rho = 0.3$, $\lambda = 0.6$, $\mu_1 = -0.1$, $\mu_2 = 0.1$, $\hat{\sigma}_1 = 0.17$, $\hat{\sigma}_2 = 0.13$, $\rho_J = -0.2$ and the boundaries x_1x_2-plane are $x_1, x_2 \in [-3, 3]$. The **RMSRE** for S_1, S_2 belonging to the set $\{90, 100, 110\}$ is calculated for $L = M = 3$ and 5. The reference values are taken from [ClEtAl08] and the **RMSRE** is obtained for three groups $A_{S_1} = \{(S_1, 90), (S_1, 100), (S_1, 110)\},\ S_1 \in \{90,\ 100,\ 110\}$. Table 14.1 reports the associated **RMSRE** and CPU time for several grids.

Acknowledgements This work has been partially supported by the European Union in the FP7-PEOPLE-2012-ITN program under Grant Agreement Number 304617 (FP7 Marie Curie Action, Project Multi-ITN STRIKE-Novel Methods in Computational Finance) and the Ministerio de Economía y Competitividad Spanish grant MTM2013-41765-P.

Table 14.1 The **RMSRE** for European put option on the minimum of two assets for several grids

		A_{90}		A_{100}		A_{110}	
	(N_1, N_2, N_τ)	**RMSRE**	CPU (sec)	**RMSRE**	CPU (sec)	**RMSRE**	CPU (sec)
$L = M = 3$	(64,32,50)	4.188*e*-3	0.17	3.561*e*-3	0.17	4.755*e*-3	0.17
	(128,64,100)	1.247*e*-3	2.63	1.197*e*-3	2.63	2.016*e*-3	2.63
	(256,128,200)	8.836*e*-4	10.72	7.158*e*-4	10.72	7.241*e*-4	10.72
	(256,256,300)	5.636*e*-4	41.28	3.913*e*-4	41.28	4.263*e*-4	41.28
$L = M = 5$	(64,32,50)	2.611*e*-3	0.31	3.558*e*-3	0.31	2.752*e*-3	0.31
	(128,64,100)	7.854*e*-4	2.72	8.205*e*-4	2.72	7.326*e*-4	2.72
	(256,128,200)	5.392*e*-4	11.12	4.916*e*-4	11.12	4.388*e*-4	11.12
	(256,256,300)	2.496*e*-4	42.55	2.227*e*-4	42.55	2.519*e*-4	42.55

References

[AbEtAl61] Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables. Dover Books on Mathematics, New York (1961)

[BaEtAl98] Barles, G., Soner, H.M.: Option pricing with transaction costs and a nonlinear Black-Scholes equation. Finance Stoch. **2**, 369–397 (1998)

[ClEtAl08] Clift, S.S., Forsyth, P.A.: Numerical solution of two asset jump diffusion models for option valuation. Appl. Numer. Math. **58**, 743–782 (2008)

[CoEtAl04] Cont, R., Tankov, P.: Financial Modelling with Jump Processes. Finance Mathematical Series. Chapman & Hall/CRC, Boca Raton (2004)

[Du06] Duffy, D.J.: Finite Difference Methods in Financial Engineering: A Partial Differential Approach. Wiley, Chichester (2006)

[Du93] Dupire, B.: Arbitrage pricing with stochastic volatility. Technical Report, Banque Paribas Swaps and Options Research Team Monograph (1993)

[EhEtAl08] Ehrhardt, M., Mickens, R.A.: A fast, stable and accurate numerical method for the Black–Scholes equation of American options. Int. J. Theor. Appl. Finance **11**, 471–501 (2008)

[FaEtAl14] Fakharany, M., Company, R., Jódar, L.: Positive finite difference schemes for partial integro-differential option pricing model. Appl. Math. Comput. **249**, 320–332 (2014)

[FaEtAl16] Fakharany, M., Company, R., Jódar, L.: Solving partial integro-differential option pricing problems for a wide class of infinite activity Lévy processes. J. Comput. Appl. Math. **296** , 739–752 (2016)

[Ga98] Garabedian, P.R.: Partial Differential Equations. AMS Chelsea Pubs. Co., Providence, R.I. (1998)

[He93] Heston, S.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)

[KaEtAl00] Kangro, R., Nicolaides, R.: Far field boundary conditions for Black–Scholes equations. SIAM J. Numer. Anal. **38**(4), 1357–1368 (2000)

[RaEtAl13] Rambeerich, N., Tangman, D.Y., Lollchund, M.R., Bhuruth, M.: High-order computational methods for option valuation under multifactor models. Eur. J. Oper. Res. **224**, 219–226 (2013)

[Sm85] Smith, G.D.: Numerical Solution of Partial Differential Equations: Finite Difference Methods, 3rd edn. Clarendon Press, Oxford (1985)

[ZvEtAl03] Zvan, R., Forsyth, P.A., Vetzal, K.R.: Negative coefficients in two-factor option pricing models. J. Comput. Finance **7 (1)**, 37–73 (2003)

Chapter 15
Online Traffic Prediction Using Time Series: A Case study

M. Karimpour, A. Karimpour, K. Kompany, and Ali Karimpour

15.1 Introduction

Crashes and traffic congestion are among the most challenging issues in traffic engineering. Road capacities and road accidents have great impacts on traffic congestion. An accurate prediction of traffic flow is one of the significant steps in Intelligent Transportation Systems (ITS) [LiWa13]. ITS have enabled the engineers to get access to real time data [SmEtA02]. Real time data not only can be helpful for the road users to decide their routes for traveling, but also can give the chance to the engineer to manage the routes more effectively [Us12].

Different programs have long been employed to anticipate the point traffic, such as fuzzy neural approach [OgEtA11]. Also, using linear multi-regression dynamic approach helps anticipate data online [QuAl08]. Another approach to predict the traffic is artificial neural network [KuEtA13]. Other methods that have been utilized are Bayesian networks and neural network [SuEtA06].

M. Karimpour
Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: mostafa.karimpour1@gmail.com

A. Karimpour
Department of Highway and Transportation Engineering, Iran University of Science and Technology, Tehran, Iran
e-mail: Abolfazl.karimpour@gmail.com

K. Kompany
Department of Highway and Transportation Engineering, Virginia Tech, Blacksburg, VA, USA
e-mail: Kian.kompany@gmail.com

Ali Karimpour (✉)
Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: karimpor@um.ac.ir

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_15

Time series are the data and information variables collected in the past and used to predict the data in the future. These data are collected in regular time intervals. One of the applications of time series is anticipating the stock exchange of big cities [CaEtA97]. Some other applications of time series include predicting water demand of metropolitan area using structural models, time series, and neural networks [JaKu07]. Forecasting the Forex financial markets is another application of time series [YaTa00].

Time series are also used to predict traffic intensity in various studies. In a recent study conducted in [CoHu14], time series utilized to predict the trucks short-term traffic by analyzing the recent traffic data of the trucks. Traffic prediction is one of the important factors in intelligent transportation systems, utilizing time series using previous data enable us to predict the forthcoming traffic and be to enhance the traffic conditions [RaEtA14].

This paper is organized as follows. Section 15.2 explains a methodology to examine the traffic in an arbitrary intersection. Section 15.3 provides an in-flow rate prediction method of the traffic data of the target intersection, and then illustrated the proposed scheme. The results are then presented in Section 15.4. Finally, Section 15.5 highlights the main results and draws concluding remakes.

15.2 Traffic Modeling by Mixed Logic Dynamic

Mixed logic dynamic (MLD) is a common framework that combines both continues states with logical states as the following general form [BeMo99]:

$$x(t+1) = A_t x(t) + B_{1t} u(t) + B_{2t} \delta(t) + B_{3t} z(t)$$
$$y(t) = C_t x(t) + D_{1t} u(t) + D_{2t} \delta(t) + D_{3t} z(t)$$
$$E_{2t} \delta(t) + E_{3t} z(t) \leq E_{4t} x(t) + E_{1t} u(t) + E_{5t}$$

where $x(t)$, $u(t)$, and $y(t)$ are the states, inputs, and outputs, respectively, with continues and logical elements as:

$$x = \begin{bmatrix} x_c \\ x_l \end{bmatrix}, x_c \in R^{n_c}, x_l \in \left(0,1\right)^{n1}, n = n_c + n_l$$

$$u = \begin{bmatrix} u_c \\ u_l \end{bmatrix}, u_c \in R^{m_c}, u_l \in \left(0,1\right)^{m_l}, m = m_c + m_l$$

$$y = \begin{bmatrix} y_c \\ y_l \end{bmatrix}, y_c \in R^{p_c}, y_l \in \left(0,1\right)^{p_l}, p = p_c + p_l$$

other variables are logic (binary) variables.

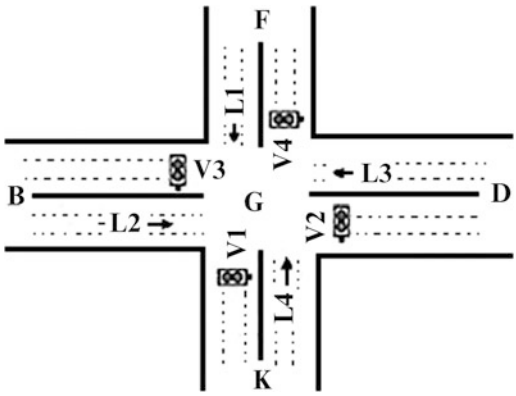


Fig. 15.1 A four connection intersection [Pa15].

Traffic modeling could be a good candidate for the MLD structure. That is, the queue length can be considered as the continuous state and the signaling of intersection can be considered as the logic values. Figure 15.1 illustrates an arbitrary intersection, the queue length for every approach of the intersection can be derived as:

$$L_i(k+1) = L_i(k) + \frac{T_s}{M_{fi}}(q_{in,i}(k) - v_i(k)q_{out,i}(k)) \quad v_i(k) \in (0,1)$$

where L_i is the queue length of ith approach, T_s is sampling period, M_{fi} is the number of lane in the ith approach, $q_{in,i}$ is the input flow rate of the ith approach, $q_{out,i}$ is the output flow rate of the ith approach, and v_i is a binary variable that shows the ith approach signal is green ($v_i = 1$) or red ($v_i = 0$). The output flow rate of the ith approach is:

$$q_{out,i}(k)v_i(k) = min\left(M_{fi}m_i v_i(k), \frac{M_{fi}L_i(k)}{T_s}v_i(k)\right)$$

where m_i is the maximum output rate of the ith approach. Since out flow rate has a nonlinear formula, it can be rewritten as [BeMo99]:

$$q_{out,i}(k)v_i(k) = z_{if}(k)\delta_i(k) + z_{3f}(k) \quad \left(\delta_i(k) = 1 \Leftrightarrow z_{if}(k) \leq 0\right)$$

where

$$z_{if} = M_{fi}m_f v_i(k) - \frac{M_{fi}L_i(k)}{T_s}v_i(k), \quad z_{3f} = \frac{M_{fi}L_i(k)}{T_s}v_i(k)$$

and δ_i is a binary variable.

In a four phase intersection at each time one light is green. The order of lights are:

$$v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \rightarrow v_1$$

To consider the order in a four phase system, the following equations must be satisfied:

$$\begin{aligned} v_1(k-1)(1-v_1(k)) - v_2(k) &\leq 0 \\ v_2(k-1)(1-v_2(k)) - v_3(k) &\leq 0 \\ v_3(k-1)(1-v_3(k)) - v_4(k) &\leq 0 \\ v_4(k-1)(1-v_4(k)) - v_1(k) &\leq 0 \end{aligned}$$

Also, to avoid interfering movements the following set of equations must be satisfied:

$$J_j v(k) \leq 1$$

where J_j is jth row of matrix J and J is:

$$J = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

By using the recent formula for output flow rate all equations can be mixed in the MLD framework. By modeling an intersection in the MLD framework, and solving the derived formulation by model predictive control [Pa15], Figure 15.2 have been achieved. Figure 15.2 demonstrates the number of cars in the approach #3 (Figure 15.1) for an arbitrary sequence (the signal order is not fixed) and regular sequence. Figure 15.3 shows the line signal of the approach #3 for an arbitrary sequence and regular sequence.

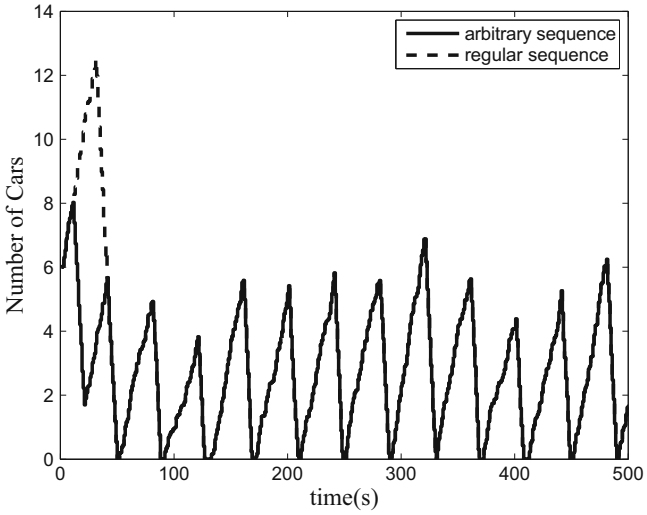


Fig. 15.2 Number of cars in approach #3 in Figure 15.1 for both arbitrary and regular sequence

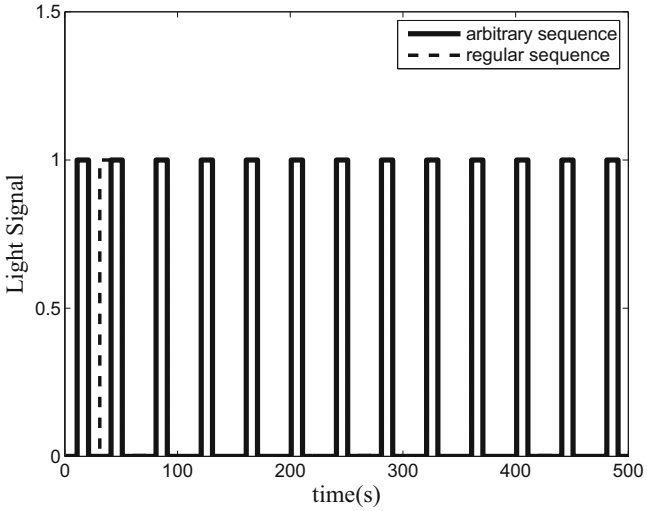


Fig. 15.3 Light signal in approach #3 in Figure 15.1 for both arbitrary and regular sequence

Mixed logic dynamic method is used to model the intersection sequences. For modeling part, the flow rate of each approach is used as the input to our model. Finally, the model is solved using model predictive control formula. Next section will provide an approach to predict input flow rate (in-flow rate) of each approach in an arbitrary intersection.

15.3 In-Flow Rate Prediction

In order to predict the in-flow traffic in an intersection, a time series model is incorporated. All the steps of building a time series for traffic prediction are explained in the following sections.

A. Time Series Models

In the physical applications of time series, when studying on a variable, there are some situations where the relevant input variable is not specified. In other words, there are some important variables which may leave effect on outputs, but one cannot manipulate them.

Statistician and economists use time series terminology to explain these systems. Some examples of time series are the world oil price, daily price of power markets, and traffic flow in an intersection in this study. For instance, traffic flow in an intersection is affected by too many parameters, i.e., ambient temperature, time of day, seasonal situation, etc. But none of these variables can be manipulated.

In time series, there is no manipulated input but there are lots of inputs that affect the output of a system. AR model is one kind of time series representation. In AR model, output of system is derived in an autoregressive (AR) manner based on the previous value of outputs:

$$y(t) + \alpha_1 y(t-1) + \alpha_2 y(t-2) + \ldots + \alpha_n y(t-n) = e(t)$$

$y(t)$ is the output in time t and $e(t)$ is white noise input with variance λ. The parameters of AR models are $\theta = [\alpha_1 \quad \alpha_2 \quad \ldots \quad \alpha_n]^T$ and can be derived by least square method (LSM):

$$\theta = [\Phi^T \Phi]^{-1} \Phi^T Y$$

Φ and Y are defined in [Lj99]. In some situation white noise assumption may be restrictive so one can use ARMA model. In ARMA model output of system is derived in an autoregressive manner and the input is in moving average (MA) such that:

$$y(t) + \alpha_1 y(t-1) + \ldots + \alpha_n y(t-n) = e(t) + c_1 e(t-1) + \ldots + c_m e(t-m)$$

$y(t)$ is the value of output in time t and $e(t)$ is white noise input with variance λ. The parameters of ARMA models are

$$\theta = [\alpha_1 \quad \alpha_2 \quad \ldots \quad \alpha_n \quad c_1 \quad \ldots \mathrm{A} \quad c_m]^T$$

and can be derived by several different methods [Lj99].

A more convenient way to define ARMA model is

$$A(q)y(t) = C(q)e(t)$$

where $A(q) = 1 + \alpha_1 q^{-1} + \ldots + \alpha_n q^{-n}$ and $C(q) = 1 + c_1 q^{-1} + \ldots + c_m q^{-m}$ and q is a backward shift operator.

B. Proposed Method

This paper uses ARMA model to predict traffic flow in the specific intersection. Here $y(t)$ is considered as the traffic flow in time t and $e(t)$ is the combined effect of variables which may affect the traffic flow $y(t)$. The data used in this paper are obtained from Mashhad Traffic Organization, for Moallem Blvd. intersection. The traffic flow for each approach of the intersection, for every 15 minute intervals are collected for several days using loop detectors.

15.4 Experimental Results

I. Experiment Number 1

The system is learned with data from the first day with 15 minute intervals. Then, the prediction was made 15 minutes ahead for the following day. Moreover, the ARMA model applied in this system that uses the data of the several 15 minutes before the traffic flow. In other words, 12 samples of traffic flow (3 hours) are applied to predict the traffic flow for the 15 minutes ahead. Coefficients used to predict this model are as follows:

$$A(q) = 1 - 0.8478q{-}1 - 0.3553q^{-2} + 0.2323q^{-4} + 0.03152q^{-5} + \ldots$$

$$C(q) = 1 - 0.3099q^{-1}$$

As it is shown in Figure 15.4, prediction is made after the 12th sample and all previous 12 samples are available data. In this condition, the model accuracy is 88.74 % (Table 15.1). To examine the prediction system more thoroughly, the anticipation is also made for one hour ahead. The result is shown in Figure 15.5.

As it is demonstrated in Figure 15.5, accuracy of the system is decreased a little. Considering Figure 15.5, it can be concluded that because the prediction interval has increased, the accuracy of the system has reduced to 81.96 % (Table 15.1).

II. Experiment Number 2

The same system in this step is learned with data from the first day with every 30 minutes intervals. Then the anticipation was made 30 minutes ahead for the following day. The result of the prediction for 30 minute ahead in the following day and 60 minute ahead is illustrated in Table 15.2. From Table 15.2 it can be inferred that the accuracy of the system decreased significantly.

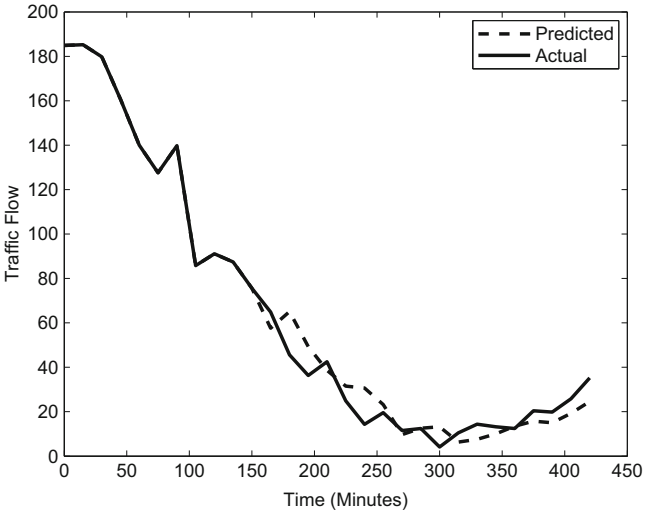


Fig. 15.4 Results for the system predicted output for 15 min later

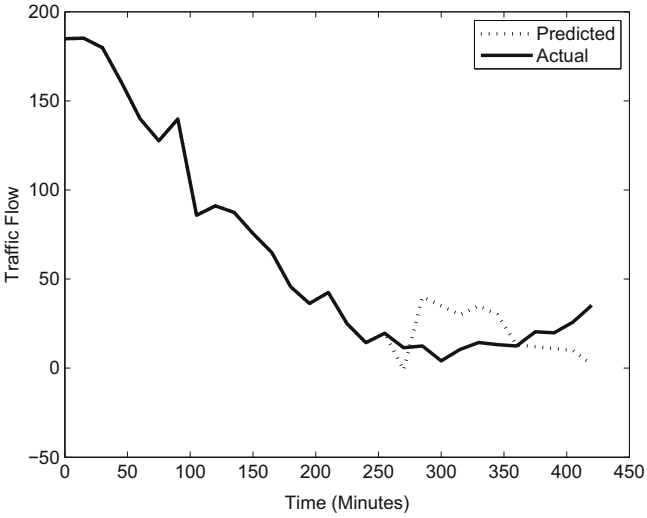


Fig. 15.5 Results for the system predicted output for 1 hour later

By comparing the two experiments which are shown in Table 15.2, it can be noted that the less the time intervals are, the more the accuracy will be achieved. Thus, it is significant to record data in shorter intervals.

Table 15.1 System accuracy (least square error) for 15 minutes interval learning data

Accuracy of the system	Forecasting ahead
88.74 %	30 min
81.96 %	60 min

Table 15.2 Different system accuracy (least square error) for two experiments

Experiment 1		Experiment 2	
Accuracy of the system	Forecasting ahead	Accuracy of the system	Forecasting ahead
88.74 %	30 min	73.76 %	30 min
81.96 %	60 min	66.01 %	60 min

15.5 Conclusion and Discussion

The burgeoning rise of car production and the following traffic increase have made it inevitable to propose a model for the online traffic control. Implementing a model for predicting traffic flow in an online method for a definite intersection would help traffic engineers to plan for the upcoming traffic of that intersection. Since the ITS systems depend on real time information, it is mandatory that to be able to have access to online models. It is important for the traffic engineers to have meticulous information about the traffic flow, before the planning and construction phase. Moreover, the ability to predict the traffic flow in a built intersection for 15 minutes or one hour later has advantages such as:

1. If the flow is predicted to be more than the intersection capacity, the approaching ways to the intersection can be limited.
2. With precise flow prediction, smaller queue length can be derived.
3. Sufficient time would be available to send police units or in case of other emergencies.
4. The intersection traffic light can be changed easier manually if the intersection is not actualized.

In this study, time series are used to create the model. Two different experiments conducted on the same data. In the first experiment, the proposed model is instructed by the data of the first day having 15 minutes intervals, and the model is tested for 15 and 60 minutes later time intervals in the following day. The average error obtained from this test is less than 18 percent for all the predictions. In the second experiment, the mode is instructed by the data of the first day having 30 minutes intervals, and the model tested for the same period as the first one. But the average error gained from the second experiment increased significantly. So using data by less interval measurements is much better than data with large interval time.

Traffic intersections may also be linked as in a network. That is, the neighbor intersections should follow a green wave to provide the most efficiency. In this

paper, we only analyzed the impact of all the four approaches in an isolated arbitrary intersection. Future research can be conducted on the impact of an intersection in a network.

Acknowledgements The authors would like to thank the Traffic Organization of Mashhad for providing the data used in this paper.

References

[BeMo99] Bemporad, A., Morari, M.: Control of systems integrating logic, dynamics, and constraints. Automatica **35**, 407–427 (1999)

[CaEtA97] Caldarelli, G., Marsili, M., Zhang, Y.C.: A prototype model of stock exchange. Europhys. Lett. **40**(5), 479 (1997)

[CoHu14] Cordova, F., Huynh, N.: Using economic indicators to perform short-term truck traffic forecasting a time series and truck traffic analysis framework. TRB Annual Meeting, Current Research in Freight Transportation and Logistics Planning and Operations (2014)

[JaKu07] Jain, A., Kumar, A.M.: Hybrid neural network models for hydrologic time series forecasting. Appl. Soft Comput. **7**(2), 585–592 (2007)

[KuEtA13] Kumar, K., Parida, M., Katiyar, V.K.: Short term traffic flow prediction for a non-urban highway using artificial neural network. Proc. Soc. Behav. Sci. **104**(2), 755–764 (2013)

[LiWa13] Liang, Z., Wakahara, Y.: City traffic prediction based on real-time traffic information for intelligent transport systems. In: IEEE, 13th International Conference on ITS Telecommunications (ITST), pp. 378–383 (2013)

[Lj99] Ljung, L.: System Identification Theory for the User. Prentice Hall Information and System Science Series. Wiley (1999)

[OgEtA11] Ogunwolu, L., Adedokun, O., Orimoloye, O., Oke, S.A.: A Neuro-Fuzzy approach to vehicular traffic flow prediction for a metropolis in a developing country. J. Ind. Eng. Int. **7**(13), 52–66 (2011)

[Pa15] Pahnabi, A.H.: Intersection traffic control with uncertainties based on mixed logical dynamical model and robust model predictive control. M.Sc. Dissertation (2015)

[QuAl08] Queen, C.M., Albers, C.J.: Forecasting traffic flows in road networks: a graphical dynamic model approach. In: Proceedings of the 28th International Symposium of Forecasting, International Institute of Forecasters (2008)

[RaEtA14] Raeesi, M., Mesgari, M.S., Mahmoudi, P.: Traffic time series forecasting by feedforward neural network: a case study based on traffic data of Monroe. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **40**(2), 219–223 (2014)

[Us12] United State Department of Transportation (USDOT-2012)

[SmEtA02] Smith, B.L., Williams B.M., Oswald K.R.: Comparison of parametric and nonparametric models for traffic flow forecasting. Trans. Res. Part C **10**, 303–321 (2002)

[SuEtA06] Sun, S., Zhang, C., Yu, G.: A Bayesian network approach to traffic flow forecasting. IEEE Trans. Intell. Trans. Syst. **7**(1), 124–132 (2006)

[YaTa00] Yao, J., Tan, C.L.: Forecasting the forex financial markets. Neurocomputing **34**, 79–98 (2000)

Chapter 16
Mathematical Modeling of One-Dimensional Oil Displacement by Combined Solvent-Thermal Flooding

T. Marotto, A. Pires, and F. Forouzanfar

16.1 Introduction

Enhanced oil recovery (EOR) is defined as a set of techniques applied to improve the recovery of hydrocarbons by the injection of materials that are not normally present in the reservoir [La89]. Most EOR methods may be classified into thermal, chemical, and miscible. The chemical methods improve the sweep efficiency through the reduction of the water mobility and/or interfacial tension. Thermal methods consist of injecting a fluid (heat source), which can be steam or hot water, that causes the reduction of the oil viscosity in the reservoir, and the miscible methods are based on the injection of a solvent to decrease the capillary and interfacial forces.

There are large heavy oil and bitumen deposits in many areas in the world [Zh04]. Under this scenario, it is important to develop new technologies to extract the vast amount of oil from these reservoirs. Carbonated water flooding (CWF) is an improved oil recovery technique that combines the advantages of waterflooding with carbon dioxide sequestration [HiEtAl60, Na89, Pi07, SoEtAl09, DoEtAl11]. In this work we present the analytical solution for the problem of 1D oil displacement by a combined thermal-solvent EOR method.

T. Marotto (✉) • A. Pires
North Fluminense State University, Macaé, RJ, Brazil
e-mail: tamires.marotto@gmail.com; adolfo.puime@gmail.com

F. Forouzanfar
The University of Tulsa, Tulsa, OK, USA
e-mail: fahim-forouzanfar@utulsa.edu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_16

16.2 Physical and Mathematical Model

The system of governing equations that models the injection of a hot fluid containing a solvent into an oil reservoir consists of oil, solvent, and water mass balance and energy conservation. The main assumptions for this model are:

- One-dimensional two-phase flow in a homogeneous porous media;
- No diffusion, no chemical reactions;
- Incompressible system;
- Gravity and capillary effects are neglected;
- Enthalpies are functions of components concentrations and temperature;
- Constant heat capacity;
- Viscosity of phases are functions of the concentration of solvent in the phase and temperature;
- Pure component density is the same in all phases;
- Only mass transfer of component solvent occurs between the phases;
- Residual saturations of phases are set to zero.

The mass conservation of components oil, solvent, and water can be written as:

$$\frac{\partial}{\partial t}\left(\phi \rho_o c_{oo} s_o\right) + \frac{\partial}{\partial x}\left(\rho_o c_{oo} u_o\right) = 0,$$

$$\frac{\partial}{\partial t}\left[\phi\left(\rho_w c_{sw} s_w + \rho_o c_{so} s_o\right)\right] + \frac{\partial}{\partial x}\left(\rho_w c_{sw} u_w + \rho_o c_{so} u_o\right) = 0$$

and

$$\frac{\partial}{\partial t}\left(\phi \rho_w c_{ww} s_w\right) + \frac{\partial}{\partial x}\left(\rho_w c_{ww} u_w\right) = 0,$$

where ϕ is the porosity, ρ_j is the density of phase j, c_{ij} is the mass fraction of component i in phase j, s_j is the saturation of phase j, u_j is the velocity of phase j, t is time, and x is the spatial variable.

The energy conservation is given by:

$$\frac{\partial}{\partial t}\left[\phi\left(\rho_o s_o H_o + \rho_w s_w H_w\right) + \left(1-\phi\right)\rho_r H_r\right] + \frac{\partial}{\partial x}\left(\rho_o H_o u_o + \rho_w H_w u_w\right) = 0,$$

where H_j is the enthalpy of phase j, ρ_r is the rock density, and H_r is the rock enthalpy.

We can define the dimensionless time (t_D) and spatial (x_D) variables by:

$$t_D = \frac{\int_0^t u_T\left(\tau\right) d\tau}{\phi L} \text{ and } x_D = \frac{x}{L}.$$

The velocity can be expressed in terms of the fractional flow function:

$$f_j = \frac{u_j}{u_T},$$

where f_j is the fractional flow of phase j and u_T is the total velocity.

Considering that Amagat's law [PrEtAl86] is valid, and that the pure component density is the same in all phases, we can replace the mass fraction by the volume fraction of component i in phase j in the conservation laws using the following equation:

$$\hat{c}_{ij} = \frac{c_{ij}\rho_j}{\hat{\rho}_i},$$

where $\hat{c}_{ij}$ is the volume fraction and $\hat{\rho}_i$ is the pure component density at system P and T.

To relate the solvent concentrations in water and oil phases we consider infinity dilution model for both phases. Therefore, using Henry's law [PrEtAl86] to calculate the fugacity of the solvent in the liquid phase, we get:

$$\hat{c}_{so}K_{s,o} = \hat{c}_{sw}K_{s,w}, \tag{16.1}$$

where $K_{s,j}$ is the Henry's constant in liquid phase j. Henry's constant was calculated using Harvey's model at water saturation pressure at system temperature [Ha96] corrected to system pressure applying Poynting's correction [PrEtAl86].

We also consider the following auxiliary relations:

$$\sum_{i=1}^{n_c}\hat{c}_{ij} = 1,\ \sum_{j=1}^{n_p}s_j = 1 \text{ and } \sum_{j=1}^{n_p}f_j = 1. \tag{16.2}$$

The unknowns of this problem are oil saturation, solvent concentration in the oil phase, and temperature. We only need to solve one concentration ($\hat{c}_{so}$), because $\hat{c}_{sw}$ can be obtained by Henry's law (Equation 16.1). The other concentrations are computed using the first auxiliary relation (Equation 16.2). The mass conservation of component water does not need to be solved as the system is incompressible.

Thus, given all these assumptions we find a hyperbolic system composed of three equations:

$$[1-\hat{c}_{so}]\frac{\partial s_o}{\partial t_D} - s_o\frac{\partial \hat{c}_{so}}{\partial t_D} + \left[(1-\hat{c}_{so})\frac{\partial f_o}{\partial s_o}\right]\frac{\partial s_o}{\partial x_D} +$$
$$\left[(1-\hat{c}_{so})\frac{\partial f_o}{\partial \hat{c}_{so}} - f_o\right]\frac{\partial \hat{c}_{so}}{\partial x_D} + \left[(1-\hat{c}_{so})\frac{\partial f_o}{\partial T}\right]\frac{\partial T}{\partial x_D} = 0$$

$$\left[1-\frac{K_{s,o}}{K_{s,w}}\hat{c}_{so}\right]\frac{\partial s_o}{\partial t_D} + \left[\frac{K_{s,o}}{K_{s,w}}(1-s_o)\right]\frac{\partial \hat{c}_{so}}{\partial t_D} + \left[\left(1-\frac{K_{s,o}}{K_{s,w}}\hat{c}_{so}\right)\frac{\partial f_o}{\partial s_o}\right]\frac{\partial s_o}{\partial x_D} +$$
$$\left[\left(1-\frac{K_{s,o}}{K_{s,w}}\hat{c}_{so}\right)\frac{\partial f_o}{\partial \hat{c}_{so}} + \frac{K_{s,o}}{K_{s,w}}(1-f_o)\right]\frac{\partial \hat{c}_{so}}{\partial x_D} + \left[\left(1-\frac{K_{s,o}}{K_{s,w}}\hat{c}_{so}\right)\frac{\partial f_o}{\partial T}\right]\frac{\partial T}{\partial x_D} = 0$$

$$\frac{\partial T}{\partial t_D} + \left\{ \begin{array}{c} f_o + \dfrac{\hat{c}_{so}\dfrac{K_{s,o}}{K_{s,w}}(M_s - M_w) + M_w}{\left\{1 + \hat{c}_{so}\left[\dfrac{K_{s,o}}{K_{s,w}}(M_w - M_s) + M_s - 1\right] - M_w\right\}} \\ s_o + \dfrac{\hat{c}_{so}\dfrac{K_{s,o}}{K_{s,w}}(M_s - M_w) + M_w + \dfrac{(1-\phi)}{\phi}M_r}{\left\{1 + \hat{c}_{so}\left[\dfrac{K_{s,o}}{K_{s,w}}(M_w - M_s) + M_s - 1\right] - M_w\right\}} \end{array} \right\} \frac{\partial T}{\partial x_D} = 0, \tag{16.3}$$

where

$$M_i = \frac{M_{Ti}}{M_{To}} = \frac{\hat{\rho}_i C_{pi}}{\hat{\rho}_o C_{po}},$$

and C_{pi} is the heat capacity of component i.

Rewriting system (Equation 16.3) in conservative form:

$$u_{t_D} + A u_{x_D} = 0,$$

we obtain an upper triangular matrix, so the characteristics velocities (eigenvalues) of this system are the elements of the main diagonal. Thus, the corresponding eigenpairs are given by:

$$\lambda^{(1)} = \frac{\partial f_o}{\partial s_o},\ r^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$\lambda^{(2)} = \frac{f_o + \dfrac{K_{s,o}}{K_{s,w}}(1 - f_o - \hat{c}_{so})}{s_o + \dfrac{K_{s,o}}{K_{s,w}}(1 - s_o - \hat{c}_{so})},\ r^{(2)} = \begin{bmatrix} 1 \\ \dfrac{\lambda^{(2)} - \lambda^{(1)}}{\dfrac{\partial f_o}{\partial \hat{c}_{so}} + C} \\ 0 \end{bmatrix},$$

$$\lambda^{(3)} = \frac{f_o + \dfrac{\hat{c}_{so}\dfrac{K_{s,o}}{K_{s,w}}(M_s - M_w) + M_w}{\left\{1 + \hat{c}_{so}\left[\dfrac{K_{s,o}}{K_{s,w}}(M_w - M_s) + M_s - 1\right] - M_w\right\}}}{s_o + \dfrac{\hat{c}_{so}\dfrac{K_{s,o}}{K_{s,w}}(M_s - M_w) + M_w + \dfrac{(1-\phi)}{\phi}M_r}{\left\{1 + \hat{c}_{so}\left[\dfrac{K_{s,o}}{K_{s,w}}(M_w - M_s) + M_s - 1\right] - M_w\right\}}},\ r^{(3)} = \begin{bmatrix} 1 \\ 0 \\ \dfrac{\lambda^{(3)} - \lambda^{(1)}}{\dfrac{\partial f_o}{\partial T}} \end{bmatrix},$$

where

$$C = -\frac{\frac{K_{s,o}}{K_{s,w}}(f_o - s_o)}{s_o + \frac{K_{s,o}}{K_{s,w}}(1 - s_o - \hat{c}_{so})}.$$

From the eigenpairs, we obtain the multipliers to calculate the rarefaction waves. So, from the first eigenpair, $\lambda^{(1)}$ and $r^{(1)}$ we get:

$$\alpha^{(1)} = \left(\frac{\partial^2 f_o}{\partial s_o^2}\right)^{-1}.$$

For the second pair, $\lambda^{(2)}$ and $r^{(2)}$ we have:

$$\alpha^{(2)} = \frac{J^2\left(\frac{\partial f_o}{\partial \hat{c}_{so}} + C\right)}{\left(\frac{\partial f_o}{\partial \hat{c}_{so}} + C\right)\left[\left(1 - \frac{K_{s,o}}{K_{s,w}}\right)(J\lambda^{(1)} - E)\right] + (\lambda^{(2)} - \lambda^{(1)})\left[\left(1 - \frac{K_{s,o}}{K_{s,w}}\right)J\frac{\partial f_o}{\partial \hat{c}_{so}} + \frac{K_{s,o}}{K_{s,w}}(E - J)\right]},$$

where

$$J = s_o + \frac{K_{s,o}}{K_{s,w}}(1 - s_o - \hat{c}_{so})$$

and

$$E = f_o + \frac{K_{s,o}}{K_{s,w}}(1 - f_o - \hat{c}_{so}).$$

And from the last eigenpair, $\lambda^{(3)}$ and $r^{(3)}$:

$$\alpha^{(3)} = 0.$$

In this case, $\nabla\lambda^{(3)} \cdot r^{(3)}$ is equal to zero for all s_o, $\hat{c}_{so}$, and T, a linearly degenerate wave [Le02].

In the first family oil saturation changes while concentration and temperature remain constant. For the second family, oil saturation and concentration change, and temperature is constant. The oil saturation and temperature change while concentration stays constant along the third family.

We also have the shock equations given by Rankine-Hugoniot conditions:

$$D = \frac{\left(1 - \hat{c}_{so}^{+}\right)f_o^{+} - \left(1 - \hat{c}_{so}^{-}\right)f_o^{-}}{\left(1 - \hat{c}_{so}^{+}\right)s_o^{+} - \left(1 - \hat{c}_{so}^{-}\right)s_o^{-}},$$

$$D = \frac{\hat{c}_{so}^{+}\left[\frac{K_{s,o}}{K_{s,w}} + \left(1 - \frac{K_{s,o}}{K_{s,w}}\right) f_o^{+}\right] - \hat{c}_{so}^{-}\left[\frac{K_{s,o}}{K_{s,w}} + \left(1 - \frac{K_{s,o}}{K_{s,w}}\right) f_o^{-}\right]}{\hat{c}_{so}^{+}\left[\frac{K_{s,o}}{K_{s,w}} + \left(1 - \frac{K_{s,o}}{K_{s,w}}\right) s_o^{+}\right] - \hat{c}_{so}^{-}\left[\frac{K_{s,o}}{K_{s,w}} + \left(1 - \frac{K_{s,o}}{K_{s,w}}\right) s_o^{-}\right]}$$

and

$$D = \frac{T^{+}\left\{\left[\hat{c}_{so}^{+}F + H\right] f_o^{+} + \hat{c}_{so}^{+} I + M_{Tw}\right\} - T^{-}\left\{\left[\hat{c}_{so}^{-}F + H\right] f_o^{-} + \hat{c}_{so}^{-} I + M_{Tw}\right\}}{T^{+}\left\{\left[\hat{c}_{so}^{+}F + H\right] s_o^{+} + \hat{c}_{so}^{+} I + G\right\} - T^{-}\left\{\left[\hat{c}_{so}^{-}F + H\right] s_o^{-} + \hat{c}_{so}^{-} I + G\right\}},$$

where

$$F = \left(\frac{K_{s,o}}{K_{s,w}}\left(M_{Tw} - M_{Ts}\right) + M_{Ts} - M_{To}\right),$$

$$G = M_{Tw} + \frac{(1-\phi)}{\phi} M_{Tr},$$

$$H = M_{To} - M_{Tw}$$

and

$$I = \frac{K_{s,o}}{K_{s,w}}\left(M_{Ts} - M_{Tw}\right).$$

16.3 Example of Solution

In this section we present a solution for this problem. Corey's model [CoEtAl56] was used to calculate the relative permeability of phases:

$$k_{rj} = k_{rj}^{0}\left(s_j^{*}\right)^{n_j},$$

where k_{rj} is the relative permeability of phase j, k_{rj}^{0} is the endpoint relative permeability of phase j, and s_j^{*} is the normalized saturation of phase j, defined as

$$s_j^{*} = \frac{s_j - s_{rj}}{1 - \sum_{j=1}^{n_p} s_{rj}},$$

where s_{rj} is the residual saturation of phase j and n_p is the number of phases. The parameters used to calculate the relative permeability are given in Table 16.1, and the curves are presented in Figure 16.1.

Table 16.1 Relative permeability parameters

Property	Oil	Water
s_{rj}	0	0
k^0_{rj}	1	1
n_j	2	2

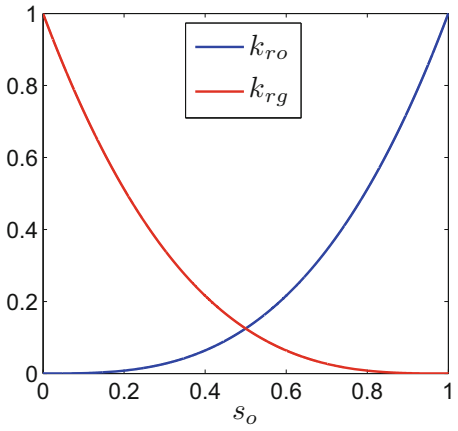


Fig. 16.1 Relative permeability curves

In this work, we will consider carbon dioxide as the solvent. Water viscosity containing carbon dioxide was calculated using the Vogel-Fulcher-Tammann (VFT) correlation [AnEtAl00]. Oil phase viscosity was determined as a function of $\hat{c}_{so}$ and T through the following polynomial function:

$$\begin{aligned}
&\mu_o[cP] \\
&= 9704 - 94.13T - 2342\hat{c}_{so} + 0.366T^2 + 18.24\hat{c}_{so}T - 11780\hat{c}^2_{so} - 7.122\text{x}10^{-4}T^3 \\
&\quad - 0.05611\hat{c}_{so}T^2 + 71.04\hat{c}^2_{so}T + 5573\hat{c}^3_{so} + 6.93\text{x}10^{-7}T^4 + 7.941\text{x}10^{-5}\hat{c}_{so}T^3 \\
&\quad - 0.1391\hat{c}^2_{so}T^2 - 27.07\hat{c}^3_{so}T + 3719\hat{c}^4_{so} - 2.696\text{x}10^{-10}T^5 - 4.293\text{x}10^{-8}\hat{c}_{so}T^4 \\
&\quad + 8.913\text{x}10^{-5}\hat{c}^2_{so}T^3 + 0.02827\hat{c}^3_{so}T^2 - 1.435\hat{c}^4_{so}T - 4023\hat{c}^5_{so},
\end{aligned}$$

where T is the absolute temperature in K and $\hat{c}_{so}$ is the volume fraction of the solvent in the oil phase.

The parameters used to build the solution are given in Table 16.2.

The solution was built from injection conditions to initial conditions (increasing self-similar variable), thus, to begin the solution path it is necessary to analyze the eigenvalues behavior at the injection conditions.

The structure of the solution path (Figures 16.2 and 16.3) is given by: $J - a - b - c \rightarrow d - e \rightarrow I$, where $(-)$ denotes a rarefaction wave and $(\rightarrow)$ indicates

Table 16.2 Physical properties

Property	Symbol	Value	Unit
System pressure	P	1.8000	MPa
Oil saturation pressure at injection conditions	$P_{sat,o}^{(J)}$	$1.31E-06$	MPa
Oil saturation pressure at initial conditions	$P_{sat,o}^{(I)}$	$2.48051E-7$	MPa
Oil saturation at injection conditions	$s_o^{(J)}$	0.0000	$\frac{m^3}{m^3}$
Oil saturation at initial conditions	$s_o^{(I)}$	1.0000	$\frac{m^3}{m^3}$
Solvent concentration at injection conditions	$\hat{c}_{so}^{(J)}$	0.1020	$\frac{m^3}{m^3}$
Solvent concentration at initial conditions	$\hat{c}_{so}^{(I)}$	0.0100	$\frac{m^3}{m^3}$
Temperature at injection conditions	$T^{(J)}$	313.1500	K
Temperature at initial conditions	$T^{(I)}$	296.1500	K
Oil heat capacity	C_{po}	1939.0015	$\frac{J}{kgK}$
Water heat capacity	C_{pw}	4527.2273	$\frac{J}{kgK}$
Solvent heat capacity	C_{ps}	2865.0307	$\frac{J}{kgK}$
Rock heat capacity	C_{pr}	720.0000	$\frac{J}{kgK}$
Oil density	ρ_o	652.2238	$\frac{kg}{m^3}$
Water density	ρ_w	861.0170	$\frac{kg}{m^3}$
Solvent density	ρ_s	911.7881	$\frac{kg}{m^3}$
Rock density	ρ_r	2200.0000	$\frac{kg}{m^3}$
Porosity	ϕ	0.3000	$\frac{m^3}{m^3}$

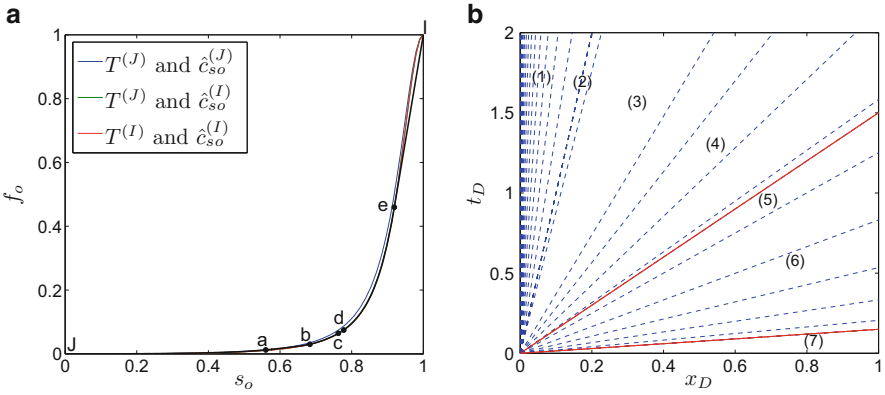


Fig. 16.2 Solution path: (**a**) Plane (s_o, f_o) (**b**) Physical plane (x_D, t_D)

a shock wave. The solution begins at point (J) with the first family rarefaction wave, where oil saturation changes from injection saturation up to (a). From (a) there is an oil saturation and concentration rarefaction (second family) connecting fractional flow at $T^{(J)}$ and $\hat{c}_{so}^{(I)}$ to point (b), where there is a constant state zone. Next, there is another oil saturation rarefaction (first family) up to point (c). From (c)

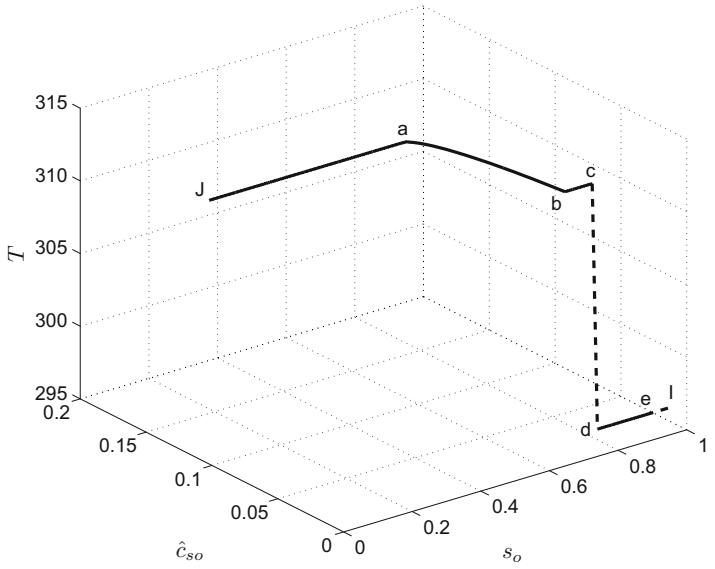


Fig. 16.3 Phase space solution

there is an oil saturation and temperature shock (third family) connecting fractional flow function at $T^{(I)}$ and $\hat{c}_{so}^{(I)}$ to point (d), the degenerate shock wave. This shock is followed by a constant state region. From (d) there is the last oil saturation rarefaction up to (e), which is connected to initial conditions (I) through a Buckley-Leverett type shock [Bu42]. The solution profiles are presented in Figure 16.4, where T_D is the dimensionless temperature, $T_D = T/T^{(J)}$. Figure 16.5 presents the solution profile in region $(J - b)$, the first saturation rarefaction and the combined saturation-concentration transition. In Figure 16.6 there is a zoom in the fractional flow curves in two regions, the first one $(a - b)$ is the saturation-concentration transition and the second one $(c \to d)$ is the degenerate shock.

16.4 Conclusions

This work presents an analytical solution for the problem of oil displacement by a hot fluid containing solvent as a combined thermal-solvent EOR method. The hyperbolic system is composed of three equations and it is solved using the method of characteristics. The solution path is composed of rarefaction, shock waves, and constant states, and this problem presents a degenerate wave. The solution presented is divided into seven regions consisting of an oil saturation rarefaction (first family)

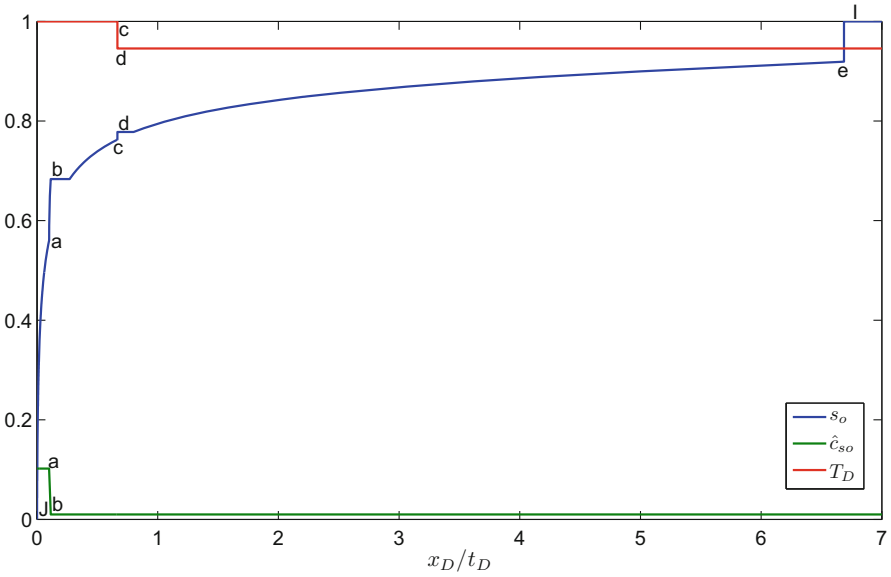


Fig. 16.4 Profiles: oil saturation, solvent concentration, and dimensionless temperature

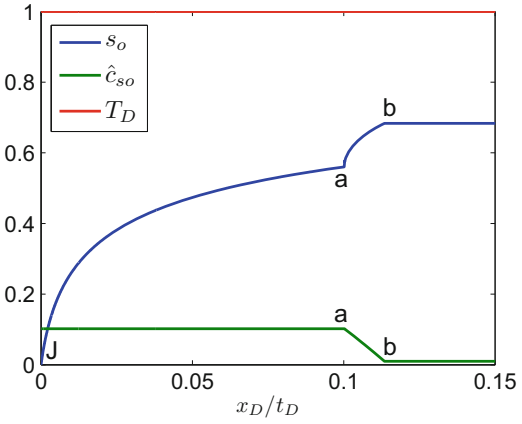


Fig. 16.5 Zoom in the solution profile in section $(J - b)$

followed by an oil saturation and solvent concentration rarefaction (second family). This rarefaction is followed by a constant state zone, then a first family rarefaction wave appears before a temperature shock (degenerate wave). It is followed by another constant state region, then an oil saturation rarefaction, and finally the first family shock (Buckley-Leverett type).

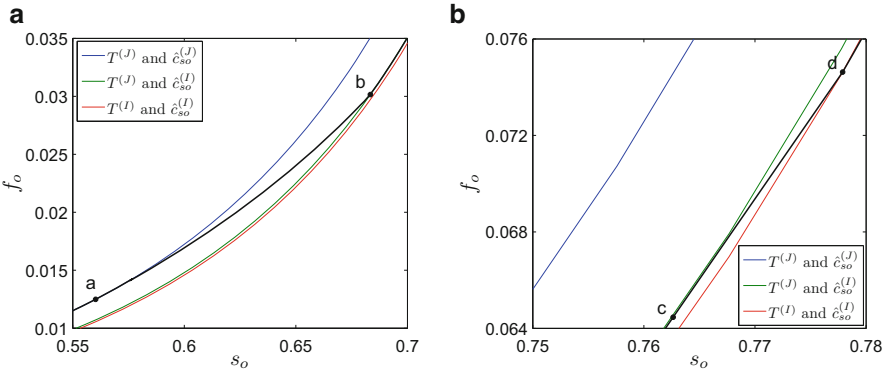


Fig. 16.6 Zoom of plane (s_o, f_o). (**a**) Region $(a - b)$. (**b**) Region $(c \rightarrow d)$

References

[AnEtAl00] Angell, C.A., Ngai, K.L., McKenna, G.B., McMillan, P.F., Martin, S.W.: Relaxation in glassforming liquids and amorphous solids. J. Appl. Phys. **88**, 3113–3157 (2000)

[Bu42] Buckley, S.E., Leverett, M.C.: Mechanisms of fluid displacement in sands. Am. Inst. Min. Metall. Pet. Eng. **146**, 107–116 (1942)

[CoEtAl56] Corey, A.T., Rathjens, C.H., Henderson, J.H., Wyllie, M.R.J.: Three-phase relative permeability. J. Can. Pet. Technol. **8**, 63–65 (1956)

[DoEtAl11] Dong, Y., Dondoruk, B., Ishizawa, C., Lewis, E.J., Kubicek, T.: An experimental investigation of carbonated water flooding. In: SPE Annual Technical Conference and Exhibition, Denver, CO (2011)

[Ha96] Harvey, A.H.: Semiempirical correlation for Henry's constants over large temperature ranges. AlChE J. **42**, 1491–1494 (1996)

[HiEtAl60] Hickok, C.W., Christensen, R.J., Ramsay, H.J.: Progress review of the K&S carbonated waterflood project. J. Pet. Sci. Technol. **12**, 20–24 (1960)

[La89] Lake, W.L.: Enhanced Oil Recovery. Prentice-Hall, Englewood Cliffs (1989)

[Le02] LeVeque, R.J.: Finite Volume Method for Hyperbolic Problems. Cambridge University Press, New York (2002)

[Na89] Nars, T.N., McKay, A.S.: Novel oil recovery processes using caustic and carbon dioxide as dual additives in hot water. In: Petroleum Conference of The South Saskatchewan Section, Regina, CA (1989)

[Pi07] Picha, M.S.: Enhanced oil recovery by hot CO_2 flooding. In: SPE Middle East Oil & Gas Show and Conference, Kingdom of Bahrain, BA (2007)

[PrEtAl86] Prausnitz, J.M., Lichtenthaler, R.N., Azevedo, E.G.: Molecular Thermodynamics of Fluid-Phase Equilibria. Prentice-Hall, Englewood Cliffs (1986)

[SoEtAl09] Sohrabi, M., Riazi, M., Jamiolahmady, M., Ireland, S., Brown, C.: Mechanisms of oil recovery by carbonated water injection. In: International Symposium of the Society of Core Analysts, Noordwijk, HO (2009)

[Zh04] Zhao, L.: Steam alternating solvent process. In: SPE International Thermal Operations and Heavy Oil Symposium and Western Meeting, Bakersfield, CA (2004)

Chapter 17
Collocation Methods for Solving Two-Dimensional Neural Field Models on Complex Triangulated Domains

R. Martin, D.J. Chappell, N. Chuzhanova, and J.J. Crofts

17.1 Introduction

The nervous system consists of approximately 10^{11} neurons and 10^{14} connections all embedded within a highly constrained anatomical space. To better understand such a complex multi-scale system, neural models are deployed that use a range of mathematical and computational techniques to explain/predict function and behaviour of the brain at a range of different scales [Am97, JiEtAl96]. One such approach, the foundations of which were laid in the 1970s by Wilson and Cowan [WiEtAl72] and Amari [Am97], is neural field theory, which employs a continuum approach to model the activity of large populations of neurons in the cortex. These techniques are of great interest, not only from a mathematical point of view, but also from an experimental neuroscience point of view since they can replicate many of the dynamic patterns of brain activity that are observed using modern neuroimaging methodologies [Co10, BoEtAl11].

Neural field models are built from neural masses and typically take the form of a nonlinear partial integro-differential equation:

$$\frac{\partial}{\partial t}u(\mathbf{x},t) = -u(\mathbf{x},t) + \int_{\Omega} w(\mathbf{x}-\mathbf{x}')S(u(\mathbf{x}'))d\mathbf{x}', \tag{17.1}$$

where $u(\mathbf{x},t)$ describes the average activity of the neuronal population at position $\mathbf{x} \in \Omega$ at time t, and S denotes the firing rate function. In our work S takes the form of a sigmoid

R. Martin (✉) • D.J. Chappell • N. Chuzhanova • J.J. Crofts
Nottingham Trent University, Nottingham, UK
e-mail: rebecca.martin022011@my.ntu.ac.uk

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_17

$$S(u) = \frac{1}{1 + e^{-\beta u}},$$

that converts population activity to firing frequency at a rate governed by the steepness parameter β. Other possibilities include, for example, the Heaviside function and piecewise linear functions [Br11]. In addition, $w(\mathbf{x}, \mathbf{x}')$ denotes the connectivity function which describes how neurons positioned at $\mathbf{x}'$ interact with neighbouring neurons at position $\mathbf{x}$. Popular connectivity functions in the literature include Gaussian, Laplace and Mexican-hat functions [Br11, SaEtAl15, RaEtAl14].

Equations of the type (17.1) have been shown to support a variety of solutions, including travelling and spiral waves, as well as spatially and temporally periodic patterns [Co10, Br11]. These pattern formations can be linked to different neurological phenomena such as bumps in models of working memory [LaEtAl02], and spiral waves that are linked to the generation of visual hallucinations [BrEtAl01]. Moreover, oscillatory and travelling waves can be the signature of neurological diseases, such as epilepsy [Br11, Er98]. Thus, understanding the types of waves, as well as mechanisms of synchrony and cortical propagation, promises to assist in the treatments of such diseases. They are also of increasing relevance in neuroimaging, interpreting (and unifying) electroencephalography, functional magnetic resonance imaging and magnetoencephalography data [Co10, BoEtAl11].

In the next section we provide details concerning the neural field model to be studied here. This is followed by a brief description of the collocation technique in §17.3. In §17.4 we show the results of applying the collocation technique to solve the neural field model defined in §17.2, and investigate the dependence of these results on the underlying mesh. We finish by giving a brief overview of the work and outlining areas for future study.

17.2 A Two-Dimensional Neural Field Model

Here we consider a two-dimensional neural field model of the type studied in [La14]:

$$\begin{aligned} \frac{\partial u(x,y,t)}{\partial t} &= A \int_0^L \int_0^L w(x-x', y-y') S(u(x', y', t) - h) dx' dy' \\ &\quad - u(x,y,t) - a(x,y,t), \\ \tau \frac{\partial a(x,y,t)}{\partial t} &= Bu(x,y,t) - a(x,y,t). \end{aligned} \tag{17.2}$$

The above includes an additional recovery variable a which acts to repolarise u via negative feedback, while the parameters A, B, h and τ are related to the sensitivities

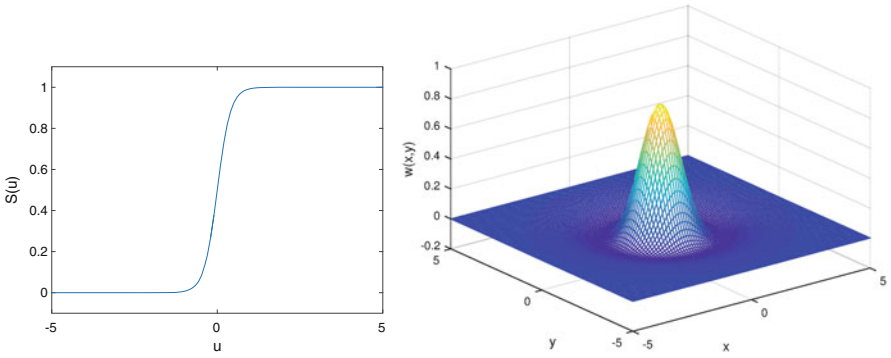


Fig. 17.1 (a) Sigmoidal firing rate function (b) A two-dimensional Mexican-hat coupling function

and time scale of the problem [La14]. As mentioned above, the integral kernel $w(x - x', y - y')$ describes interactions between neighbouring neurons. We take the following functional form for w in our work:

$$w(x, y) = e^{-(x^2+y^2)} - 0.17e^{-0.2(x^2+y^2)},$$

which is a Mexican-hat type function; see Figure 17.1b. Note that Equation (17.2) is typically solved over a square domain $\Omega = [-L, L]^2$ with periodic boundary conditions in both x and y.

In Ref. [La14], the neural field model in (17.2) is shown to admit a stable bump solution travelling from right to left, using Fast Fourier transforms (FFTs) to evaluate the integral form of the equation in a highly idealised setting. For comparative purposes, Figure 17.2 displays a travelling bump solution for parameter values matching those in [La14]. In the following, we investigate such solutions when employing collocation techniques that, unlike FFTs, can be deployed on the more general, typically asymmetric domains, that result from modern neuroimaging studies.

17.3 The Collocation Method

Collocation is an example of a projection method that approximates the infinite dimensional problem in (17.2) by a finite dimensional one, via a suitably defined projection operator $\mathcal{P}_n$. Below we provide brief details of the method as applied to Equation (17.2), the interested reader, however, should consult the excellent text by Atkinson [At97] for further details.

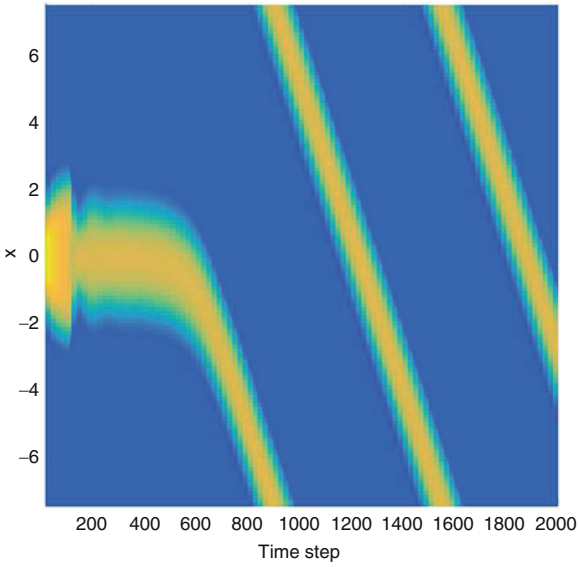


Fig. 17.2 Bump solution of Equation (17.2) computed using FFTs and parameter values matching those in [La14]

Consider the following triangulation $\mathcal{T}_n = \{\triangle_1, \ldots, \triangle_n\}$ of the square $[-L, L]^2$ and suppose that on each triangle $\triangle_k$ we employ a piecewise linear approximation of the unknown functions $u(x, y, t)$ and $a(x, y, t)$. In this case the projection operator takes the form

$$\begin{aligned}\mathcal{P}_n u(x, y, t) &= u_n(x, y, t) \\ &= \sum_{j=1}^{3} u(\mathbf{v}_{k,j}, t) l_j(x, y), \quad (x, y) \in \triangle_k,\ k = 1, 2, \ldots, n.\end{aligned}$$

Here, $v_{k,j}$ denotes the (x, y) coordinates of the j^{th} interpolation point of the k^{th} triangle $\triangle_k$, while l_j denotes the linear Lagrange basis functions [At97]. Note that a similar equation holds for $a(x, y, t)$.

The above allows us to formulate the following approximation to (17.2):

$$\begin{aligned}\frac{\partial u_n(x, y, t)}{\partial t} &= A\mathcal{P}_n \left\{ \int_{-L}^{L} \int_{-L}^{L} w(x - x', y - y') S(u - h) dx' dy' \right\} \\ &\quad - u_n(x, y, t) - a_n(x, y, t), \\ \tau \frac{\partial a_n(x, y, t)}{\partial t} &= B u_n(x, y, t) - a_n(x, y, t).\end{aligned} \tag{17.3}$$

Assuming this expression holds exactly at the node values $v_1, v_2, \ldots, v_{n_v}$, where n_v refers collectively to a global numbering of the node points $v_{k,j}$, we obtain a collocation scheme for (17.2).

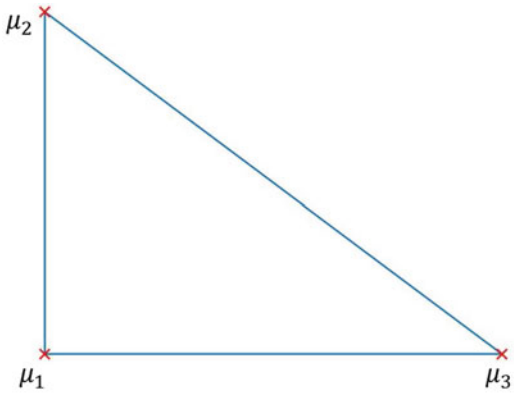


Fig. 17.3 The unit simplex and the three linear interpolation nodes. Here, $\mu_1 = (0, 0)$, $\mu_2 = (0, 1)$, $\mu_3 = (1, 0)$

To make the above collocation scheme more tractable we perform the integration in (17.3) by applying a quadrature rule over each triangle and summing the result. More specifically, we employ the transformation $T_k : \sigma \to \triangle_k$, given by

$$(x, y) = T_k(r, s) = (1 - r - s)v_{k,1} + sv_{k,2} + rv_{k,3},$$

which maps the unit simplex σ (see Figure 17.3) on to each triangle $\triangle_k$. This enables us to integrate an arbitrary function, g say, over the triangle $\triangle_k$ as follows:

$$\int_{\triangle_k} g(x, y)dxdy = 2\text{Area}(\triangle_k) \int_{\sigma} g(T_k(r, s))drds.$$

Substituting this expression into (17.3) gives

$$\begin{aligned} \frac{du_n(v_i)}{dt} &= 2A \sum_{k=1}^{n} \text{Area}(\triangle_k) \int_{\sigma} w(v_i - T_k(r, s))S\left(\sum_{j=1}^{3} u(v_{k,j})l_j(r, s) - h\right) drds \\ &\quad - u_n(v_i) - a_n(v_i), \\ \tau \frac{da_n(v_i)}{dt} &= Bu_n(v_i) - a_n(v_i), \end{aligned} \tag{17.4}$$

for $i = 1, \ldots, n_v$, which is a system of $2n_v$ ordinary differential equations (ODE) that can be solved to determine approximate solutions to (17.2).

17.4 Results

Numerical computations were performed on the domain $\Omega = [-L, L]^2$ with periodic boundary conditions in both x and y, and the parameters in (17.2) set equal to those in [La14], *i.e.* $A = 2, B = 0.4, h = 0.8, \tau = 3, \beta = 5$ and $L = 7.5$. The system of

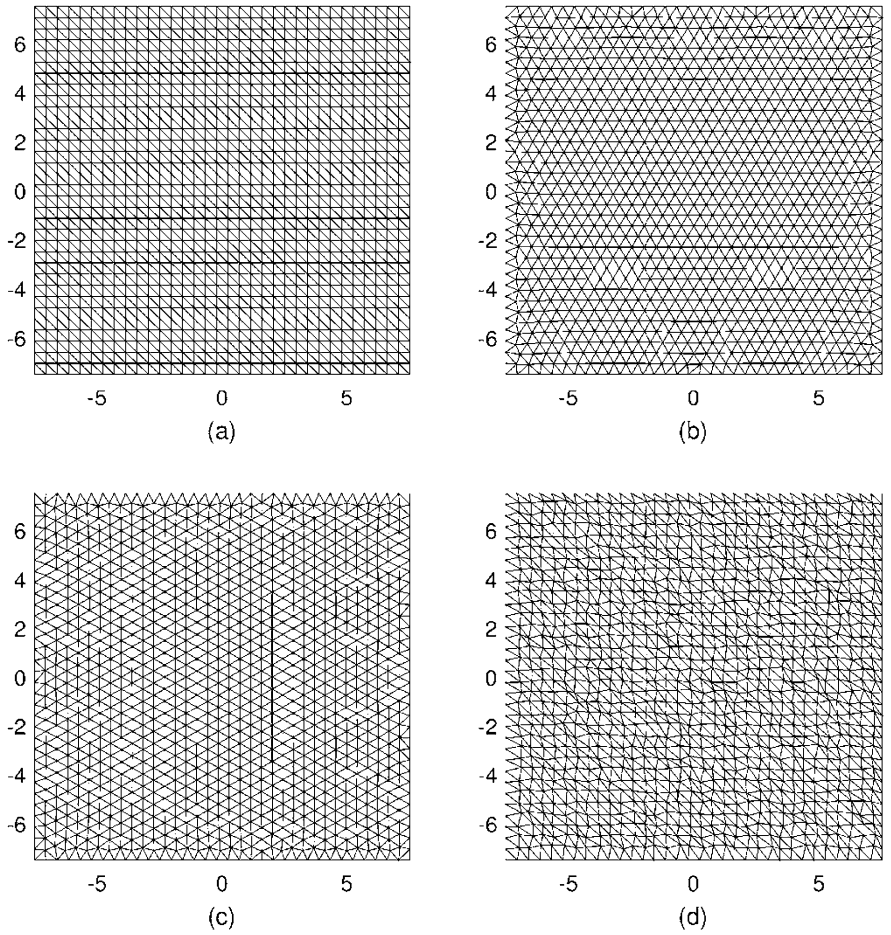


Fig. 17.4 Illustrative examples of the four different triangulations of the domain $[-L, L]^2$ (with $L = 7.5$) used in this work: (a) a regular mesh based upon Cartesian grid points; (b)–(c) triangulations constructed using the DistMesh software package, optimised so as to maximise the number of equilateral triangles in the mesh; and (d) an irregular mesh obtained by randomly perturbing the interior points of the regular mesh in (a)

ODEs in (17.4) was solved using Euler's method to step forward in time with step size $\Delta t = 0.2$ in all of our experiments. Note that similar results were attained for a range of step sizes (results omitted for brevity).

We are interested not only in the ability of collocation techniques to reproduce the FFT solution shown in Figure 17.2, but also in the ramifications that mesh regularity has on both the accuracy of solutions and the efficiency with which we obtain them. Thus we consider four different meshes in our experiments: three regular and one random (see Figure 17.4 for an illustration of the meshes deployed in this

study). The first mesh (Figure 17.4(a)) is the result of discretising the square domain $[-L, L]^2$ into $N = 129$ equally spaced points in both the x and y directions, in identical manner to that used to produce the FFT solution shown in Figure 17.2, followed by a subdivision of each square in the grid into two triangular elements. The resulting triangulation consists of $n = 8192$ triangles and $n_v = 16,641$ node values. The meshes displayed in Figures 17.4(b) and (c) were constructed using the DistMesh package [PeEtAl04] employing horizontal and vertical segments, respectively. The package was set to optimise the node positions so that the mesh consisted mainly of equilateral triangles, and the number of triangles chosen to match as close as possible the numbers for the regular mesh. In all of our experiments, the number of triangles in the DistMesh grids (Figures 17.4(b) and (c)) was such that the number of ODEs in (17.4) was in the range $n_v = 16,641 \pm 100$. Finally, we considered a randomised mesh (Figure 17.4(d)), which we constructed by perturbing at random the interior points of the regular mesh given in Figure 17.4(a). Note that boundary nodes were fixed constant for all four meshes in order to implement the periodicity of the problem more easily.

The results for each of the four domains are shown in Figure 17.5. In all cases we find spatially localised solutions in the form of a travelling pulse, or bump, similar to that obtained using FFTs (see Figure 17.2). In the FFT solution the bump is centred at $y = 0$, and whilst we found that we could accurately reproduce this solution employing collocation techniques on a regular grid, we observed a drift in the y axis when using an irregular grid, such as the ones shown in Figure 17.4(b)–(d). In particular, domain (b) exhibits a slight drift around the 700th time step, whilst the random mesh in (d) exhibits a slight drift almost immediately (approximately at the 200th time step), as shown in Figure 17.6. Note that we have conducted experiments with varying numbers of spatial grid points and have observed a relationship between the number of grid points and the time step at which the bump solution drifts from $y = 0$. In particular, the larger the number of grid points the longer the bump will travel before drifting. Moreover, we have considered higher-order polynomial approximations and early indications are that these techniques result in more reliable solutions that more closely match that of the standard FFT one. This gives us confidence in the method as it implies that with enough grid points and computational power it can reproduce the same types of solution as that obtained using FFTs, regardless of the underlying mesh.

17.5 Conclusions

In this work, we employed collocation techniques to solve a two-dimensional neural field model on the periodic, square domain $\Omega = [-L, L]^2$. Importantly, we found that these techniques were capable of reproducing solutions found by standard Fourier based methods, and also, that these results were not dependent upon the underlying mesh (for large enough grid size). The significance of the aforementioned results are twofold: firstly, unlike Fourier based methods,

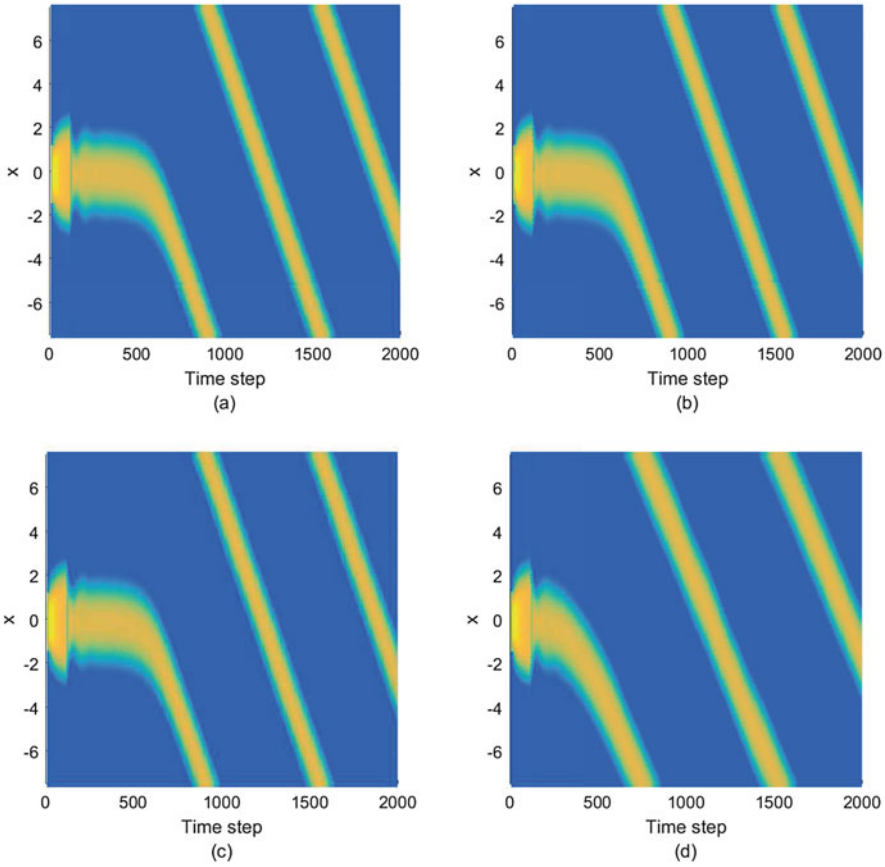


Fig. 17.5 Figures (a)–(d) show travelling bump solutions of Equation (17.2) computed using a series of increasingly irregular meshes, as illustrated in Figure 17.4(a)–(d)

collocation techniques can be deployed on complex triangulated domains, more akin to the types of geometries resulting from neuroimaging studies; and secondly, such techniques have the ability to handle more general connectivity kernels that better reflect physiology – including, for example, longer range connections (or 'short-cuts') that can result due to the convoluted nature of the cortex [HeEtAl14, OdEtAL13, LoEtAl15]. Future work shall deploy the methods discussed here, in conjunction with efficient numerical schemes for computing geodesic distances, to solve neural field models on two-dimensional curved geometries such as a sphere or torus, with the overarching aim of extending these methods to more physiologically realistic cortical domains.

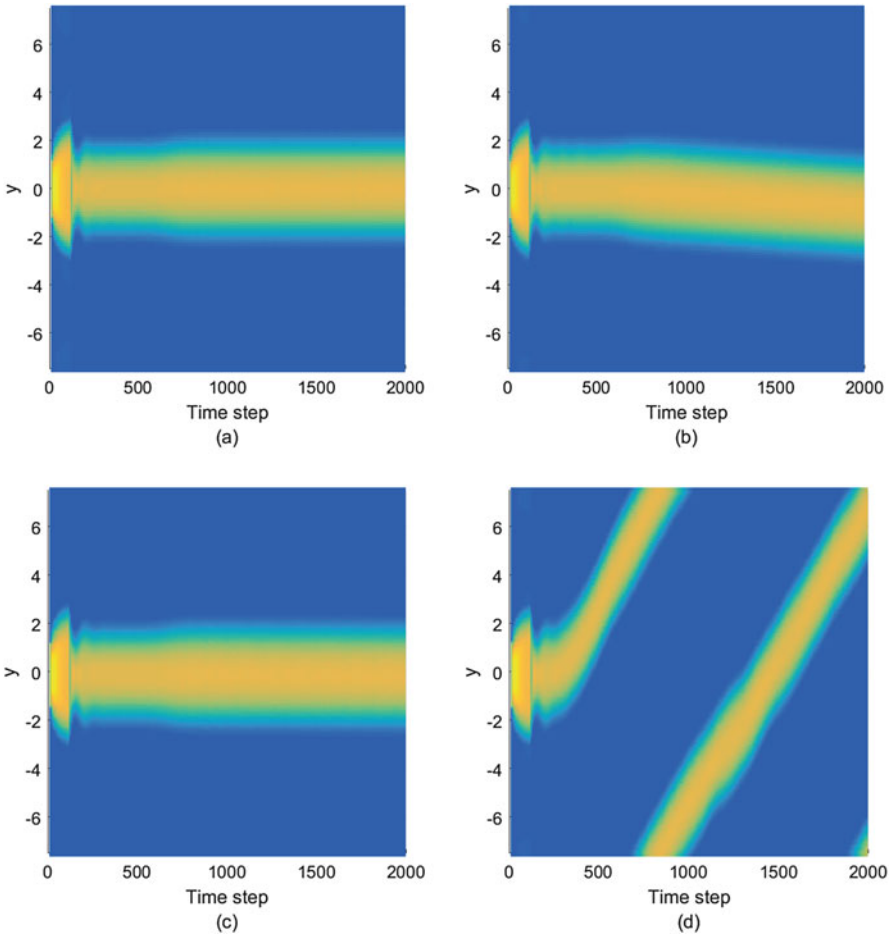


Fig. 17.6 Figures (a)–(d) show that for increasing mesh irregularity (see Figure 17.4(a)–(d)), solutions of Equation (17.2) can exhibit a drift about the y axis due to numerical errors

References

[Am97] Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biol. Cybern. **27**(2), 77–87 (1977)

[At97] Atkinson, K.E.: The Numerical Solution of Integral Equations of the Second Kind, vol. 4. Cambridge University Press, Cambridge (1997)

[BoEtAl11] Bojak, I., Oostendorp, T.F., et al.: Towards a model-based integration of co-registered electroencephalography/functional magnetic resonance imaging data with realistic neural population meshes. Phil. Trans. R. Soc. A **369**, 3785–3801 (2011)

[Br11] Bressloff, P.C.: Spatiotemporal dynamics of continuum neural fields. J. Phys. A Math. Theor. **45**(3), 033001 (2011)

[BrEtAl01] Bressloff, P.C., Cowan, J.D., et al.: Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. Phil. Trans. R. Soc. B: Biol. Sci. **356**(1407), 299–330 (2001)

[Co10] Coombes, S.: Large-scale neural dynamics: simple and complex. NeuroImage **52**(3), 731–739 (2010)

[Er98] Ermentrout, B.: Neural networks as spatio-temporal pattern-forming systems. Rep. Progress Phys. **61**(4), 353–430 (1998)

[HeEtAl14] Henderson, J.A., Robinson, P.A.: Relations between geometry of cortical gyrification and white matter network architecture. Brain Connect. **4**(2), 112–130 (2014)

[JiEtAl96] Jirsa, V.K., Hermann H.: Field theory of electromagnetic brain activity. Phys. Rev. Lett. **77**(5), 960–963 (1996)

[La14] Laing, C.R.: Numerical bifurcation theory for high-dimensional neural models. J. Math. Neurosci. **4**, 21908567 (2014)

[LaEtAl02] Laing, C.R., Troy, W.C., et al.: Multiple bumps in a neuronal model of working memory. SIAM J. Appl. Math. **63**(1), 62–97 (2002)

[LoEtAl15] Lo, Y.-P., O'Dea, R., et al.: A geometric network model of intrinsic grey-matter connectivity of the human brain. Sci. Rep. **5**, 15397 (2015)

[OdEtAL13] O'Dea, R., Crofts, J.J., et al.: Spreading dynamics on spatially constrained complex brain networks. J. R. Soc. Interface **10**(81), 20130016 (2013)

[PeEtAl04] Persson, P.O., Strang, G.: A simple mesh generator in MATLAB. SIAM Rev. **46**(2), 329–345 (2004)

[RaEtAl14] Rankin, J., Avitabile, D., et al.: Continuation of localized coherent structures in nonlocal neural field equations. SIAM J. Sci. Comput. **36**(1), B70–B93 (2014)

[SaEtAl15] Sanz-Leon, P., Knock, S.A., et al.: Mathematical framework for large-scale brain network modelling in the virtual brain. Neuroimage **111**, 385–430 (2015)

[WiEtAl72] Wilson, H.R., Cowan, J.D.: Excitatory and inhibitory interactions in localized populations of model neurons. Biophys. J. **12**(1), 1–24 (1972)

Chapter 18
Kulkarni Method for the Generalized Airfoil Equation

A. Mennouni

18.1 Mathematical Background

We consider the following generalized airfoil equation:

$$\frac{1}{\pi}\oint_{-1}^{1}\frac{\omega(t)x(t)}{t-s}dt+\int_{-1}^{1}\omega(t)k(s,t)x(t)dt=f(s),\quad -1<s<1, \tag{18.1}$$

where the first integral is a Cauchy principal value, $k(\cdot,\cdot)$ is a Fredholm kernel and

$$\omega(t):=\sqrt{\frac{1-t}{1+t}},\quad -1<t<1.$$

We introduce the following inner products:

$$\langle f,g\rangle_{\omega}:=\int_{-1}^{1}\omega(t)f(t)\overline{g(t)}dt,$$

$$\langle f,g\rangle_{\omega^{-1}}:=\int_{-1}^{1}\omega(t)^{-1}f(t)\overline{g(t)}dt,$$

and the following weighted spaces:

$$\mathcal{H}_{\omega}:=L^2_{\omega}((-1,1),\mathbb{C})=\left\{\varphi:(-1,1)\to\mathbb{C},\quad \int_{-1}^{1}\omega(t)\left|\varphi(t)\right|^2dt<\infty\right\},$$

A. Mennouni (✉)
Department of Mathematics, LTM, University of Batna 2, Batna, Algeria
e-mail: aziz.mennouni@yahoo.fr

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_18

$$\mathcal{H}_{\omega^{-1}} := L^2_{\omega^{-1}}((-1,1),\mathbb{C})$$
$$= \left\{\varphi : (-1,1) \to \mathbb{C}, \quad \int_{-1}^{1} \omega(t)^{-1} |\varphi(t)|^2 \, dt < \infty\right\}.$$

Norms in $\mathcal{H}_\omega$ and $\mathcal{H}_{\omega^{-1}}$ are defined by

$$\|\varphi\|_w := \left(\int_{-1}^{1} \omega(t) |\varphi(t)|^2 \, dt\right)^{\frac{1}{2}},$$
$$\|\varphi\|_{\omega^{-1}} := \left(\int_{-1}^{1} \omega(t)^{-1} |\varphi(t)|^2 \, dt\right)^{\frac{1}{2}},$$

respectively.

Let $(\psi_j)_{j\geq 1}$ denote the normalized sequence of Chebyshev polynomials:

$$\psi_j(s) := \frac{1}{\sqrt{\pi}} \frac{\sin(j+\frac{1}{2})\theta}{\sin\frac{\theta}{2}}, \quad \theta := \cos^{-1} s, \quad j \geq 1.$$

The sequence $(\psi_j)_{j\geq 1}$ is an orthogonal basis of $\mathcal{H}_\omega$.

Let $(\pi_{\omega,n})_{n\geq 1}$ be the sequence of bounded finite rank orthogonal projections in $\mathcal{H}_\omega$ defined by

$$\pi_{\omega,n} x := \sum_{j=1}^{n} \langle x, \psi_j \rangle_\omega \, \psi_j, \quad x \in \mathcal{H}_\omega.$$

Let be

$$Jx(s) := \frac{1}{\pi} \oint_{-1}^{1} \frac{\omega(t)x(t)}{t-s} dt, \quad x \in \mathcal{H}_\omega, \quad -1 < s < 1,$$
$$Lx(s) := -\int_{-1}^{1} \omega(t)k(s,t)x(t)dt, \quad x \in \mathcal{H}_\omega, \quad -1 < s < 1.$$

We assume that L is Hilbert-Schmidt operator from $\mathcal{H}_\omega$ into $\mathcal{H}_{\omega^{-1}}$, that is,

$$\int_{-1}^{1}\int_{-1}^{1} \frac{\omega(t)}{\omega(s)} |k(s,t)|^2 \, dsdt < \infty,$$

so L is compact.

Note that J: $\mathcal{H}_\omega \to \mathcal{H}_{\omega^{-1}}$ is unitary, i;e; $J^{-1} = J^*$, and hence

$$\|J\| = \left\|J^{-1}\right\| = 1.$$

We approximate the solution of a first kind operator equation:

$$(J - L)x = f,$$

using the orthogonal projections $(\pi_{\omega,n})_{n\geq 1}$ based on the Chebyshev polynomials. For this purpose let us consider the following finite rank operator:

$$(J^*L)^K_{\omega,n} := \pi_{\omega,n}(J^*L) + (J^*L)\pi_{\omega,n} - \pi_{\omega,n}(J^*L)\pi_{\omega,n}.$$

18.2 Description of the Method

The approximate equation is

$$\left[I - \pi_{\omega,n}(J^*L) - (J^*L)\pi_{\omega,n} + \pi_{\omega,n}(J^*L)\pi_{\omega,n}\right] x_n^K = J^*f,$$

Following Kulkarni, let

$$u_n := \pi_{\omega,n} x_n^K.$$

Since $\pi_{\omega,n}u_n = u_n$, there exist scalars $c_{n,j}$ such that

$$u_n = \sum_{j=1}^{n} c_{n,j}\psi_j.$$

Let $Q_{\omega,n}$ be defined by

$$Q_{\omega,n} := I - \pi_{\omega,n}.$$

Following [Ku03, Me12],

$$u_n - \left[\pi_{\omega,n}(J^*L)\pi_{\omega,n} + \pi_{\omega,n}(J^*L)Q_{\omega,n}(J^*L)\pi_{\omega,n}\right] u_n = \pi_{\omega,n}J^*f + \pi_{\omega,n}(J^*L)Q_{\omega,n}J^*f,$$

so

$$\sum_{j=1}^{n} c_{n,j}\left[\psi_j - (\pi_{\omega,n}(J^*L)\psi_j + \pi_{\omega,n}(J^*L)Q_{\omega,n}(J^*L)\psi_j)\right] = \pi_{\omega,n}J^*f + \pi_{\omega,n}(J^*L)Q_{\omega,n}J^*f,$$

and

$$\sum_{j=1}^{n} c_{n,j}\left[\psi_j - \sum_{k=1}^{n}\left(\left\langle (J^*L)\psi_j, \psi_k\right\rangle_\omega + \left\langle (J^*L)Q_{\omega,n}(J^*L)\psi_j, \psi_k\right\rangle_\omega\right)\psi_k\right]$$
$$= \sum_{k=1}^{n} \langle J^*f, \psi_k\rangle_\omega\, \psi_k + \sum_{k=1}^{n} \langle (J^*L)Q_{\omega,n}J^*f, \psi_k\rangle_\omega\, \psi_k.$$

Performing the inner product with ψ_i, we obtain the linear system for any i in $[\![1, n]\!]$,

$$c_{n,i} - \sum_{j=1}^{n} c_{n,j}\left[\left\langle J^*L\psi_j, \psi_i\right\rangle_\omega + \left\langle J^*LQ_{\omega,n}J^*L\psi_j, \psi_i\right\rangle_\omega\right]$$
$$= \langle J^*f, \psi_i\rangle_\omega + \langle J^*LQ_{\omega,n}J^*f, \psi_i\rangle_\omega, \quad i \in [\![1, n]\!],$$

which becomes

$$c_{n,i} - \sum_{j=1}^{n}\left[\left\langle J^*L\psi_j, \psi_i\right\rangle_\omega + \left\langle (J^*L)^2\psi_j, \psi_i\right\rangle_\omega - \sum_{k=1}^{n}\left\langle J^*L\psi_j, \psi_k\right\rangle_\omega \langle J^*L\psi_k, \psi_i\rangle_\omega\right]c_{n,j}$$
$$= \langle J^*f, \psi_i\rangle_\omega + \langle J^*LJ^*f, \psi_i\rangle_\omega - \sum_{k=1}^{n} \langle J^*f, \psi_k\rangle_\omega \langle J^*L\psi_k, \psi_i\rangle_\omega, \quad i \in [\![1, n]\!]. \tag{18.2}$$

Once the system (18.2) is solved, x_n^K is built as

$$x_n^K = u_n + Q_{\omega,n}J^*Lu_n + Q_{\omega,n}J^*f$$
$$= u_n + J^*Lu_n - \pi_{\omega,n}J^*Lu_n + J^*f - \pi_{\omega,n}J^*f.$$

18.3 Convergence Analysis

For $r \geq 0$, consider the following inner product:

$$\langle f, g\rangle_{\omega,r} := \sum_{i=0}^{\infty} (1+i)^{2r} \langle f, \psi_i\rangle_\omega \langle g, \psi_i\rangle_{\omega,r}.$$

We define the subspace $\mathcal{H}_{\omega,r}$ of $\mathcal{H}_\omega$ by

$$\mathcal{H}_{\omega,r} := L^2_{\omega,r}((-1,1),\mathbb{C}) = \left\{\varphi \in \mathcal{H}_\omega, \quad \|\varphi\|_{\omega,r} < \infty\right\}.$$

Its norm is given by

$$\|\varphi\|_{\omega,r} := \sqrt{\langle \varphi, \varphi \rangle_{\omega,r}} = \left[\sum_{i=0}^{\infty} (1+i)^{2r} |\widehat{\varphi}_i|^2 \right]^{\frac{1}{2}},$$

where $\widehat{\varphi}_i$ is the Fourier-Jacobi coefficient of φ :

$$\widehat{\varphi}_i := \int_{-1}^{1} \omega(t)\varphi(t)\psi_i(t)dt.$$

Following [BeEtAl92],

$$\|x - \pi_{\omega,n}x\|_{\omega} \leq cn^{-r}\|x\|_{\omega,r} \quad \text{for all} \quad x \in \mathcal{H}_{\omega,r}. \tag{18.3}$$

In this paper we assume that $k(.,.) \in L^2_{\omega,r}((-1,1)^2, \mathbb{C})$.

Proposition 1 *For $x \in \mathcal{H}_{\omega,r}$, the following estimate holds:*

$$\|J^* L Q_{\omega,n} x\|_{\omega} \leq \frac{\alpha}{\pi} n^{-2r} \|x\|_{\omega,r} \quad \textit{for some positive constant } \alpha.$$

Proof For all $x \in \mathcal{H}_{\omega,r}$,

$$\begin{aligned} |J^* L Q_{\omega,n} x| &= \left| \frac{1}{\pi} \oint_{-1}^{1} \frac{L Q_{\omega,n} x(t)}{\omega(t)(s-t)} dt \right| \\ &= \frac{1}{\pi} \left\langle L Q_{\omega,n} x, h_s \right\rangle_{\omega^{-1}}, \quad -1 < s < 1, \end{aligned}$$

where

$$h_s(t) := \frac{1}{s-t}.$$

Hence

$$\begin{aligned} |J^* L Q_{\omega,n} x| &= \frac{1}{\pi} \left\langle Q_{\omega,n} x, L^* h_s \right\rangle_{\omega} \\ &= \frac{1}{\pi} \left\langle Q_{\omega,n} x, Q_{\omega,n} L^* h_s \right\rangle_{\omega}. \end{aligned}$$

This leads to

$$\|J^* L Q_{\omega,n} x\|_{\omega} \leq \frac{1}{\pi} \|Q_{\omega,n} x\|_{\omega} \|Q_{\omega,n} L^* h_s\|_{\omega}.$$

Hence, by (18.3),

$$\|J^*LQ_{\omega,n}x\|_\omega \leq \frac{1}{\pi}c_0 n^{-2r}\|x\|_{\omega,r}\|L^*h_s\|_{\omega,r}, \quad \text{for some positive constant } c_0,$$

which completes the proof.
The convergence order of Kulkarni method is given in the following theorem.

Theorem 1 *The following estimate holds:*

$$\|x_n^K - x\|_\omega \leq \frac{\beta}{\pi}\, n^{-3r}\|x\|_{\omega,r} \quad \textit{for some positive constant } \beta.$$

Proof Since

$$x_n^K - x = (f + (J^*L)_n^K x_n^K) - (f + J^*Lx) = (J^*L)_n^K(x_n^K - x) + ((J^*L)_n^K - J^*L)x,$$
$$(I - (J^*L)_n^K)(x_n^K - x) = ((J^*L)_n^K - J^*L)x,$$

and

$$x_n^K - x = (I - (J^*L)_n^K)^{-1}((J^*L)_n^K - J^*L)x,$$

which lends to

$$\left\|x_n^K - x\right\|_\omega \leq \left\|(I - (J^*L)_n^K)^{-1}\right\| \left\|((J^*L)_n^K - J^*L)x\right\|_\omega.$$

Since J^*L is compact, [AhEtAl01] shows that $(I - (J^*L)_{\omega,n}^K)^{-1}$ exists and is uniformly bounded for n large enough.

Hence

$$\left\|x_n^K - x\right\|_\omega \leq C_1 \left\|((J^*L)_n^K - J^*L)x\right\|_\omega \quad \text{for some positive constant } C_1.$$

Also,

$$\left((J^*L)_n^K - J^*L\right)x = \left[\pi_{\omega,n}J^*LQ_{\omega,n} - J^*LQ_{\omega,n}\right]x = -Q_{\omega,n}J^*LQ_{\omega,n}x,$$

and using (18.3),

$$\|x_n^K - x\|_\omega \leq C_1C_2 n^{-r}\, \|J^*LQ_{\omega,n}x\|_{\omega,r} \quad \text{for some positive constant } C_2.$$

The result follows.

Table 18.1 Absolute errors

n	$\|x - x_n^K\|_\omega$
3	4.27e-5
5	5.52e-6
7	4.91e-6
8	4.75e-6
10	4.60e-6
15	4.39e-6

18.4 Numerical Example

In this example we consider the airfoil equation (18.1) with $k(s,t) = s + t$ and f such that the exact solution be

$$x(s) = (1-s)\left(\frac{1}{2}s^3 + s\right).$$

Table 18.1 shows the corresponding absolute errors for different values of n and confirms the theoretical results.

References

[AhEtAl01] Ahues, M., Largillier, A., Limaye, B.V.: Spectral Computations for Bounded Operators. CRC, Boca Raton (2001)

[BeEtAl92] Berthold, D., Hoppe, W., Silbermann, B.: A fast algorithm for solving the generalized airfoil equation. J. Comput. Appl. Math. **43**, 185–219 (1992)

[Ku03] Kulkarni, R.: A superconvergence result for solutions of compact operator equations. Bull. Aust. Math. Soc. **68**, 517–528 (2003)

[Me12] Mennouni, A.: Two projection methods for skew-Hermitian operator equations. Math. Comput. Modell. **55**, 1649–1654 (2012)

Chapter 19
Droplet Deposition and Coalescence in Curved Pipes

H. Nguyen, R. Mohan, O. Shoham, and G. Kouba

19.1 Introduction

Wet gas separation is a process in which entrained liquid droplets are separated from gas-liquid flow to ensure the gas quality for use. The separation process also aims at eliminating liquid carryover, a phenomenon in which liquid droplets are carried out in the gas outlet, in order to ensure no failure of downstream devices, such as compressors, meters, and scrubbers. Installation of a piping system upstream of separators, which are typically configured for layout convenience, can compromise the separation efficiency.

Piping components can be utilized as flow conditioning devices upstream of separators. Different configurations have been used in the field, such as short and long elbow bends, cushion tee bends, and impact tee bends. These components coalesce and remove droplets from a wet gas flow, before it enters separators, resulting in improvement of the separator efficiency. Only a few studies have been conducted on piping components utilized as flow conditioning devices. There is a lack of data and predictive methods on the design and performance of piping components used as flow conditioning devices. This is the gap that the present study attempts to address.

Several investigators studied flow in curved pipes. These include a study on gas-liquid two-phase flow in a helical coil conducted by Banerjee et al. [Ba61]. Shoham et al. [ShBr87] conducted an experimental and theoretical study on gas-liquid two-phase flow splitting in a horizontal regular pipe tee. Feng [Fe09] conducted an

H. Nguyen • R. Mohan (✉) • O. Shoham
The University of Tulsa, Tulsa, OK, USA
e-mail: hungnhu-nguyen@utulsa.edu; ram-mohan@utulsa.edu; ovadia-shoham@utulsa.edu

G. Kouba
Chevron Energy Technology Company (Retired), Houston, TX, USA
e-mail: genekouba1@gmail.com

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_19

experimental study to investigate droplet deposition and coalescence in straight and curved pipes under annular flow. No mechanistic modeling was attempted. A model for predicting droplet size distribution in annular flow was proposed by Pereyra [Pe11]. The model is based on log-normal distribution.

The aim of this study is to conduct experiments on droplet deposition and coalescence in curved pipes and to provide guidelines for piping layouts upstream of separators. Also, a model is developed for the prediction of droplet deposition and coalescence in curved pipes.

19.2 Experimental Program

19.2.1 Test Facility

A schematic of the test facility is shown in Figure 19.1. The test facility consists of four main sections. These include the storage and metering section, the flow loop, the measurement section, and the data acquisition system. The air and water are introduced at the inlet of the flow loop, forming a fully developed annular flow before entering the test sections. A wet gas separator is installed downstream of the test sections, where the air is vented to the atmosphere and the water is re-circulated.

19.2.1.1 Flow Loop

The flow loop is constructed of a Harvel® 2−inch clear schedule 80 pipe. It consists of three sections, namely a long run inlet section and two test sections. It also

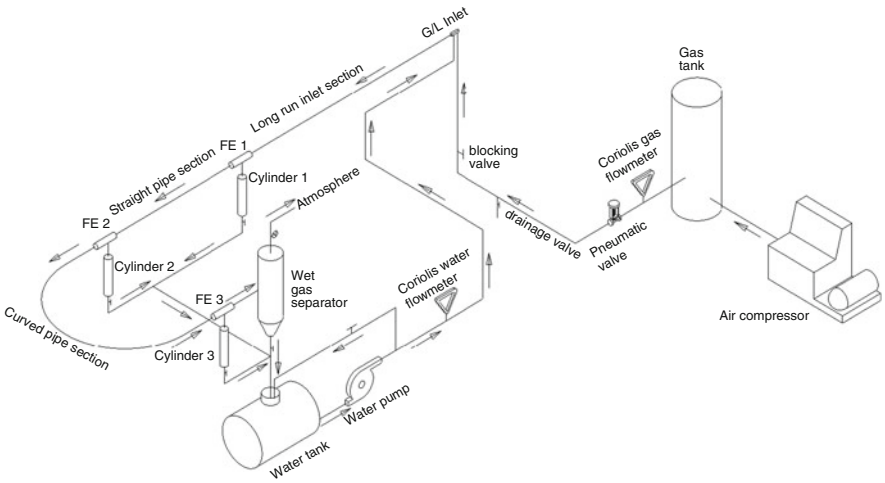


Fig. 19.1 Overall facility schematic

includes a wet gas cyclone separator at the end of the flow loop, and three liquid film extractors: FE1, FE2, and FE3, which are designed to remove the liquid film of the annular flow. The long run inlet section, from gas and liquid mixing tee at the flow loop inlet to FE1, is 12.3 m long (2-inch ID) which promotes fully developed annular flow before entering the test sections.

The two test sections include a straight pipe section (for generating baseline data), which extends from FE1 to FE2, and the 180° return curved pipe bend section between FE2 and FE3. The former is horizontally fixed to the flow loop, whereas the latter is interchangeable. These two sections have the same length of 2.215 m, enabling comparison of the droplet deposition phenomenon in the straight pipe and in the curved pipe.

Different curved bends can be installed including a standard short elbow bend, long elbow bend, 180° pipe bend, cushion tee bend, and impact tee bend. In the long elbow bend, the radius of curvature is six times the diameter of the pipe (6D). The curved pipe bends can be inclined downward to −10°.

19.2.1.2 Film Extractor

A schematic of the film extractor is shown in Figure 19.2. It consists of an incoming fixed pipe (#1) and outgoing pipe, which is movable (#2). The gap between the incoming and outgoing pipes can be adjusted between 0 and 10 cm by moving the latter. When the gap is 0 cm, the FE is fully closed, while for data acquisition, the gap is kept at 1.75D ($\tilde{8}$ cm) ensuring that the liquid film is completely extracted.

The liquid film flow rate measurement section consists of three cyclonic cylinders located downstream of the respective liquid film extractors. The removed film accumulates in the respective cylinder, enabling measurement of the liquid film flow rate.

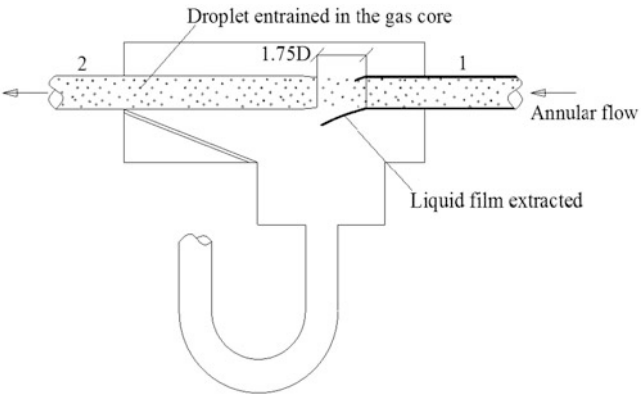


Fig. 19.2 Schematic of Film Extractor in Operation

Table 19.1 Test Matrix

Liquid Loading ($m^3/MMsm^3$)	v_{SL} (cm/s)	v_{SG} (m/s)
1400	2.8	20
1400	3.5	25
1400	4.2	30
1400	4.9	35
1400	5.6	40
1400	6.3	45
700	1.4	20
700	1.75	25
700	2.1	30
700	2.45	35
700	2.8	40
700	3.15	45

19.2.1.3 Test Matrix

The test matrix is presented in Table 19.1. As can be seen in the table, six different superficial gas velocities, v_{SG}, of 20, 25, 30, 35, 40, and 45 m/s are used. Two liquid loadings (LL) are utilized, namely 700 and 1400 $m^3/MMsm^3$ (cubic meter per million standard cubic meter), resulting in corresponding superficial liquid velocities, v_{SL}, as given in Table 19.1, resulting in a total of 12 operational points (pairs of v_{SG} and v_{SL}).

Experimental runs are conducted for 5 different curved pipe sections at 3 angles of 0°, 5°, and 10° and 1 straight pipe section only at 0°, leading to 16 ($5 \times 3 + 1$) different geometrical configurations. Each configuration is run with the 12 operational points, resulting in a total of 192 (12×16) test runs.

Droplet deposition is investigated both in the straight pipe and the curved pipe, which have the same length, under the same flow conditions making it possible to compare the deposition rates in both sections.

19.2.2 Experimental Results

19.2.2.1 Measurement of Droplet Deposition in Straight Pipe

For this case, FE3 is closed, FE1 and FE2 are open (refer to Figure 19.3). The liquid film is removed by FE1, providing the liquid film flow rate, q_1. The droplets continue flowing into the test section with the gas core and deposit on the pipe wall and form a new liquid film. FE2 removes the new liquid film, designated by q_2, which is the droplet deposition flow rate in the straight pipe. The surviving droplets exit the test section and flow with the gas core straight into the separator. The droplet deposition

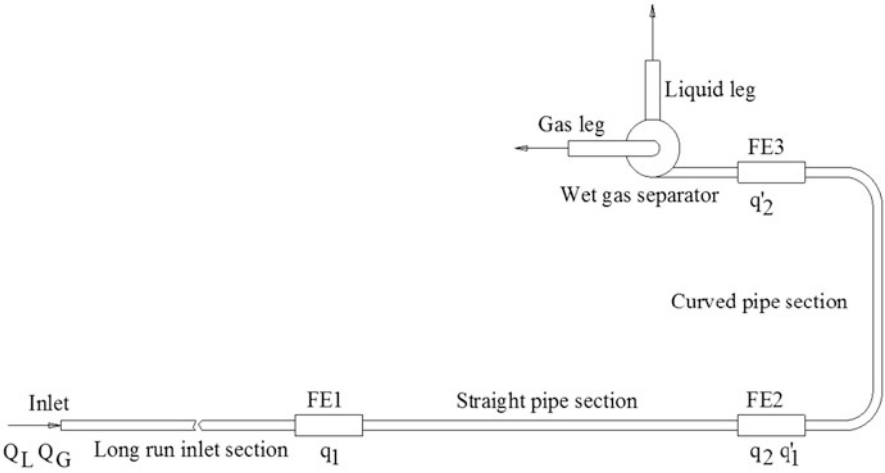


Fig. 19.3 Schematic of Flow Loop Experimental and Setup

percentage (DDP) in the straight pipe is defined as the ratio of the droplet deposition rate in straight pipe to the total droplet flow rate, as given by

$$f_D = \frac{q_2}{Q_L - q_1}.100\% \tag{19.1}$$

19.2.2.2 Measurement of Droplet Deposition Percentage in Curved Pipe Section

Following Figure 19.3, for this case, FE1 is closed, FE2 and FE3 are open. FE2 removes the liquid film from which the liquid film flow rate, $q_1{}'$, can be determined. Thus, only droplets flow into the curved pipe section, deposit on the pipe wall and form a new liquid film. FE3 removes the new liquid film, which provides the droplet deposition flow rate in the curved pipe, $q_2{}'$. The DDP in the curved pipe is defined

$$f_D{}' = \frac{q_2{}'}{Q_L - q_1{}'}.100\% \tag{19.2}$$

19.2.2.3 Experimental Results

Droplet Deposition Percentage (DDP) data are collected for LL of 700 $\left(\mathrm{m}^3/\mathrm{MMsm}^3\right)$ and 1400 $\left(\mathrm{m}^3/\mathrm{MMsm}^3\right)$, for the straight pipe section and all the curved pipe sections. Please refer to Nguyen [Ng15] and Nguyen et al. [Ng14] for the results of all the experimental runs. Presented next are typical results and summaries of all data for the two different liquid loading runs.

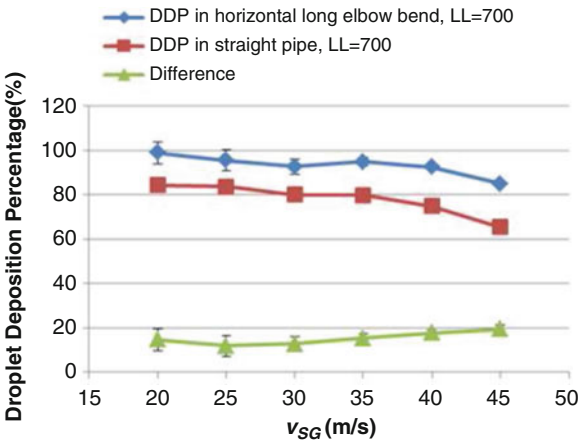


Fig. 19.4 Horizontal Long Elbow Bend, LL = 700

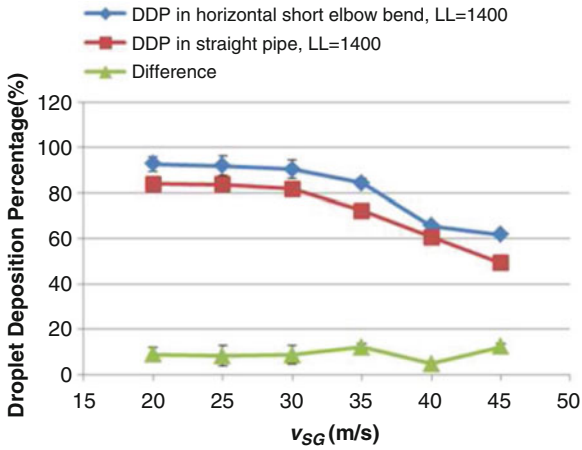


Fig. 19.5 Horizontal Long Elbow Bend, LL = 1400

Figures 19.4 and 19.5 present the results for the long elbow bend for LL of 700 and 1400 $m^3/MMsm^3$, respectively. Figure 19.4 shows a comparison between DDP in the horizontal long elbow bend and the straight pipe section. As can be seen in the figure, the DDP for the long elbow bend is about 20% higher than that of the straight pipe section. Similar results are presented in Figure 19.5 for LL of 1400 $(m^3/MMsm^3)$.

Similar data were acquired in the case of the short elbow bend for LL of 700 $(m^3/MMsm^3)$ and 1400 $(m^3/MMsm^3)$. When compared to the long elbow bend, it can be observed that for the short elbow bend, the DDP improvement over the straight pipe section is only around 10%, as compared to the long elbow bend results of 20% improvement. Also, for the short elbow bend, at superficial gas velocity higher than 35 m/s, the DDP improvement reduces and completely diminishes at the highest velocities.

A dimensionless number can be utilized for scale up and data presentation. A velocity ratio is selected to account for changes in gas density and kinetic energy, which is defined by:

$$v_{ratio} = \frac{v_{SG}}{v_{ann}} \tag{19.3}$$

where v_{ann} is the gas velocity at the onset to liquid carryover, given by

$$v_{ann} = 2.3351 \left[\sigma We \frac{\rho_L - \rho_G}{\rho_G^2} \right]^{0.25} \tag{19.4}$$

The velocity ratio given in Equation (19.3) will be used to present a comprehensive comparison of the results obtained for all the curved pipe bends, as given next.

A comprehensive performance comparison among all the curved bends used in the horizontal configuration is presented in Figures 19.6 and 19.7. Including are the long elbow bend, short elbow bend, 180° pipe bend, cushion tee bend, impact tee bend, and straight pipe section results for $LL = 700$ and 1400 $m^3/MMsm^3$, respectively.

As can be seen in Figure 19.6, all the curved pipe bends have better DDP than that of the straight pipe for the entire range of v_{ratio} between 2.1 to 4.5 for $LL = 700$ $m^3/MMsm^3$. Among all the curved pipe bends, the short elbow bend and cushion tee bend have the lowest performance, whereas for the 180° pipe bend and

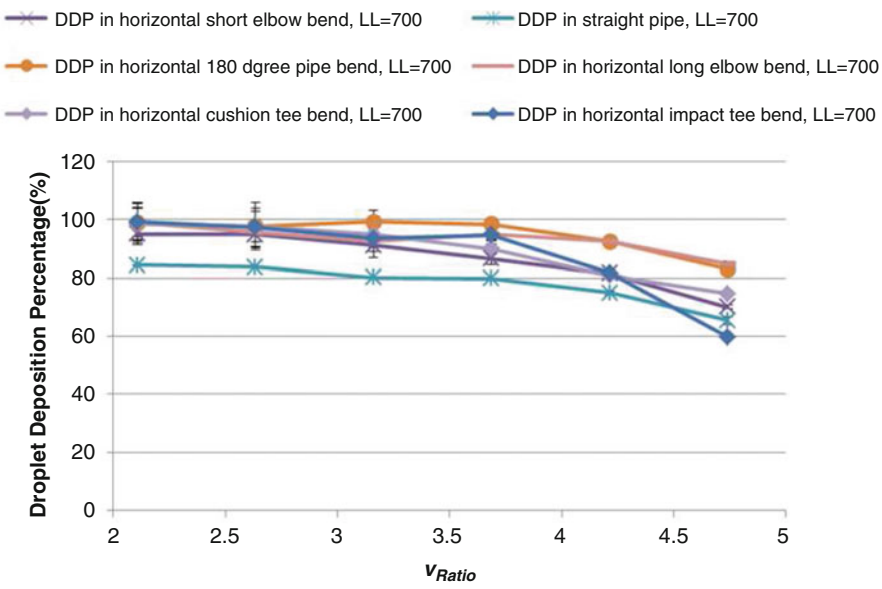


Fig. 19.6 Comprehensive Comparison of Results, LL = 700

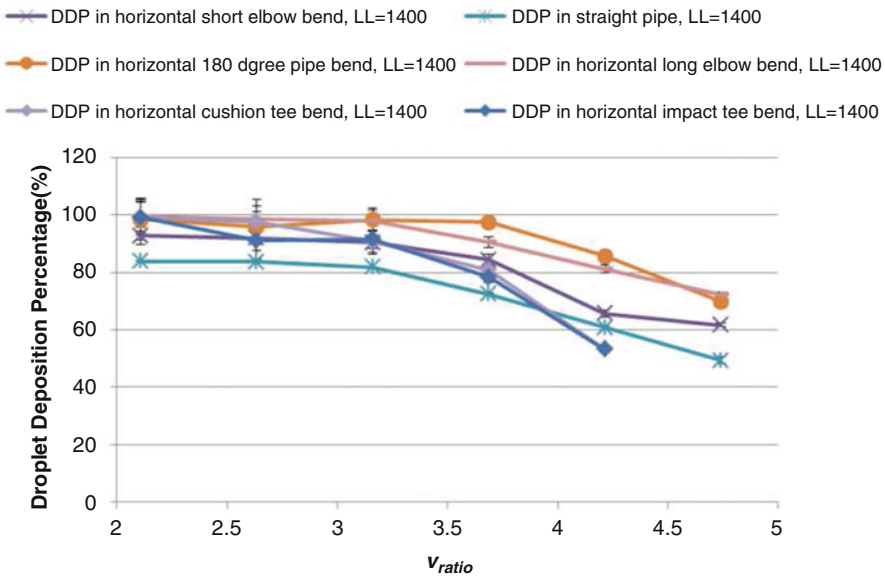


Fig. 19.7 Comprehensive Comparison of Results, LL = 1400

long elbow bend the performance is the highest. Also note that the performances of both the 180° and the long elbow bends are similar. The impact tee bend shows an intermediate performance, but it falls below the straight pipe section performance for $v_{ratio} > 4.5$.

Similarly for liquid loading of 1400 $m^3/MMsm^3$ (see Figure 19.7, all the curved pipe bends have better DDP than that of the straight pipe as long as v_{ratio} in between 2.1 to 3.9. Among all the curved pipe bends, the 180° pipe bend and long elbow bend performances have the highest and similar performance while the impact tee bend and cushion tee bend have the lowest and they even falls below the straight pipe as $v_{ratio} > 3.9$.

19.3 Modeling and Results

19.3.1 Physical Model

The physical model is presented schematically in Figure 19.8. The droplet is subjected to several forces, which are indicated in the free body diagram. The forces are the gravity (W), the buoyant force (F_B), and the drag force (F_D).

Carrying out a force balance on a droplet, the force vertical component and horizontal component are given, respectively, by

$$F_V = (\rho_G - \rho_L)\frac{\pi d_d^3}{6}g \tag{19.5}$$

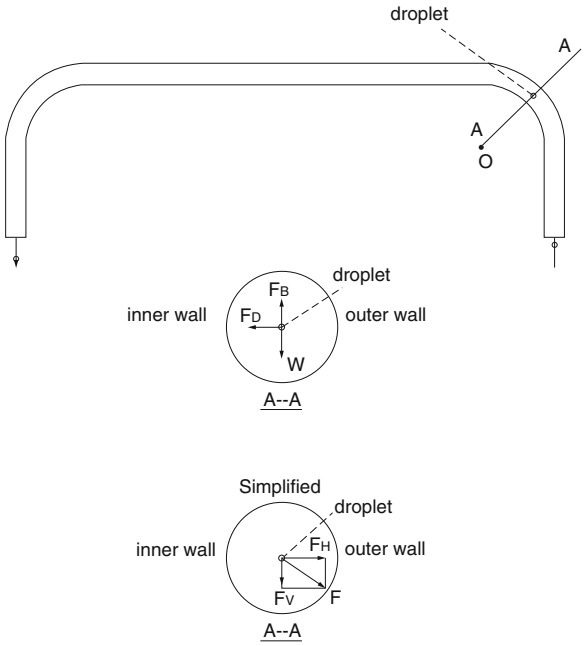


Fig. 19.8 Physical Model Schematic

and

$$F_H = F_D = \frac{1}{2} C_D A_p \rho_G v_s^2 \tag{19.6}$$

In Equations (19.5), (19.6) ρ_L is the droplet density, ρ_G is the gas density, g is the gravitational acceleration, v_s is the slip velocity between gas and droplet, C_D is the drag coefficient, V_d is the droplet volume, d_d is the droplet diameter, A_p is the droplet projected area, v_θ is the droplet axial velocity.

The slip velocity, v_s, can be determined by equating the drag force (Equation 19.6) to the centripetal force yielding

$$v_s = \sqrt{\frac{\rho_L}{\rho_G} \frac{\pi d_d^3}{3 C_D A_p} \frac{v_\theta^2}{R}} \tag{19.7}$$

where v_θ is the particle tangential velocity.

19.3.2 Conservation of Angular Momentum

The conservation of angular momentum of a droplet about O (refer to Figure 19.9) which is the center of the curvature of the bend yields

$$\vec{H}_O = \vec{R} \times m\vec{v} = \vec{R} \times m\left(\vec{v}_r + \vec{v}_\theta\right) = \vec{R} \times m\vec{v}_r + \vec{R} \times m\vec{v}_\theta = \vec{R} \times m\vec{v}_\theta = \text{constant} \tag{19.8}$$

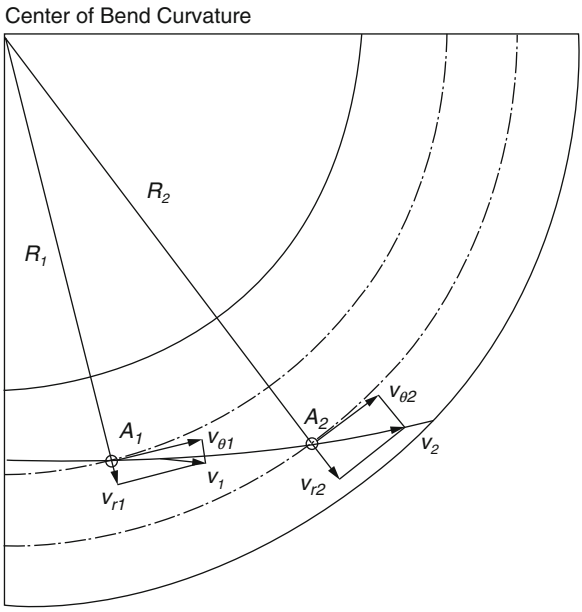


Fig. 19.9 Conservation of Angular Momentum

Note that in Equation (19.8) $\vec{R} \times m\vec{v_r} = \vec{0}$. Thus, $H_O = R_1 m v_{\theta 1} = R_2 m v_{\theta 2}$ and the relationship between the tangential velocities at A_2 and A_1 is

$$v_{\theta 2} = \frac{R_1}{R_2} v_{\theta 1} \tag{19.9}$$

19.3.3 Droplet Size Distribution

The droplet distribution at the bend inlet is predicted by Pereyra [Pe11] for annular flow. The maximum droplet size is

$$d_{max} = We_{CRIT} \frac{\left[\sigma + 2^{-\frac{3}{2}} \mu_L \left(\varepsilon_0 d_{max}\right)^{\frac{1}{3}}\right]^{\frac{3}{5}}}{\rho_L^{\frac{1}{5}} \rho_G^{\frac{2}{5}}} \varepsilon_0^{-\frac{2}{5}} \tag{19.10}$$

For the droplet size distribution, expressions for S_V and M_V, are provided, namely

$$S_V = -1.645 + \sqrt{1.645^2 - 2\ln(0.5)} = 0.378 \tag{19.11}$$

and

$$M_V = \ln(d_{max}) - 0.622 \tag{19.12}$$

Solving for S_v and M_v and substituting into the log-normal volumetric distribution

$$fv(d) = \frac{1}{dS_v\sqrt{2\pi}} \exp\left(-\frac{(\ln(d) - M_v)^2}{2S_v^2}\right) \tag{19.13}$$

Equation (19.13) is the probability density function, which gives the droplet size distribution in annular flow.

19.3.4 Droplet Deposition Criterion

In the droplet deposition process, it is assumed that all droplets are distributed evenly over the pipe cross section. When the droplets entrained by gas enter the elbow, part of them survive the bend exiting the bend with the gas core, while the rest of the droplets hit the wall and deposit into the liquid film. Criteria are developed for determination of whether or not a droplet deposits or survives. Figure 19.10 shows schematically the movement of a droplet across the elbow cross section in a period of time, Δt, from t_1 to t_2, $\Delta t = t_2 - t_1$.

The distance r_i from a droplet at a point $A(x_i, y_i)$ to the center O is given by

$$r_i = \sqrt{x_i^2 + y_i^2} \tag{19.14}$$

When the droplet moves from A_1 to A_2, it would hit the wall and deposit if the condition below is met

$$r_2 = \sqrt{x_2^2 + y_2^2} \geq \frac{D-d}{2} \tag{19.15}$$

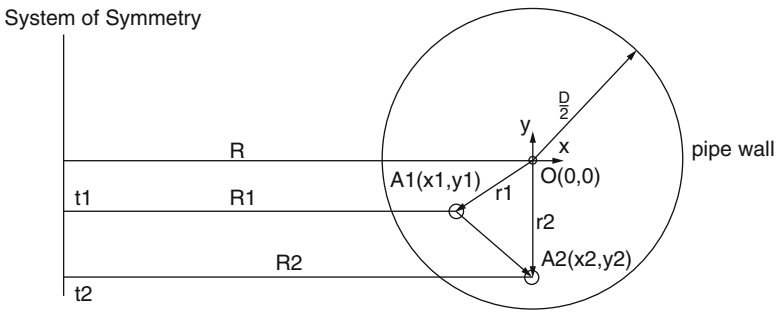


Fig. 19.10 Droplet movement in the cross section

where D is the pipe diameter, r_2 is the distance from A_2 to O, and d is droplet diameter.

A droplet moving in the elbow is assumed to be entrained by gas at an axial velocity of $v_\theta = v_{SG}$. If the droplet does not hit the elbow wall, the residence time it takes to go through the elbow is defined as follows:

$$t_{Res} = \frac{\pi R}{2v_\theta} = \frac{\pi R}{2v_{SG}} \tag{19.16}$$

The deposition criterion in Equation (19.15) is checked as long as the cumulative droplet time in the elbow is less than the residence time.

19.3.5 Results and Discussion

This section presents a comparison between the developed model predictions and the experimental data. Figure 19.11 presents a comparison between model predictions and experimental data for droplet deposition at liquid loading of 700 $m^3/MMsm^3$ ($LL = 700$). As can be seen, a fair agreement is observed, whereby the model consistently underpredicts the experimental data by 20%.

Similarly, a comparison between model predictions and experimental data for liquid loading of 1400 $m^3/MMsm^3$ ($LL = 1400$) is shown in Figure 19.12 . For this case, too, a good agreement occurs between the data and model predictions. Again, a fair agreement is observed exhibiting a consistent underprediction of 20%.

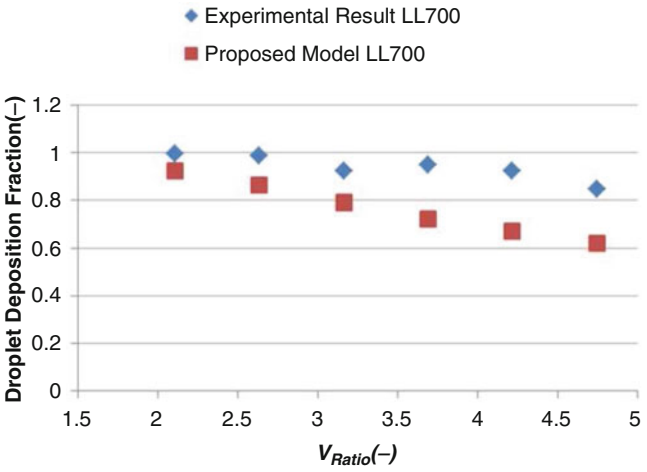


Fig. 19.11 Comparison between Model Predictions and Experimental Data for LL = 700

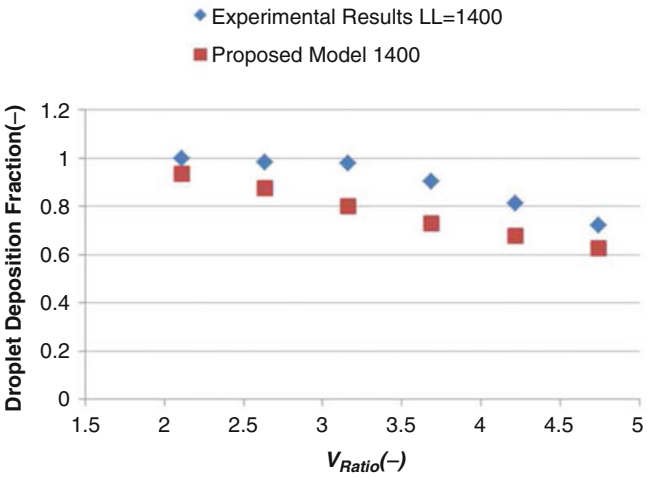


Fig. 19.12 Comparison between Model Predictions and Experimental Data for LL = 1400

Acknowledgements The authors thank the Tulsa University Separation Technology Projects (TUSTP) and The University of Tulsa where the project was conducted; and Chevron Energy Technology Company (CETC) and the Tulsa University Center of Research Excellence (TU-CoRE) for the financial support

References

[Ba61] Banerjee, S., Rhodes, E., Scott, D.S.: Film inversion of co-current two-phase film in helical coils. AIChE J. **13**, 189–191 (1961)

[Fe09] Feng, X.: A study of droplet deposition in curved pipes. M.S. Thesis, The University of Tulsa (2009)

[Ng14] Nguyen, H., Wang, S., Mohan, R.S., Shoham, O., Kouba, G.: Experimental investigations of droplet deposition and coalescence in curved pipes. ASME J. Energ. Resour. Technol. **136**, 22902 (2014)

[Ng15] Nguyen, H.: Droplet deposition and coalescence in curved pipes. Ph.D. Dissertation, The University of Tulsa (2015)

[Pe11] Pereyra, E.: Modeling of integrated compact multiphase separation system (CMSS©). Ph.D. Dissertation, The University of Tulsa (2011)

[ShBr87] Shoham, O., Brill, J.P.: Two-phase flow splitting in a tee junction - experiment and modeling. Chem. Eng. Sci. **42**, 2667–2676 (1987)

Chapter 20
Shifting Strategy in the Spectral Analysis for the Spectral Green's Function Nodal Method for Slab-Geometry Adjoint Transport Problems in the Discrete Ordinates Formulation

J.P. Curbelo, O.P. da Silva, C.R. García, and R.C. Barros

20.1 Introduction

It is well known that the adjoint angular flux, i.e., the solution of the equation which is adjoint to the Boltzmann transport equation, can be viewed as a measure of the importance of a particle to the objective function, e.g., a detector response [BeGl70, PrLa10]. This physical interpretation makes the adjoint angular flux well suited for use as an importance function in source-detector problems.

Reference [MiEtAl12] describes and tests a spectral nodal method for monoenergetic slab-geometry adjoint problems in the discrete ordinates (S_N) formulation with isotropic scattering and a prescribed interior adjoint source. That method is based on the standard spatially discretized S_N balance adjoint equations and a nonstandard adjoint auxiliary equation expressing the adjoint node-average angular flux, in each discretization node, as a weighted combination of the adjoint node-edge outgoing fluxes. The weights in the auxiliary equation act as Green's functions for the adjoint node-average angular fluxes and they are determined by a spectral analysis to yield the local general solution of the S_N equations within each node of the discretization grid; therefore, that method, that we refer to as the adjoint spectral Green's function (Adjoint-SGF) method, converges numerical solutions that are completely free from spatial truncation errors.

In this chapter, recent advances in the Adjoint-SGF method are presented. The method is extended to S_N problems considering arbitrary $L'th$ order of scattering anisotropy, provided $L < N$, and non-zero prescribed boundary conditions for the

J.P. Curbelo (✉) • O.P. da Silva • R.C. Barros
Polytechnic Institute, State University of Rio de Janeiro, Nova Friburgo, RJ, Brazil
e-mail: jcurbelo86@gmail.com; odairpds@gmail.com; rcbarros@pq.cnpq.br

C.R. García
Institute of Technology and Applied Sciences, La Habana, Cuba
e-mail: cgh@instec.cu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_20

forward S_N transport problem. In addition, we present the positive features in the shifting strategy that we use in the homogeneous component of the general solution of the monoenergetic, slab-geometry, adjoint S_N equations inside each discretization node for neutral particle source-detector transport problems. The shifting strategy scales the N exponential functions of the local solution in the interval (0, 1). One advantage is to avoid the overflow in computational finite arithmetic calculations in high-order angular quadrature and/or coarse-mesh calculations.

20.2 The Adjoint S_N Problem

First we consider a discretization grid on the slab of thickness H, as represented in Figure 20.1. Each discretization node Υ_j has width h_j and constant material parameters. Now we write the equations which are adjoint to the monoenergetic, slab-geometry S_N transport equations in Υ_j with anisotropic scattering

$$-\mu_m \frac{d}{dx}\psi_m^\dagger(x)+\sigma_{T_j}\,\psi_m^\dagger(x) = \sum_{n=1}^{N} \omega_n\,\psi_n^\dagger(x) \sum_{l=0}^{L} \frac{2l+1}{2}\,\sigma_{S_j}^{(l)}\,P_l(\mu_m)\,P_l(\mu_n) + Q_j^\dagger\,,$$

$$x_{j-1/2} < x < x_{j+1/2}\,, \qquad (20.1)$$

with boundary conditions

$$\psi_m^\dagger(0) = 0, \quad \mu_m < 0 \qquad \text{and} \qquad \psi_m^\dagger(H) = 0, \quad \mu_m > 0\,.$$

In Equation (20.1) the angular quadrature is defined by the set $\{\mu_m, \omega_m, m = 1 : N\}$. The values of μ_m represent the discrete directions and ω_m are the weights of the angular quadrature. In this chapter we use even-order sets of Gauss-Legendre quadratures [LeMi93].

The notation used in Equation (20.1) is standard [LeMi93]: $\psi_m^\dagger(x)$ is the adjoint angular flux in the direction μ_m, σ_T is the total macroscopic cross section, and $\sigma_S^{(l)}$ is the $l'th$ order component of the scattering macroscopic cross section. The quantity $Q_j^\dagger$ is the adjoint interior source, which is perfectly arbitrary [DuMa79].

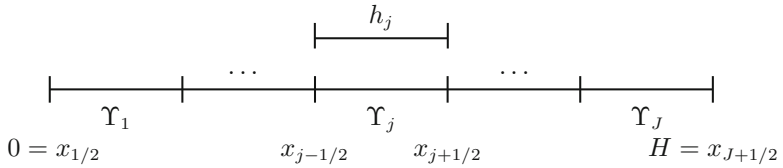


Fig. 20.1 Spatial grid on slab $D : 0 \le x \le H$

20.2.1 Detector Response for Adjoint Problems

According to [LeMi93] the detector response for monoenergetic problems can be obtained by

$$R = \langle \psi^{\dagger}, Q \rangle - \int_{\Gamma} d\Gamma \int_{4\pi} d\Omega \;\; \widehat{n} \circ \widehat{\Omega} \;\; \psi^{\dagger} \, \psi \, , \tag{20.2}$$

where Γ is the contour surface of volume V, $d\Omega$ is the differential surface element of the unit sphere and we have defined the integral operation

$$\langle \cdot, \cdot \rangle = \int_{V} dV \int_{4\pi} d\Omega \, .$$

Assuming prescribed boundary conditions for the forward problem and remarking that we have considered, for the adjoint problem, boundary conditions of no outgoing adjoint flux, in one-speed, slab-geometry S_N models, Equation (20.2) can be written as

$$R = \langle \psi^{\dagger}, Q \rangle \; + \sum_{n=1}^{N/2} \mu_n \, \omega_n \, \psi_n^{\dagger}(0) \, \widetilde{\psi}_0 \; + \sum_{n=N/2+1}^{N} \mu_n \, \omega_n \, \psi_n^{\dagger}(H) \, \widetilde{\psi}_H \, , \tag{20.3}$$

where magnitudes $\widetilde{\psi}_0$ and $\widetilde{\psi}_H$ are the forward flux values considering prescribed isotropic boundary conditions at $x = 0$ and $x = H$, respectively.

20.3 Spectral Analysis

The general solution of the system of N ordinary differential equations shown in Equation (20.1) can be written as

$$\psi_m^{\dagger}(x) = \psi_{m\,,j}^{\dagger\,P} + \psi_m^{\dagger H}(x) \, ,$$

where $\psi_{m\,,j}^{\dagger\,P}$ is a particular solution and $\psi_m^{\dagger H}(x)$ is the homogeneous component of the local general solution of Equation (20.1). Substituting the spatially constant $\psi_{m\,,j}^{\dagger\,P}$ into Equation (20.1) we obtain

$$\psi_{m\,,j}^{\dagger\,P} = \frac{Q_j^{\dagger}}{\sigma_{T_j} - \sigma_{S_j}^{(0)}} \, . \tag{20.4}$$

To determine the homogeneous component we consider the expression

$$\psi_m^{\dagger H}(x) = a_m^{\dagger}(\xi) \, e^{\dfrac{-(x-\lambda_j)}{\xi}} \quad , \quad \lambda_j = \begin{cases} x_{j+1/2} \;\; , \xi < 0 \\ \\ x_{j-1/2} \;\; , \xi > 0 \end{cases} \, , \; x \in \Upsilon_j \, . \tag{20.5}$$

Here we note that the shifting strategy used in the exponential term of Equation (20.5) bounds the N exponential functions of the local solution in the interval (0, 1) and was proposed by [Pi04]. In the next sections we give more details of the advantages of this strategy.

Substituting Equation (20.5) into the homogeneous equation corresponding to (20.1), i.e., $Q_j^{\dagger} = 0$, after some algebraic manipulations, we obtain

$$\sum_{n=1}^{N} \frac{\sigma_{Tj}}{\mu_m} \left\{ -\delta_{m,n} + \omega_n \sum_{l=0}^{L} \frac{2l+1}{2} c_{S_j}^{(l)} P_l(\mu_m) P_l(\mu_n) \right\} a_n^{\dagger}(\xi) = \frac{1}{\xi} a_m^{\dagger}(\xi) , \qquad (20.6)$$

where $\delta_{m,n}$ is the *Kronecker delta* and $c_{S_j}^{(l)} \equiv \dfrac{\sigma_{S_j}^{(l)}}{\sigma_{T_j}}$ is the anisotropic scattering ratio of order l. For $m = 1 : N$, Equation (20.6) represents an eigenvalue problem. In case $0 < c_{S_j}^{(l)} < \ldots < c_{S_j}^{(0)} < 1$, we obtain N real distinct eigenvalues which are symmetric about the origin. Therefore, for $x \in \Upsilon_j$ we obtain a linearly independent set of N eigenfunctions defined in Equation (20.5) and we write the general solution for Equation (20.1) in node Υ_j as

$$\psi_m^{\dagger}(x) = \sum_{k=1}^{N} \beta_k \, a_m^{\dagger}(\xi_k) \, e^{\frac{-(x-\lambda_j)}{\xi_k}} + \psi_{m,j}^{\dagger P} , \qquad (20.7)$$

where $a_m^{\dagger}(\xi_k)$ is the $m'th$ component of the eigenvector corresponding to eigenvalue ξ_k^{-1}; β_k are arbitrary constants, and $\psi_{m,j}^{\dagger P}$ is calculated by Equation (20.4). We remark here that for heterogeneous slabs, it is necessary to set such a general solution for each region of the domain.

20.4 The Adjoint Spectral Green's Function Method (Adjoint-SGF)

Integrating Equation (20.1) within an arbitrary spatial node Υ_j by using the operator

$$\frac{1}{h_j} \int_{x_{j-1/2}}^{x_{j+1/2}} (\cdot) \, dx ,$$

we obtain the discretized spatial balance S_N adjoint equations

$$-\frac{\mu_m}{\sigma_{T_j} h_j} \left(\psi_{m,j+1/2}^{\dagger} - \psi_{m,j-1/2}^{\dagger} \right) + \overline{\psi}_{m,j}^{\dagger} = S_{m,j}^{\dagger} + \frac{Q_j^{\dagger}}{\sigma_{T_j}} , \quad m = 1 : N , \qquad (20.8)$$

where we have defined the node-average adjoint angular flux in node Υ_j

$$\overline{\psi}^{\dagger}_{m,j} \equiv \frac{1}{h_j} \int_{x_{j-1/2}}^{x_{j+1/2}} \psi^{\dagger}_{m}(x) dx \,,$$

the node-edge adjoint angular flux

$$\psi^{\dagger}_{m,j\pm 1/2} \equiv \psi^{\dagger}_{m}(x_{j\pm 1/2})$$

and the term which is adjoint to the anisotropic scattering source, that we term the adjoint anisotropic scattering source

$$S^{\dagger}_{m,j} \equiv \sum_{n=1}^{N} \omega_n \, \overline{\psi}^{\dagger}_{n,j} \left[\sum_{l=0}^{L} \frac{2l+1}{2} \, c^{(l)}_{S_j} \, P_l(\mu_m) \, P_l(\mu_n) \right].$$

Equation (20.8), within an arbitrary spatial node Υ_j, represents a system of N algebraic linear equations into $3N$ unknowns. In order to guarantee uniqueness of the system solution, we need to use auxiliary equations. In the Adjoint-SGF method we use an auxiliary equation, which relates the node-average adjoint angular flux to the outgoing node-edge adjoint fluxes. This auxiliary equation has the form

$$\overline{\psi}^{\dagger}_{m,j} = \sum_{\mu_n<0} \Lambda^{j}_{m,n} \psi^{\dagger}_{n,j-1/2} + \sum_{\mu_n>0} \Lambda^{j}_{m,n} \psi^{\dagger}_{n,j+1/2} + B_m(Q^{\dagger}_j) \,, \tag{20.9}$$

where $\Lambda^{j}_{m,n}$ plays the role of the *Green's function* of the adjoint S_N operator discretized in space, and $B_m(Q^{\dagger}_j)$ is a function of the interior adjoint source to be determined such that the particular solution is automatically preserved. The quantities involved in the auxiliary equation (20.9) are illustrated in Figure 20.2 for an arbitrary node Υ_j.

To determine the term $B_m(Q^{\dagger}_j)$, we substitute Equation (20.4) into Equation (20.9) and the result appears as

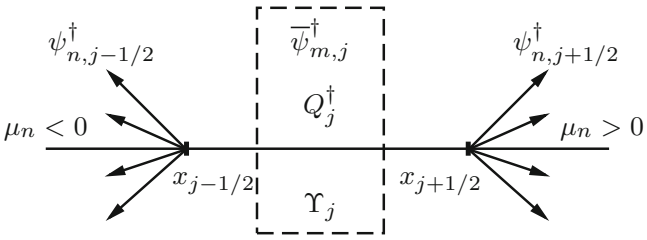


Fig. 20.2 Arbitrary discretization node (Υ_j) with outgoing node-edge adjoint fluxes ($\psi^{\dagger}_{n,j\pm 1/2}$), node-average adjoint angular flux ($\overline{\psi}^{\dagger}_{m,j}$) in direction μ_m, and interior adjoint source ($Q^{\dagger}_j$)

$$B_m(Q_j^\dagger) = \left(1 - \sum_{n=1}^{N} \Lambda_{m,n}^j\right) \frac{Q_j^\dagger}{\sigma_{T_j} - \sigma_{S_j}^{(0)}}\,.$$

To proceed further, we determine the parameters $\Lambda_{m,n}^j$ by requiring that the homogeneous component of the general solution be preserved by using Equation (20.5) in Equation (20.9) and, after some algebraic manipulations, we obtain the following linear systems:

$$\frac{\xi_k\, a_m^\dagger(\xi_k)}{h_j}\left(1 - e^{-\frac{h_j}{\xi_k}}\right) = \sum_{\mu_n<0} a_n^\dagger(\xi_k)\, \Lambda_{m,n}^j + e^{-\frac{h_j}{\xi_k}} \sum_{\mu_n>0} a_n^\dagger(\xi_k)\, \Lambda_{m,n}^j\,,\ (\xi_k > 0) \tag{20.10a}$$

and

$$\frac{|\xi_k|\, a_m^\dagger(\xi_k)}{h_j}\left(1 - e^{-\frac{h_j}{|\xi_k|}}\right) = e^{-\frac{h_j}{|\xi_k|}} \sum_{\mu_n<0} a_n^\dagger\, \Lambda_{m,n}^j + \sum_{\mu_n>0} a_n^\dagger\, \Lambda_{m,n}^j\,,\ (\xi_k < 0). \tag{20.10b}$$

Requiring this to hold for $m = 1 : N$, we obtain a linear system of N^2 equations in the N^2 unknowns $\Lambda_{m,n}^j$. Each entry $\Lambda_{m,n}^j$ represents the node-average adjoint angular flux in direction μ_m due to a unit outgoing node-edge adjoint flux in direction μ_n. It should be noted that in heterogeneous domains, one must have a Λ matrix for each region where the node thickness is constant.

We remark that with these choices, the exponential in Equations (20.10a) and (20.10b) are always decreasing functions, and hence, are restricted to the interval (0, 1). These convenient choices for the scaling parameters prevent from possible overflow when solving the system (20.10) on a digital computer that typically causes the computations to halt [Pi04, MeEtAl14]. The shifting strategy allows solving problems with high-order angular quadratures and/or coarse spatial grids.

In next section, we describe a sweeping algorithm for iteratively solving the Adjoint-SGF equations, which consists of three steps: first a sweep from left to right to calculate estimates for the adjoint angular fluxes in directions $\mu_m < 0$, then a sweep from right to left to calculate estimates for the adjoint angular fluxes in directions $\mu_m > 0$ and finally a check to see if the stopping criterion is satisfied.

20.5 The Partial One-Node Block Inversion Iterative Scheme

The iteration procedure for iteratively solving the Adjoint-SGF equations can be described as an adjoint one-node block inversion (NBI) scheme. Partial NBI scheme uses the most recent estimates for the node-edge adjoint angular fluxes outgoing a given discretization node (dashed arrows in Figure 20.3), to solve the resulting

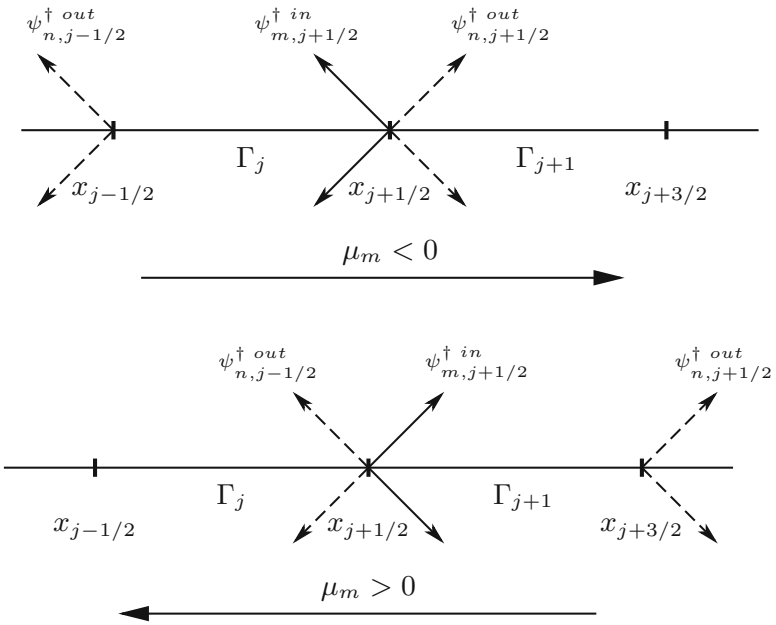


Fig. 20.3 Sweeping scheme for adjoint partial NBI algorithm

adjoint S_N problem in that node for all the incoming adjoint angular fluxes, which constitute the outgoing adjoint angular fluxes for the adjacent node in the sweeping direction (solid arrows in Figure 20.3).

In order to determine the sweeping equations, we first substitute the adjoint auxiliary equation (20.9) into the terms which contain the average adjoint angular fluxes in the spatially discretized adjoint S_N balance equations (20.8). Then we manipulate the scattering source terms so we obtain terms involving outgoing adjoint fluxes from both node edges, and we group the terms containing the node-interior adjoint source. At this point, we sweep from left to right to estimate the incoming adjoint fluxes on the right node-edge due to all outgoing adjoint fluxes and the interior adjoint source. Following this procedure, we obtain

$$\boldsymbol{\Psi}^{\dagger\, in}_{j+1/2} = \mathbf{G}^{\dagger\,+}_j \boldsymbol{\Psi}^{\dagger\, out}_{j-1/2} + \mathbf{G}^{\dagger\,-}_j \boldsymbol{\Psi}^{\dagger\, out}_{j+1/2} + \mathbf{F}^{\dagger}_j \,,$$

which is the sweeping equation for the adjoint partial NBI scheme from left to right, represented in matrix form. Following analogous procedure we obtain the sweeping equation in the opposite direction, i.e., from right to left, which appears as

$$\boldsymbol{\Psi}^{\dagger\, in}_{j-1/2} = \mathbf{G}^{\dagger\,+}_j \boldsymbol{\Psi}^{\dagger\, out}_{j+1/2} + \mathbf{G}^{\dagger\,-}_j \boldsymbol{\Psi}^{\dagger\, out}_{j-1/2} + \mathbf{F}^{\dagger}_j \,.$$

20.6 Numerical Examples

Let us consider a multilayer slab composed of seven regions and four different material zones, as illustrated in Figure 20.4. Total (σ_T) and scattering ($\sigma_S^{(l)}$) macroscopic cross sections for each material zone are displayed in Table 20.1. This model problem simulates the detection of neutrons by a detector D_1 ($\sigma_A = 0.1\, cm^{-1}$) located in the fourth region ($45 \leq x \leq 47$ cm) of the 100 cm slab, due to two neutron sources $Q_1 = 1$ and $Q_2 = 2$ located in the second region ($30 \leq x \leq 35$ cm) and in the sixth region ($57 \leq x \leq 60$ cm), respectively. To solve the adjoint problem, we set the adjoint source numerically equal to the detector absorption macroscopic cross section (σ_A), i.e., $Q^{\dagger} = 0.1$, as illustrated in Figure 20.5.

To model the forward transport problem, represented in Figure 20.4, and the adjoint transport problem, represented in Figure 20.5, we used the S_{32}, S_{64}, and S_{128} Gauss-Legendre angular quadrature set [LeMi93]. The stopping criterion for each run required that the discrete maximum norm of the relative deviation between two consecutive estimates for the node-average scalar fluxes (forward and adjoint) did not exceed 10^{-6}.

Table 20.2 displays the absorption rate density (absorption rate per unit cross section area) R_{Q1}, only due to the neutron source $Q_1 = 1$, the absorption rate density R_{Q2} only due to the source $Q_2 = 2$, the absorption rate density R_{bc} only due to the isotropic unit incident flux on the left boundary, and the total absorption rate density R due to both sources and incoming flux (Figure 20.4). To generate these results, we

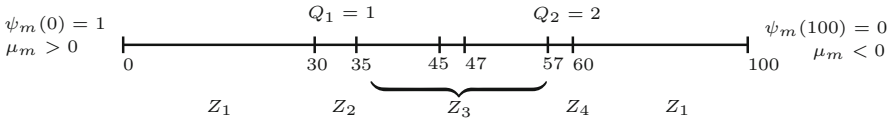


Fig. 20.4 Slab for the Model Problem (forward problem)

Table 20.1 Material parameters for the model problem

	Cross sections						
Material zones	$\sigma_T(cm^{-1})$	$\sigma_S^{(0)}(cm^{-1})$	$\sigma_S^{(1)}(cm^{-1})$	$\sigma_S^{(2)}(cm^{-1})$	$\sigma_S^{(3)}(cm^{-1})$	$\sigma_S^{(4)}(cm^{-1})$	$\sigma_S^{(5)}(cm^{-1})$
Zone 1	1.0	0.97	0.6	0.3	0.1	0.07	0.02
Zone 2	0.9	0.8	0.4	0.2	0.08	0.02	0.01
Zone 3	0.95	0.9	0.45	0.3	0.1	0.05	0.01
Zone 4	0.8	0.7	0.3	0.1	0.07	0.01	0.05

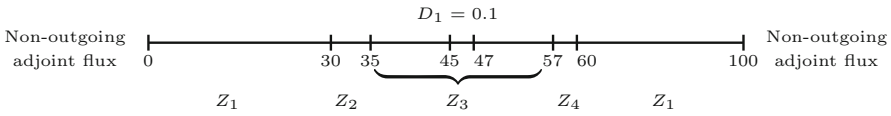


Fig. 20.5 Slab for the Model Problem (adjoint problem)

Table 20.2 Neutron detection for the Model Problem ($cm^{-2}s^{-1}$)

Forward problem				Adjoint problem			
	S_{32}	S_{64}	S_{128}		S_{32}	S_{64}	S_{128}
R_{Q1}	0.1036401	0.1036399	0.1036394	$R^{\dagger}_{Q1}$	0.1036400	0.1036398	0.1036393
R_{Q2}	0.1717053	0.1717046	0.1717036	$R^{\dagger}_{Q2}$	0.1717052	0.1717044	0.1717035
R_{bc}	8.244×10^{-6}	8.244×10^{-6}	8.244×10^{-6}	$R^{\dagger}_{bc}$	8.244×10^{-6}	8.244×10^{-6}	8.244×10^{-6}
R	0.2753536	0.2753527	0.2753513	$R^{\dagger}$	0.2753534	0.2753525	0.2753510

ran four distinct forward fixed-source problems using the coarse-mesh SGF method. On the other hand, the use of the adjoint technique to calculate the detector response is very convenient as it is possible to run the adjoint problem just once, provided we do not move the detector D_1 or replace it by a different one. Absorption rate densities $R^{\dagger}_{Q1}$, $R^{\dagger}_{Q2}$, $R^{\dagger}_{bc}$, and $R^{\dagger}$ are obtained, as defined in Equation (20.3), by running the adjoint problem with $Q^{\dagger} = 0.1$, just once.

Here we remark that using S_{64} and S_{128} Gauss-Legendre angular quadrature sets, for this model problem, required the application of the shifting strategy. This was due to the fact that for the seventh region ($h = 40\,cm$), the exponentials were out of range and occurred overflow errors in the computational calculations. Table 20.2 shows the results for the detector response obtained by running both the forward and adjoint methods. As we see, in all cases, the results generated with forward and adjoint techniques do agree, at least, up to the sixth decimal place.

20.7 Conclusions and Perspectives

In this work we have extended the SGF method to adjoint one-speed, slab-geometry S_N problems considering anisotropic scattering and non-zero prescribed boundary conditions for the forward problem. The shifting strategy is applied to the Adjoint-SGF method in order to avoid the overflow in computational finite arithmetic calculations in high-order angular quadrature and/or coarse-mesh calculations.

According to the model problem considered in this chapter, the numerical results for the detector response, as generated by the forward and the adjoint techniques, were identical up to the sixth decimal place. We note that the use of the adjoint technique to calculate the detector response is convenient as it is possible to run the adjoint problem just once for various interior source distributions and/or prescribed incident flux of particles, provided we do not change the location or the type of the detector.

A negative feature of the Adjoint-SGF method is that it requires more storage than standard discretization methods. The Adjoint-SGF requires the storage of as many matrices $\Lambda^{j}_{m,n}$ as sub-domains, and the iteration scheme requires the storage of the adjoint node-edge angular fluxes in all discrete ordinates directions. This extra storage requirement is compensated by the possibility to use coarse spatial meshes.

We intend to apply the present method to an arbitrary order L of scattering anisotropy in energy multigroup adjoint S_N problems to account for the energy transfer in scattering events. The present Adjoint-SGF method can be used to improve the accuracy of multidimensional adjoint S_N nodal methods, similarly to the steps followed previously for forward S_N problems [BaLa92]. However, this must await future work.

Acknowledgements The authors acknowledge the financial support of the project National Institute of Science and Technology on Innovative Nuclear Reactors, Brazil, for the ongoing development of this work. The work by Jesús Pérez Curbelo was supported by CAPES and FAPERJ.

References

[BaLa92] Barros, R.C., Larsen, E.W.: A spectral nodal method for one-group X, Y-geometry discrete ordinates problems. Nucl. Sci. Eng. **111**(1), 34–45 (1992)

[BeGl70] Bell, G.I., Glasstone, S.: Nuclear Reactor Theory. Van Nostrand Reinhold, New York (1970)

[DuMa79] Duderstadt, J.J., Martin, W.R.: Transport Theory. Wiley-Interscience, New York (1979)

[LeMi93] Lewis, E.E., Miller, W.F.: Computational Methods of Neutron Transport. American Nuclear Society, Illinois (1993)

[MeEtAl14] Menezes, W.A., Alves, H., Barros, R.C.: Spectral Green's function nodal method for multigroup S_N problems with anisotropic scattering in slab-geometry non-multiplying media. Ann. Nucl. Energ. **64**, 270–275 (2014)

[MiEtAl12] Militão, D.S., Alves, H., Barros, R.C.: A numerical method for monoenergetic slab-geometry fixed-source adjoint transport problems in the discrete ordinates formulation with no spatial truncation error. Int. J. Nucl. Energ. Sci. Technol. **7**, 151–165 (2012)

[Pi04] Pimenta de Abreu, M.: Mixed singular-regular boundary conditions in multislab radiation transport. J. Comput. Phys. **197**, 167–185 (2004)

[PrLa10] Prinja, A.K., Larsen, E.W.: In: Cacuci, D.G. (ed.) General Principles of Neutron Transport. Handbook of Nuclear Engineering, Cap 5. Springer Science+Business Media, New York (2010)

Chapter 21
A Metaheuristic Approach for an Optimized Design of a Silicon Carbide Operational Amplifier

M. Pourreza and S. Kargarrazi

21.1 Introduction

Emergence of wide bandgap semiconductor technologies such as Gallium Nitride (GaN) and Silicon Carbide (SiC) has enabled electronics to operate in extreme (e.g., high-temperature and high radiation) environments. Such electronics has been promoted in down-hole drilling, automobile, aerospace, and also space applications [CrMa12]. In particular, on high-temperature SiC electronics side, recent advances in the fabrication process [ThEtAl11, LaEtAl13, SpEtAl16] have paved the way for realization of integrated electronics that inherit the advantages that SiC material can provide. In the last decades, a multitude of integrated circuits have been demonstrated in SiC NMOS [VaEtAl14, XiEtAl94], CMOS [ChKo98, RyEtAl98, RaEtAl16, HaEtAl16], JFET [PaEtAl09, SpEtAl16], MESFET [AlEtAl15], and BJT [KEtAl15, KaEtAl16, KaEtAl15, KarEtAl15] technologies. Due to the fabrication process uncertainties and incomplete models for the devices in such technologies that can vary from batch-to-batch, wafer-to-wafer, and die-to-die, the integrated circuit design encounters many challenges.

Compared to mature technologies such as Silicon CMOS processes, the lack of process corner models and statistical data targeting the solution space limits the integrated circuit designer to optimize the circuits for various objectives. Considering constraints such as temperature range of operation which is 2–3x higher than Silicon ICs and analog design trade-offs (such as voltage gain and power consumption), a metaheuristic algorithm helps in finding a smart design solution in

M. Pourreza (✉)
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
e-mail: pourreza@ce.sharif.edu

S. Kargarrazi
The Royal Institute of Technology, Kista, Sweden
e-mail: salehk@kth.se

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_21

a reasonable time. These types of algorithms use special techniques to visit different areas of the problem's solution space and hence, can help the designer to find a near optimal solution in a reasonable amount of time. Previous works have been focused on using metaheuristics to optimize silicon analog integrated electronics, and especially CMOS circuits [DeTl15, JoEtAl15, CoEtAl07, DeEtAl10, LoEtAl02]. However, this study is the first attempt to use metaheuristics for designing in emerging technologies such as SiC ICs. It is worth mentioning that an operational amplifier is an essential building block for analog electronics in which gain and power consumption are the two most critical performance metrics. Therefore, this circuit in SiC technology is analyzed along with its important performance metrics throughout this paper. The optimization approach presented for this circuit can be highly beneficial for the designer who faces the mentioned challenges in SiC technology.

Thus, in this paper first the design of a SiC operational amplifier is introduced along with its influential parameters. Next, the optimization problem is elaborated formally and a Tabu Search algorithm is presented for finding a suboptimal solution for the mentioned problem. Finally, the results obtained from using this algorithm are substantiated and the conclusion is given in the last section.

21.2 Circuit Design

A high-temperature SiC operational amplifier (opamp) with two amplification stages has been previously demonstrated in [KaEtAl16, KaEtAl15], fabricated and tested in the temperature range of 25 °C - 500 °C. As previously mentioned, open-loop gain and power consumption are the two critical performance metrics in this circuit which need to be optimized. The open-loop gain is mainly determined by the amplification of the first and second stage, as illustrated by the simplified schematics of Figure 21.1 and can be approximated as:

$$A_{OL} = g_{m_1} \cdot [R_{C1} \| r_{\pi 2}] \cdot g_{m2} \cdot R_{C2} \tag{21.1}$$

where g_{m_1} and g_{m2} are the trans-conductances of the first and second stage, respectively.

Equation (21.1) can be expressed in terms of the semiconductor device parameters:

$$A_{OL} = \frac{\beta \cdot R_{C1} \cdot R_{C2} \cdot I_{C1}}{[R_{C1} + \beta V_T / I_{C2}] \cdot V_T}$$

where β is the forward current gain of the bipolar junction transistors (BJTs) and V_T is the thermal voltage which can be defined as:

$$V_T = \frac{kT}{q}$$

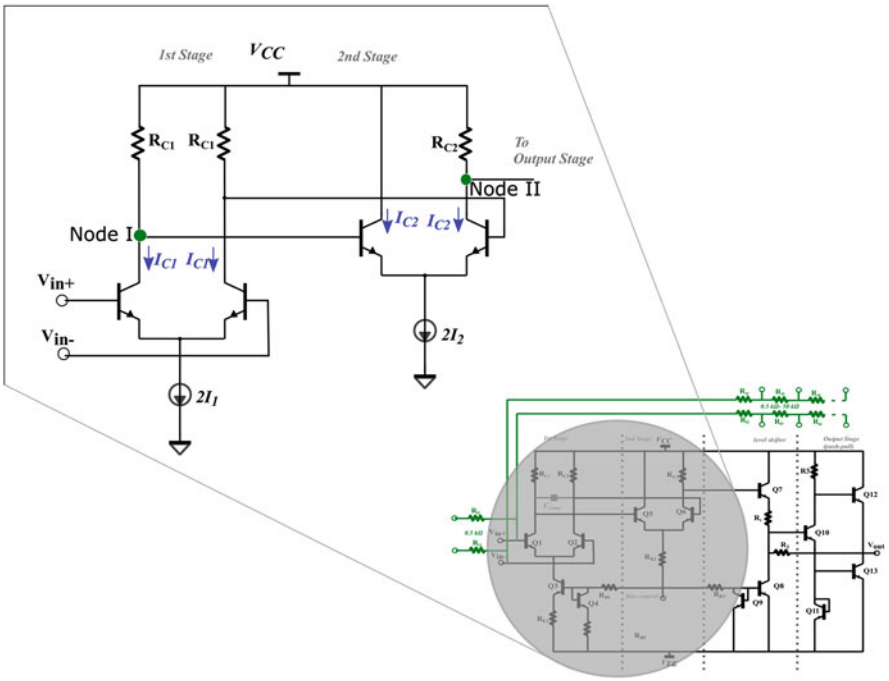


Fig. 21.1 The simplified opamp schematics highlighting the first and second amplification stages (top) and a complete SiC bipolar opamp schematic, reproduced from [KaEtAl16] (bottom)

Table 21.1 Design constraints and specifications

Biasing currents	$1mA < I_{C1,C2} < 10\ mA$
Power supply	$10V < V_{CC} < 20\ V$
Node voltage I	$V_{CC} - I_{C1} \cdot R_{C1} > 6\ V$
Node voltage II	$V_{CC} - I_{C2} \cdot R_{C2} > 6\ V$
Open-loop gain	$A_{OL} = \frac{\beta \cdot R_{C1} \cdot R_{C2} \cdot I_{C1}}{[R_{C1} + \beta V_T / I_{C2}] \cdot V_T} > 1000\ V/V$

where k is the Boltzmann constant and q is the electron's electrical charge.

Designing circuits in infant technologies such as SiC involves limitations such as the availability of the device type in the technology which in turn limits the choices of circuit topologies. In order to achieve high open-loop gain (>60 dB) with reasonable power consumption, considering design specifications and technology limitations, the constraints can be defined according to Table 21.1. In this regard, and based on the previously shown results on the selection of biasing point [KEtAl15, KaEtAl16] for having high enough current gain for the NPN devices, a safe range for the biasing current of the gain stages (I_1 and I_2) is 1 mA - 10 mA. Moreover, with a single supply opamp, the power supply V_{CC} has to be as low as possible to reduce the power consumption of the circuit. A safe minimum for this voltage is 10 V. Another important constraint is also the minimum voltage needed to guarantee that the amplifying devices are in *active* region. Using the

output characteristics of the BJTs, a safe biasing voltage at the output of the first and second amplifying stages can be chosen. Table 21.1 summarizes the variables and defined constraints.

Due to applications of this SiC opamp, this circuit must be designed for operation in $300^{\circ K} < T < 773^{\circ K}$. Moreover, the design must consider die-to-die and wafer-to-wafer variations of the device parameters. Since the forward current gain (β) of the device plays an important role on the performance of the whole circuit, it has been given a special attention. Thus, taking the previous fabrication processes into account and with a glimpse on the future outlook for bipolar SiC technology, β should be in the range of 10 to 100, and the design targets to meet the optimum performance at each selected current gain value.

21.3 Metaheuristic Optimization

In the previous section, gain and power consumption were introduced as two important performance metrics for operational amplifiers. Based on the equations provided, there is a trade-off between these two metrics which makes it difficult to find an optimal solution for both of them in all temperatures. Thus, in this section we first introduce an objective function for this optimization problem which considers both of these performance metrics. Then, we introduce an algorithm for finding a near optimal solution for the mentioned problem which also considers problem's constraints.

The optimization problem for getting the best gain and power consumption in all temperatures is presented in (21.2). In this problem, γ is the importance of the circuit's gain versus its power consumption. Therefore, in $\gamma = 1$ the only important metric is gain, while in $\gamma = 0$ power consumption is the main goal of the optimization. In (21.2), the constraints for specifying currents, resistors, and the voltage are also presented.

$$\begin{aligned} \max \quad & \sum_{t=300, t\in\mathbb{N}}^{773} \gamma * A_{OL} + \frac{1-\gamma}{V_{CC}(I_{C1}+I_{C2})} \\ s.t \quad & 0.001 < I_{C1}, I_{C2} < 0.01 \\ & V_{CC} - I_{C1} \cdot R_1 > 6 \\ & V_{CC} - I_{C2} \cdot R_2 > 6 \\ & 10 < V_{CC} < 20 \end{aligned} \tag{21.2}$$

In the above optimization problem, the circuit's currents and also the voltage have the resolution of 0.1 mA and 0.1 V, respectively. Therefore, the solution space in this problem is very large which makes it impossible to search the whole space for finding the optimal solution. It should also be mentioned that variations of resistance have also been considered based on [Ka17] for different temperatures.

Algorithm 1 Tabu Search Algorithm

```
function TABUALGM(Maximum Number of Iterations, Selection Probability)
    Generate a random feasible solution
    bestSolution ← ∅
    iterationNumber ← 0
    while iterationNumber < maxIteration do
        Generate neighbors of the current solution which are not in the tabu list
        Find the best neighbor
        if The best solution is better than the current solution then
            currentSolution ← bestNeighbor
        else
            Update the currentSolution based on the selection probability
        Update tabu list
        if The currentSolution is better than the bestSolution then
            Update the bestSolution
            iterationNumber ← 0
        else
            iterationNumber ← iterationNumber + 1
    return The best solution found
```

Metaheuristic algorithms [Ge09] search for the optimal solution based on a combination of techniques for exploring the solution space. Most of these algorithms are inspired by natural phenomena and provide a suboptimal solution in a reasonable amount of time. Tabu Search algorithm is a widely used metaheuristic algorithm which was created in 1986 by Fred W. Glover [Gl86]. This algorithm uses some techniques called diversification and intensification for searching for the optimal solution. In the former technique, the algorithm accepts worse solutions compared with the current solution in order to explore various parts of the solution space. However, the later technique helps the algorithm to find the best solution in a local region.

It is worth mentioning that the Tabu Search algorithm maintains a list of previously visited solutions in order to avoid visiting them again in the future. This is the reason why this algorithm is called Tabu Search. Our proposed Tabu Search algorithm for the optimization problem (21.2) is presented in Algorithm 1.

In the proposed algorithm, first a random feasible solution for the problem is generated. Then in every iteration of the algorithm, neighbors of the current solution are found. Neighbor of a solution in this algorithm is defined as a solution with the same parameters except that one random parameter is changed to another feasible value.

Among the computed neighbors in each iteration of the algorithm, the best one is compared with the current solution. If this neighbor provides a better solution based on the objective function of (21.2), then the current solution is replaced with this neighbor. Else with a probability which is the input of the algorithm, the best neighbor of the current solution is accepted as a new solution. This probability gives the algorithm the opportunity to explore more in the solution space and this mechanism is called Probabilistic Move Selection [WuHa13]. At the end of each

iteration, the tabu list is updated and the current solution is also compared with the best solution found so that it can be updated for a better solution. Finally, the best solution found in the algorithm is returned.

21.4 Results

In this section, the results of running the algorithm presented in previous section are introduced. The Tabu Search algorithm proposed for the optimization problem was programmed in Java and used for different values of β and γ.

Different design scenarios were investigated: At γ=1, all the algorithm's effort was spent to maximize the voltage gain, considering the power consumption as a less significant spec of the design. On the contrary, γ=0 describes a case where the overall power efficiency is the main goal.

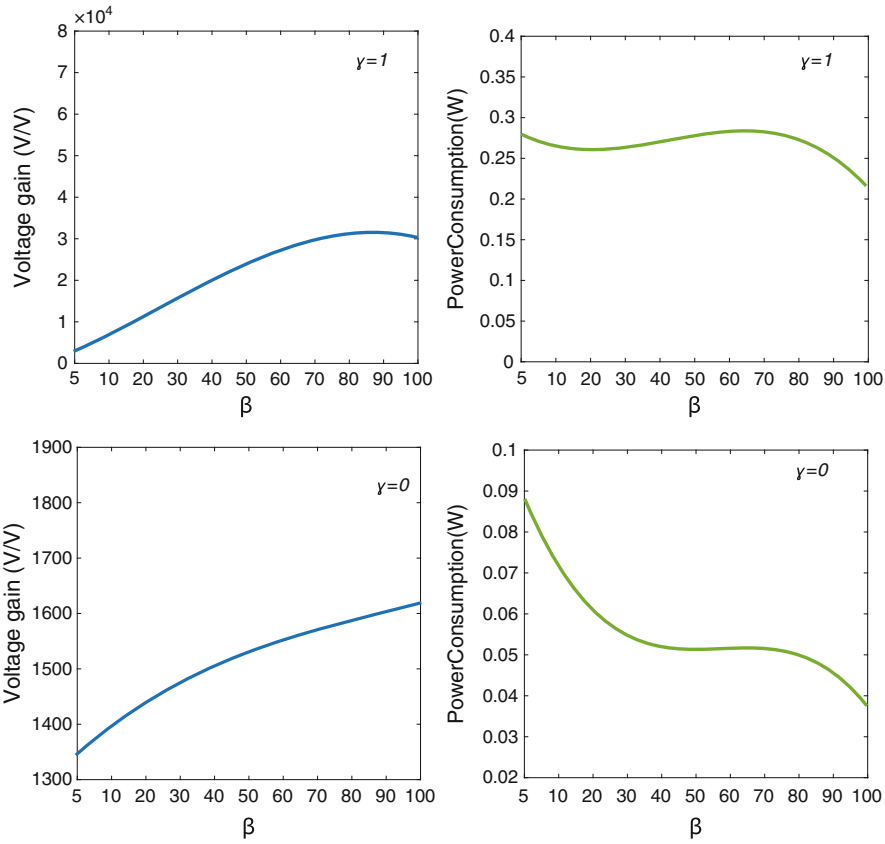


Fig. 21.2 Average open-loop voltage gain and power consumption of the opamp for different values of γ

Figure 21.2 illustrates the voltage gain and power consumption values found by the algorithm in the mentioned scenarios. It can be seen that voltage gain of 10000–30000 (80–90 dB), even at low β values, is achievable when power consumption is not the main concern. However, in this case the power consumption of the two gain stages are $\approx 200 - 270$ mW. On the other hand, when power consumption is the main goal, it is possible to reduce it as low as 40 mW at the voltage gain of 1350–1650 (62–64 dB). Table 21.2 also shows the optimized variables along with their resulting average voltage gain and also power consumption for sample values of β and γ.

The reported simulated voltage gain of [KaEtAl16, Ka14], when the power consumption was not a concern, was around 70 dB. Comparing this with the results of the proposed metaheuristic algorithm reveals that using metaheuristics provides at least 10dB higher voltage gain than the designer's choice for the circuit.

21.5 Conclusions

A circuit designer in emerging technologies such as SiC has to tackle numerous challenges such as variations in the fabrication process and the device parameters. Therefore, this paper aimed at optimizing voltage gain and power consumption of a SiC operational amplifier. For solving this problem, a Tabu Search algorithm was proposed as a metaheuristic for finding a near optimal solution in a reasonable amount of time. The results of running this algorithm reveal that it can produce at least 10dB higher voltage gain when compared with the designer's simulation. Therefore, the proposed metaheuristic can clearly aid designers to enhance the voltage gain of their circuits, while considering their power consumption.

Table 21.2 Optimized variables for sample β and γ values along with the resulting average voltage gains and also power consumptions

γ	β	R_{C1}	R_{C2}	I_{C1}	I_{C2}	V_{CC}	Average Voltage Gain	Power Consumption
1	20	1390	1990	0.01	0.007	20	7870.814	0.34
1	40	1390	3880	0.01	0.0036	20	24625.35	0.272
1	60	1390	6990	0.01	0.002	20	46061.98	0.24
1	80	1390	3490	0.01	0.004	20	36606.19	0.28
1	100	4110	2410	0.0034	0.0058	20	14876.9	0.184
0	20	676	2879	0.002	0.0019	11.5	1463.113	0.04485
0	40	1476	2177	0.0011	0.0027	11.9	1423.884	0.04522
0	60	870	570	0.0044	0.0027	10	1517.695	0.071
0	80	1060	1243	0.0033	0.001	10.1	1637.682	0.04343
0	100	1360	460	0.0029	0.0036	10	1504.305	0.065

In future works, optimizing other performance metrics of the SiC operational amplifier can also be considered. Moreover, the proposed method can be extended for other analog circuits in SiC and similar emerging semiconductor technologies.

References

[AlEtAl15] Alexandru, M., Banu, V., Jordá, X., Montserrat, J., Vellvehi, M., Tournier, D., Millán, J., Godignon, P.: SiC integrated circuit control electronics for high-temperature operation. IEEE Trans. Ind. Electron. **62**, 3182–3191 (2015)

[ChKo98] Chen, J.S., Kornegay, K.T.: Design of a process variation tolerant CMOS opamp in 6H-SiC technology for high-temperature operation. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **45**(11), 1159–1171 (1998)

[CoEtAl07] Cooren, Y., Fakhfakh, M., Loulou, M., Siarry, P.: Optimizing second generation current conveyors using particle swarm optimization. In: 2007 International Conference on Microelectronics, pp. 365–368. IEEE (2007)

[CrMa12] Cressler, J.D., Mantooth, H.A.: Extreme Environment Electronics. CRC Press, Boca Raton (2012)

[DeEtAl10] Delican, Y., Vural, R.A., Yildirim, T.: Artificial bee colony optimization based CMOS inverter design considering propagation delays. In: 2010 XIth International Workshop on Symbolic and Numerical Methods, Modeling and Applications to Circuit Design (SM2ACD), pp. 1–5 (2010)

[DeTl15] de la Fraga, L.G., Tlelo-Cuautle, E.: Optimizing operational amplifiers by meta-heuristics and considering tolerance analysis. In 2015 16th Latin-American Test Symposium (LATS), pp. 1–4. IEEE (2015)

[Ge09] Geem, Z.W.: Studies in Computational Intelligence, vol. 191. Springer Berlin Heidelberg (2009)

[Gl86] Glover, F.: Future paths for integer programming and links to artificial intelligence. Comput. Oper. Res. **13**(5), 533–549 (1986)

[HaEtAl16] Harris, R.K., McCue, B.M., Roehrs, B.D., Roberts, C., Blalock, B.J., Costinett, D.J., Sariri, K., Megyei, G., Chen, C.P., Kashyap, A., Ghandi, R.: A silicon carbide integrated circuit implementing nonlinear-carrier control for boost converter applications. In: 2016 IEEE Applied Power Electronics Conference and Exposition (APEC), pp. 3255–3258 (2016)

[JoEtAl15] Joshi, D., Dash, S., Agarwal, U., Bhattacharjee, R., Trivedi, G.: Analog circuit optimization based on hybrid particle swarm optimization. In: 2015 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 164–169 (2015)

[Ka14] Kargarrazi, S.: Bipolar silicon carbide integrated circuits for high temperature power applications, Licentiate Thesis (2014). http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-156212

[Ka17] Kargarrazi, S.: High temperature bipolar SiC power integrated circuits. Ph.D. Thesis (2017). http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-201618

[KarEtAl15] Kargarrazi, S., Lanni, L., Rusu, A., Zetterling, C.M.: A monolithic SiC drive circuit for SiC power BJTs. In: 2015 IEEE 27th International Symposium on Power Semiconductor Devices IC's (ISPSD), pp. 285–288 (2015)

[KaEtAl15] Kargarrazi, S., Lanni, L., Zetterling, C.M.: Design and characterization of 500°C Schmitt trigger in 4H-SiC. Mater. Sci. Forum **821**, 897–901 (2015)

[KaEtAl16] Kargarrazi, S., Lanni, L., Zetterling, C.M.: A study on positive-feedback configuration of a bipolar SiC high temperature operational amplifier. Solid State Electron. **116**, 33–37 (2016)

[KEtAl15] Kargarrazi, S., Lanni, L., Saggini, S., Rusu, A., Zetterling, C.M.: $500°C$ bipolar SiC linear voltage regulator. IEEE Trans. Electron Devices **62**, 1953–1957 (2015)

[LaEtAl13] Lanni, L., Malm, B.G., Östling, M., Zetterling, C.M.: $500°C$ bipolar integrated OR/NOR gate in 4H-SiC. IEEE Electron Device Lett. **34**, 1091–1093 (2013)

[LoEtAl02] Loulou, M., Ali, S.A., Fakhfakh, M., Masmoudi, N.: An optimized methodology to design CMOS operational amplifier. In: The 14th International Conference on 2002 - ICM Microelectronics, pp. 14–17. IEEE (2002)

[PaEtAl09] Patil, A.C., Fu, X.A., Mehregany, M., Garverick, S.L.: Fully-monolithic, $600°C$ differential amplifiers in 6H-SiC JFET IC technology. In: Custom Integrated Circuits Conference, 2009 (CICC '09), pp. 73–76. IEEE (2009)

[RaEtAl16] Rahman, A., Roy, S., Murphree,R., Kotecha, R., Addington, K., Abbasi, A., Mantooth, H.A., Francis, A.M., Holmes, J., Di, J.: High temperature SiC CMOS comparator and op amp for protection circuits in voltage regulators and switch-mode converters. IEEE J. Emerging Sel. Top. Power Electron. **PP**(99), 1–1 (2016)

[RyEtAl98] Ryu, S.H., Kornegay, K.T., Cooper, J.A., Melloch, M.R.: Digital CMOS IC's in 6H-SiC operating on a 5-V power supply. IEEE Trans. Electron Devices **45**(1), 45–53 (1998)

[SpEtAl16] Spry, D.J., Neudeck, P.G., Chen, L., Lukco, D., Chang, C.W., Beheim, G.M.: Prolonged $500°C$ demonstration of 4H-SiC JFET ICs with two-level interconnect. IEEE Electron Device Lett. **37**(5), 625–628 (2016)

[ThEtAl11] Thompson, R.F., Clark, D.T., Murphy, A.E., Ramsay, E.P., Smith, D.A., Young, R.A.R., Cormack, J.D., McGonigal, J., Fletcher, J., Zhu, C., Finney, S., Martin, L.C., Horsfall, A.B.: High temperature silicon carbide CMOS integrated circuits. Additional Papers and Presentations (2011). HiTEN, 000115–000119

[VaEtAl14] Valle-Mayorga, J.A., Rahman, A., Mantooth, H.A.: A SiC NMOS linear voltage regulator for high-temperature applications. IEEE Trans. Power Electron. **29**, 2321–2328 (2014)

[WuHa13] Wu, Q., Hao, J.K.: An adaptive multistart tabu search approach to solve the maximum clique problem. J. Comb. Optim. **26**(1), 86–108 (2013)

[XiEtAl94] Xie, W., Cooper, J.A., Melloch, M.R.: Monolithic NMOS digital integrated circuits in 6H-SiC. IEEE Electron Device Lett. **15**(11), 455–457 (1994)

Chapter 22
Severe Precipitation in Brazil: Data Mining Approach

H. Musetti Ruivo, H.F. de Campos Velho, and S.R. Freitas

22.1 Introduction

Heavy precipitation associated with severe weather systems has been one of the largest economic and social impacts. On 6–7 April 2010, the city of Rio de Janeiro and several neighboring municipalities were victims of extreme weather conditions. More than 150 mm of accumulated rainfall in a 24-hour period was recorded, and landslides occurred in the city. Many people died (233) in the Rio de Janeiro and Niteroi cities, at least 14,000 people have been made homeless, and vast stretches of road in various parts of the city and surrounding areas were partially ruined [MoEtAl13]. Other episodes to be cited are a series of floods and mudslides taking place on 12 January 2011 in the mountainous region of the State of Rio de Janeiro. More than 900 people have died. Thousands of people have been made homeless. In a 24-hour period between 11 and 12 January 2011, the local weather service registered more rainfall than what is expected for the entire month. Around 2,960 people had their homes destroyed [G111]. Figure 22.1 illustrates the location of the tragedies and destruction images.

Here is presented an innovative data mining (DM) approach to investigate the climatic condition linked with the extreme precipitation events occurred in Rio de Janeiro (Brazil), coupling two different techniques – one from statistical analysis and another one from artificial intelligence.

Our data mining approach was employed in previous research [RuEtAl14, RuEtAl15], applying a class-comparison technique also used as a tool to analyze

H.M. Ruivo (✉) • H.F. de Campos Velho
National Institute for Space Research (INPE), São José dos Campos, SP, Brazil
e-mail: helomusettir@gmail.com; haroldo@lac.inpe.br

S.R. Freitas
National Aeronautics and Space Administration (NASA), Washington, DC, USA
e-mail: saulo.freitas@noaa.gov

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_22

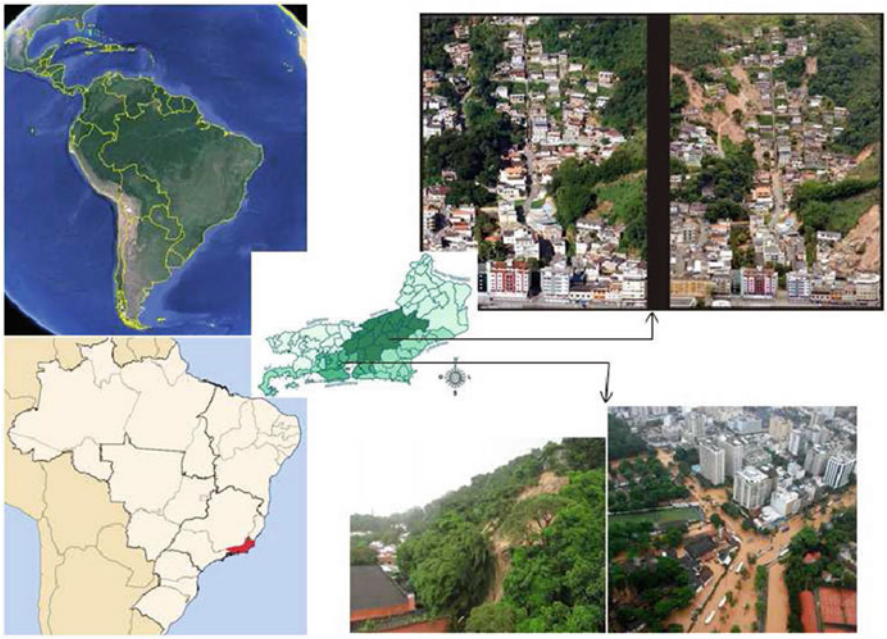


Fig. 22.1 Extreme precipitation event location in the state of Rio de Janeiro and its consequences

large datasets of genome-wide studies. The result of the statistical analysis is presented in a p-values map with the most relevant variables for the climate analysis. The maps show the climatological variables with higher correlation to the precipitation intensity.

The statistical analysis also serves as data reducing to apply a decision tree (DT) during the learning algorithm. DTs are used as a predictive model. The DT identifies the hierarchy of the influence of climatological variables associated with extreme precipitation in the analyzed region. Section 22.2 presents the methodology and datasets used in this investigation. Section 22.3 presents our results, and finally Section 22.4 draws some conclusions and discusses further developments.

22.2 Methodology

This work employs the data mining approach that comprises two steps of knowledge extraction: class-comparison and decision trees. Class-comparison uses a statistical approach for reducing the complexity of the original dataset. In the sequence, the DT method is applied as a severe weather predictor.

22.2.1 Class-Comparison

The class-comparison method is used for comparing two or more pre-defined classes in a time series of climatic grid box values. The purpose is to determine which variables in the dataset behave differently across pre-defined classes of precipitation. The "no-difference" case corresponds to a null hypothesis. The null hypothesis is the hypothesis of no effect, no correlation, or no association, whatever the case may be. The classes are defined in such a way to capture in the correct class the main episodes of extreme precipitation that occurred during the period being evaluated. There are several methods for checking whether differences in variable values are statistically significant [SiEtAl03]. The F-test is a generalization of the well-known t-test, which measures the distance between two samples in units of standard deviation. Large absolute values of the F-test suggest that the observed differences among classes are not due to chance, and that the null hypothesis can therefore be rejected. Supposing there are J1 data points of class 1 and J2 data points of class 2, the t-test score is computed as:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2 \left(\frac{1}{J_1} + \frac{1}{J_2}\right)}} \tag{22.1}$$

where:

$$s_p^2 = \frac{(J_1 - 1)s_1^2 + (J_2 - 1)s_2^2}{J_1 + J_2 - 2}\ , \quad \text{and} \quad s_i^2 = \frac{1}{J_i - 1}\sum_{j=1}^{J_i}(x_{ij} - \overline{x_i})^2 \ \ (i = 1, 2)\ .$$

The used notation in the above equation is expressed as $\overline{x_1}$= mean of samples class 1, $\overline{x_2}$= mean of samples class 2, J_1= quantity of samples class 1, J_2= quantity of samples class.

A F-test shall be computed for more than two classes. In this case, the alternative to the null hypothesis is that at least one of the classes has a distribution that is different from the others. The t-test and F-test computed are then converted into probabilities, known as p-values. The p-value is the probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming that the model is true. The *p*-value is a measure of statistical significance, meaning – under the null hypothesis – the p-values are less than 0.01 only 1% of the realization. Permutations methods, where Gaussianity is not assumed, are commonly used for computing *p*-values [SiEtAl03, HaEtAl07]. After calculating t-test scores for each variable, the class labels of the J_1 and J_2 are randomly permuted, so that a random J_2 of the samples are temporarily labeled as class 1, and the remaining J_2 samples are

labeled as class 2. Using these temporarily labels, a new t-test score is calculated, say t*. The labels are then reshuffle many times again, with a t* being computed at each permutation. The p-value from the permutation t-test is given by:

$$\text{p-value} = \frac{1 + \#\text{random permutation where } |t^*| \geq |t|}{1 + \#\text{random permutation}}.$$

22.2.2 Decision Tree

There are several decision tree (DT) algorithms available. The J4.8 algorithm is available from the WEKA package [WiEtAl00]. DTs are tree-like recursive structures made of leafs, labeled with a class value, and test nodes with two or more outcomes, each linked to a sub-tree.

The DT algorithm construction consists of a collection of training cases, each having a tuple of values for a fixed set of attributes (independent variables) and a class attribute (dependent variable). The aim is to generate a map that relates an attribute value to a given class. The classification task is performed following down from the root the path dictated by the successive test nodes, placed along the tree, until a leaf containing the predicted class.

The analyzed problem is successively divided into smaller subproblems until each subgroup addresses only one class, or until one of the classes shows a clear majority not justifying further divisions. Most algorithms attempt to build the smallest trees without loss of predictive power. To this end, the J4.8 algorithm relies on a partition heuristic that maximizes the *information gain ratio*, the amount of information generated by testing a specific attribute. This approach permits to identify the attributes with the greatest discrimination power among classes, and select those that will generate a tree that is both simple and efficient.

The information gain is measured in terms Shannon's entropy reduction. Given a set A with two classes P and N, the information content (in bits) of a message that identifies the class of a case in A is then

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \tag{22.2}$$

where p is the total number of objects belonging to class P, and n is the total number of the objects into the class N. If A is partitioned into subsets $A_1,\ A_2, \ldots, A_m$ by a given test T, the information gained is given by

$$G(A;T) = I(A) - \sum_{i=1}^{V} \frac{p_i + n_i}{p+n} I(A_i) \tag{22.3}$$

where A_i has p_i objects from the class P, and n_i from the class N. The algorithm chooses the test T that maximizes the information gain ratio $G(A;T)/P(A;T)$, with

$$P(A;T) = -\sum_{i=1}^{V} \frac{p_i + n_i}{p + n} \log_2 \frac{p_i + n_i}{p + n} \tag{22.4}$$

being the information gain from the partition itself. The process is repeated recursively to obtain the other nodes, structuring the decision tree with the rest of the subsets [Qu93].

22.3 Results

The entire dataset used in this study comprises 8,398 time series. Gridded data cover a region delimited by latitudes 21^o S and 24^o S, and longitudes 45^o W and 41^o W. Pentad-averaged anomalies were used in the analysis. Anomalies were computed relative to the mean values over the period 2000–2011 (12 years). Surface- and pressure-level atmospheric fields have a spatial resolution of 0.25×0.25 degrees taken to 12 UTC and were extracted from the ECMWF climate reanalysis (www.ecmwf.int/en/forecasts/datasets). ECMWF uses its forecast models and data assimilation systems to "reanalyze" archived observations, creating global datasets describing the recent history of the atmosphere, land surface, and oceans. The list of the climatic variables is:

- Air temperature at the height level 2 m and pressure levels of 300, 500, 600, 700, 850, and 925 hPa;
- Geopotential, vertical velocity (Omega), specific humidity, zonal and meridional wind components at pressure levels of 300, 500, 600, 700, 850, and 925 hPa;
- Sea Surface Temperature.

The goal of this study is to determine which variables in the dataset behave differently across pre-defined classes of precipitation intensity. The "no-difference" case corresponds to the null hypothesis for the applications considered here.

22.3.1 Extreme Rainfall Event Over the City of Rio de Janeiro

The focus here is to identify variables that might correlate with observed differences among classes of precipitation in the region of Rio de Janeiro city (red dot in Figure 22.2). The rainfall dataset was provided by *Alerta Rio* system implemented by the Instituto de Geotecnica do municipio do Rio de Janeiro (GEORIO) (http://www.sistema-alerta-rio.com.br) and its precipitation network has 32 several rain gauges installed on different areas within the city.

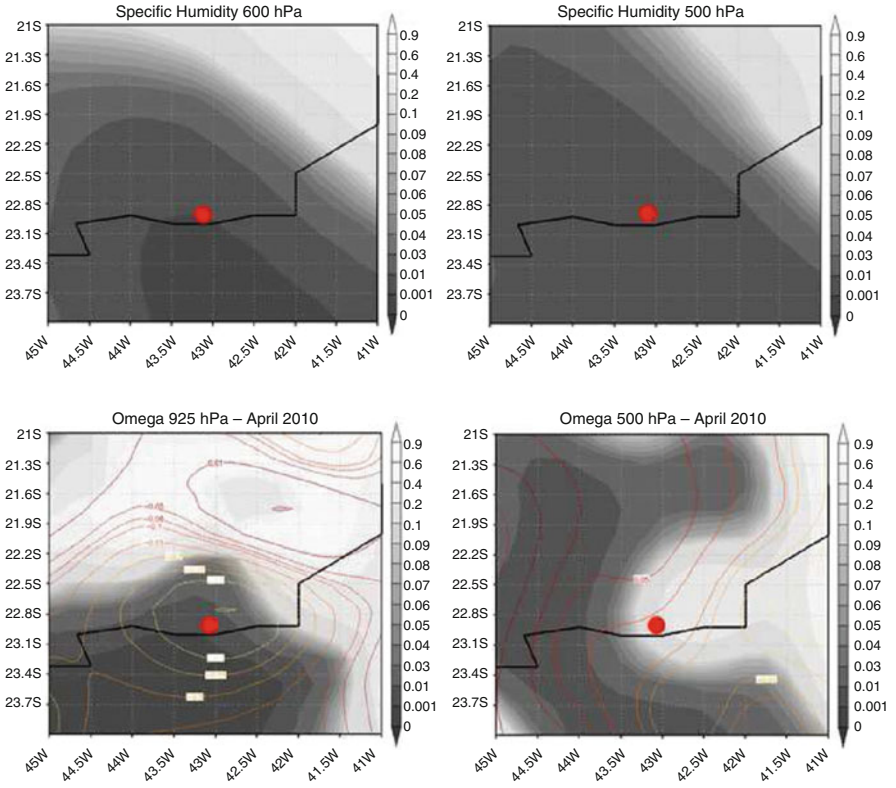


Fig. 22.2 p-value field for specific humidity at 600 and 500 hPa (upper), and omega at 925 and 500 hPa (bottom)

For classification purposes, the pentads of this time series were divided into three classes of precipitation intensity: "strong," "moderate," and "light" rainfall. The standard t-test (22.1) was applied, as recommended for applications with two classes: "strong" (precipitation greater than 8) and "moderate" (precipitation between 0 and 8). Fields of p-values for eight gridded climatic variables are presented in Figure 22.2. In this figure, the p-values at a given grid point can be interpreted as the probability that the observed difference between classes for this variable is the product of mere chance. Clearly, coherent patterns of low p-values are identified by darker areas. A p-value < 0.01, for example, indicates probability lower than 1 of being a false positive. The isolines in Figure 22.2 correspond to Omega anomalies averaged over the pentad April 6th up to 10th, 2010, the period of most intense precipitation in Rio de Janeiro. These results represent p-value fields, where coherent spatial patterns of low p-values indicate the existence of a possible links between specific humidity, Omega and zonal/meridional wind anomalies, at different levels, and the precipitation intensity in the region of Rio de Janeiro (Figure 22.2).

In the upper part of Figure 22.2, there is a dense dark area of low p-values for specific humidity coming from the ocean toward the continent to medium altitudes. It is also observed a dense dark area for omega at 925 hPa on the ocean that spreads to the mainland at 500 hPa. During the extreme rainfall episode, it was also observed (see the isolines on the bottom of Figure. 22.2) that Omega values are negative over the most affected region (red dot). It is well known that upward vertical motion over the continent can result in precipitation, under certain conditions (moisture, pressure field, for example). The low p-values in the fields of meridional wind appear in the Southern of Rio de Janeiro state at low altitudes (not shown), on the other hand, for the zonal wind it is observed low p-values at medium altitudes at the opposite side (not shown)

According to the Center for Weather Forecasting and Climate Studies (CPTEC) synoptic analysis[1], there was the presence of a cyclone with high pressure over the Atlantic, which was associated with a cold front acting primarily on the Atlantic to the South of Rio de Janeiro and Southern Brazil. Linking the cyclone presence and interaction with the unstable hot and wet mass air happening for days on the Southeast of Brazil was sufficient to trigger intense prefrontal activity from the ocean to the littoral of Rio de Janeiro.

The decision tree (DT) configured by using the J4.8 algorithm was created with confidence factor used for pruning (0.25), with number of instances per leaf equal to 2. The p-values were computed for all attributes – attribute here is considered a meteorological variable for a given coordinate (x, y, z). Only the 45 smallest p-vales were taken into account to feed the DT. Pentad anomaly was adopted for DT configuration – see [Ru13]. The best result for the DT was obtained with the 9 different climatological variables, considering 5 different coordinates for each variable. The designed DT classifies the Rio de Janeiro precipitation into two classes: "light" (values below 5) and "strong" (values above 5).The training set comprised annual data from 2000 up to 2006. The years of 2007 to 2010 were used to evaluate the DT performance.

The resulting DT, displayed in Figure 22.3, has 11 leafs (5 "strong" and 6 "light") and 10 decision nodes. The variable with the highest information gain is omega at 850 hPa (at coordinates 44.5^o W and 23.5^o S). Considering precipitation levels above 5, number of 39 cases were expressed (between 2007 up to 2010), and the DT hits in 13 cases (33.3%). For the considered period (2007 up to 2010), five pentads have rainfall above 5. In these 5 cases, the DT (Figure 22.3) hits the extreme rainfall.

[1] See the web-page: www.cptec.inpe.br/~rupload/arquivo/Notatec\RJ\060410.pdf~.

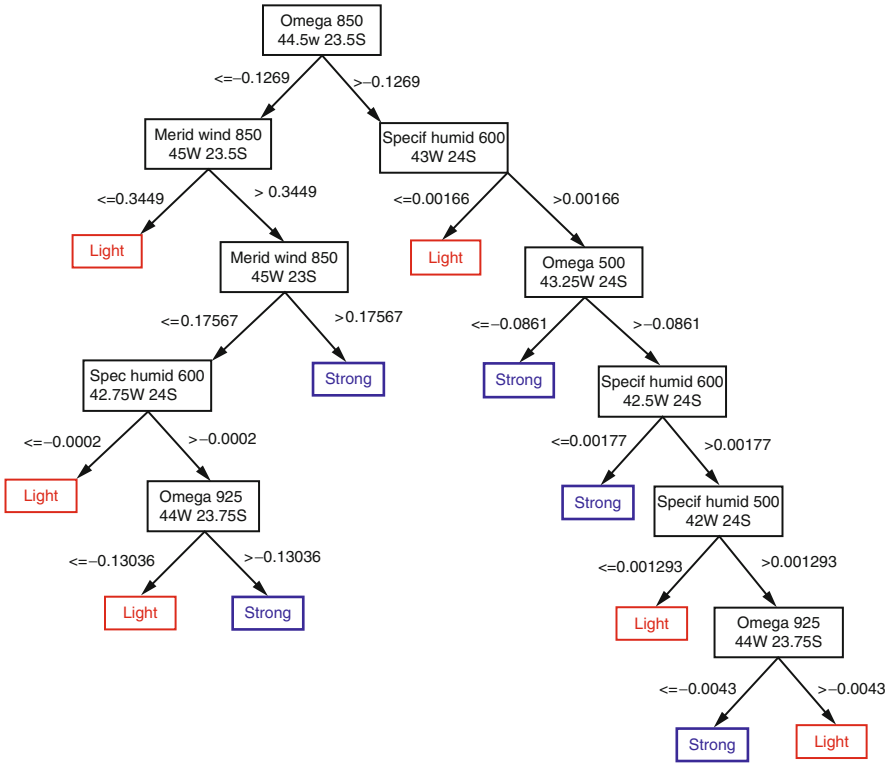


Fig. 22.3 DT using training set from 2000 up to 2006, and test set: from 2007 up to 2010

22.3.2 Extreme Rainfall Event Over Mountainous Region of the State of Rio de Janeiro

Data was computed the average of eight measurement stations from the Brazilian National Water Agency (ANA: Agencia Nacional de Águas), all placed on the region most affected by flooding.

For classification purposes, the pentads of these time series were divided into three classes of precipitation intensity: "strong," "moderate," and "light" rainfall. The standard t-test (22.1) was applied, as recommended for applications with two classes: "strong" (precipitation greater than 8) and "moderate" (precipitation between 0 and 8). Fields of p-values for seven gridded climatic variables are presented in Figure 22.4. The wind fields in Figure 22.4 are also anomalies averaged over the same period.

The low p-values for specific humidity (not shown) are once again noticeable at 850 hPa at the ocean. The low p-values (Figure 22.4, upper) of Omega at 500 hPa are highlighted on the continent, but at 300 hPa this dark area extends to whole region of analysis. It is also observed that the isolines in the pentad of the event remained

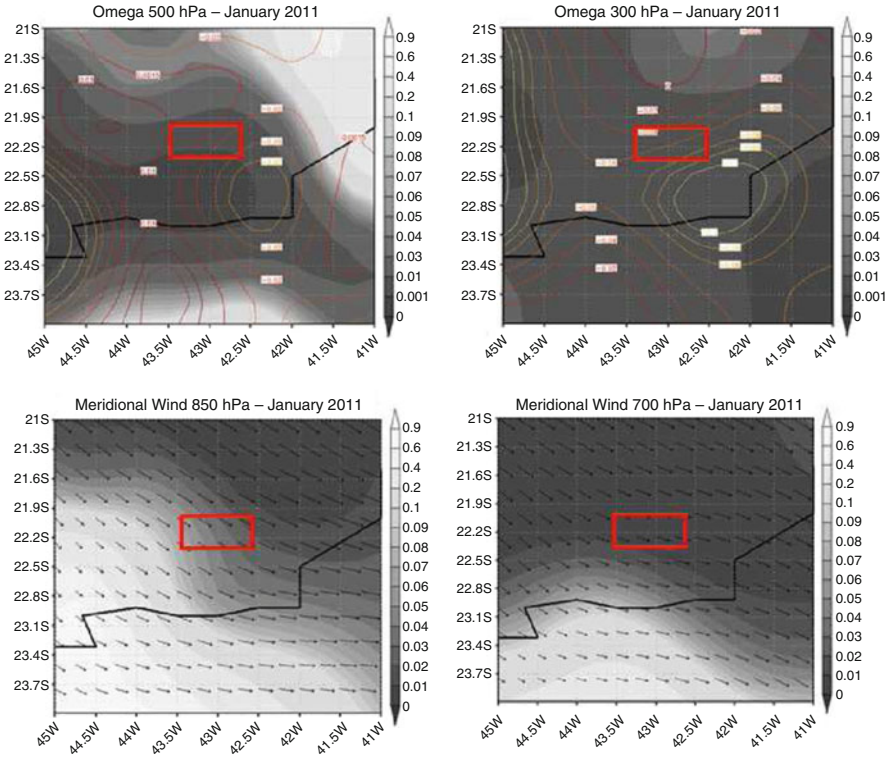


Fig. 22.4 p-value field for Omega at 500 hPa and 300 hPa (upper), meridional wind at 850 and 700 hPA (bottom)

negative. Different from the previous episode, Figure 22.4 bottom illustrates low p-values of the meridional wind at 850 and 700 hPa in the Northern part of the region, with winds blowing from the continent to the ocean. CPTEC's report (www.cptec.inpe.br/~ruplоad/arquivo/120111.pdf) mentions a strong divergent flow at high altitude over the state of Rio de Janeiro, which favors the occurrence of this event. Another determinant factor for the occurrence of this catastrophe was the complex orography of the region.

Several tests were performed with smallest p-values and the decision tree (confidence factor used for pruning =0.25, and number of instances per leaf =2) to generate the decision tree. The tree was obtained with the 15 different climatological variables, considering 5 different coordinates for each variable, with smallest p-values (total 75 attributes). The precipitation anomaly time series over the area were divided into two classes: "light" (values below 10) and "strong" (values above 10), corresponding to episodes of low and high precipitation, respectively. The training set comprised data from 2000 up to 2006. The years of 2007 to 2010 were used to evaluate the tree performance. The resulting tree, displayed in Figure 22.5, has 13 leafs (6 "strong" and 7 "light") and 12 decision nodes. The variable with the highest

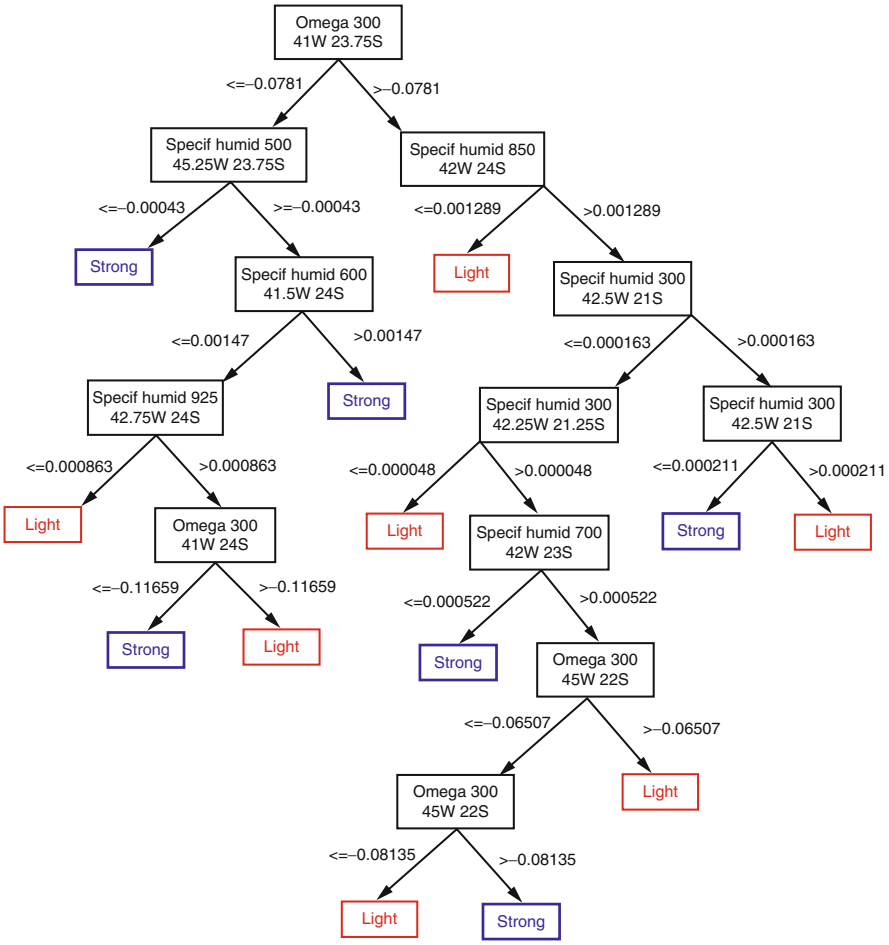


Fig. 22.5 DT using training set from 2000 up to 2006, and test set: from 2007 up to 2010

information gain is omega at 300 hPa, and at coordinates 41^o W and 23.75^o S. Analyzing only precipitation levels above 10, there was 21 cases (between 2007 up to 2010), and the tree hits in 8 cases (38%).

22.4 Conclusion

In this study, two steps of knowledge discovery were used to reduce the size of the input database. The aim was to investigate the climatic condition behind of two extreme events of rainfall occurred in Rio de Janeiro, Brazil in April 2010 and January 2011. The episode at 2011 was the most dangerous natural disaster recorded in the Brazil, killing more than 900 people.

The class-comparison applied methodology allows to mapping the influence (probability to be associated with the event) of different attributes to the event under study. The p-values maps become easier the interpretation for experts, pointing out the climatological variables directly related to the extreme event. In addition, it was able to promote a dramatic reduction on the size of the original dataset – for the current case from the order of thousands of variables to a few tenths. The decision trees generated from the results of the class-comparison step were able to correctly classify/predict cases of extreme rainfall in Rio de Janeiro city, and mountain region in Rio de Janeiro state. Overall, the data mining procedure has shown to be a promising approach in the investigation of climatic extreme events from the extraction of knowledge from large and complex datasets, with the potential of to be applied in real time weather forecast in the operational centers like CPTEC.

References

[G111] Freire, A., Lauriano, C., Araújo, G., Leta, T.: G1 RJ Número de mortos na Região Serrana do Rio passa de 400. http://g1.globo.com/rio-de-janeiro/chuvas-no-rj/noticia/2011/01/numero-de-mortos-na-regiao-serrana-do-rio-passa-de-400.html. Accessed 20 Apr 2014

[HaEtAl07] Hardin, J., Mitani, A., Hicks, L., VanKoten, B.: A robust measure of correlation between two genes on a microarray. BMC Bioinformatics **8**, 220 (2007)

[MoEtAl13] Moura, C.R.W., Escobar, G.C.J., Andrade, K.M.: Padrões de circulacão em superfície e altitude associados a eventos de chuva intensa na Região Metropolitana do Rio de Janeiro. Revista Brasileira de Meteorologia **28**(3), 267–280 (2013)

[Qu93] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)

[Ru13] Ruivo, H.M.: Metodologias de Mineracão de Dados em Análise Climática. Ph.D. Thesis, Applied Computing (CAP), National Institute for Space Research (INPE), São José dos Campo, São Paulo, Brazil (2013)

[RuEtAl14] Ruivo, H.M., Sampaio, G., Ramos, F.M.: Knowledge extraction from large climatological data sets using a genome-wide analysis approach: application to the 2005 and 2010 Amazon droughts, pp. 1–15. Climatic Change (2014)

[RuEtAl15] Ruivo, H.M., Campos Velho, H.F., Sampaio, G., Ramos, F.M.: Analysis of extreme precipitation events using a novel data mining approach. Am. J. Environ. Eng. **5**, 96–105 (2015)

[SiEtAl03] Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., Zhao, Y.: Design and Analysis of DNA Microarray Investigations, vol. 209. Springer, New York (2003)

[WiEtAl00] Witten, I.H., Frank, E.S.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann Publishers, San Mateo (2000)

Chapter 23
Shifting the Boundary Conditions to the Middle Surface in the Numerical Solution of Neumann Boundary Value Problems Using Integral Equations

A.V. Setukha

23.1 Introduction

The main idea of the boundary integral equation methods for boundary value problems is based on an integral representation of the solution. The integral equation for the unknown density of this integral representation appears as a result of satisfaction of the boundary conditions in the boundary value problem. This approach is the basis for many numerical methods.

In three-dimensional boundary value problems typical difficulties arise when the boundary problem exterior to a body of small thickness is solved: degeneration of the integral equation, when the thickness tends to zero; the need to use a small step of partition for numerical solution (the partitioning step must be much less than thickness of the body); there is a large error of numerical solution near the edges.

Replacing the object by a thin screen often simplifies the problem. But some of the physical effects cannot be simulated in this way.

The 3-D Neumann boundary value problem for the Laplace equation exterior to a body of small thickness is considered in this article. An approach is proposed to solve the problem in which the boundary conditions are transferred to the middle surface of the body. As a result, a new boundary value problem on the screen (the middle surface) is solved. Note that for two-dimensional problems of aerodynamics of the wing profiles such an idea has been developed in [LiEtAl92]. The idea from article [LiEtAl92] is developed for the three-dimensional case in the present paper.

A.V. Setukha (✉)
Lomonosov Moscow State University, Moscow, Russia
e-mail: setuhaav@rambler.ru

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_23

23.2 Shifting the Boundary Conditions to the Middle Surface and Numerical Method

The classical exterior Neumann boundary value problem for the Laplace equation is considered:

$$\Delta u = 0 \ \ \Omega, \ \frac{\partial u}{\partial n} = f \ on \ \Sigma, \ u(\mathbf{x}) \to 0 \ as \ \mathbf{x} \to \infty, \tag{23.1}$$

where Ω is a domain which lies outside its boundary Σ, where Σ is a closed surface.

Let's assume that the surface has the following structure. Let Σ_0 – some smooth open oriented surface with an edge $\partial\Sigma_0$. Let $\Sigma = \Sigma^+ \cup \Sigma^-$ where

$$\Sigma^{\pm} = \left\{ \mathbf{z}^{\pm}(\mathbf{z}) = \mathbf{z} \pm \frac{1}{2}\lambda(\mathbf{z})\mathbf{n}(\mathbf{z}), \ \mathbf{z} \in \Sigma_0 \right\},$$

$\lambda(\mathbf{z})$ is some function on the surface Σ_0 such that $\lambda(\mathbf{z}) \geq 0$, $\lambda(\mathbf{z}) = 0$ at the edge of Σ_0, $\mathbf{n}(\mathbf{z})$ is the unit normal vector to Σ_0. Let $\mathbf{n}^{\pm} = \mathbf{n}^{\pm}(\mathbf{z})$ are the unit outward normals to the surface Σ at the points $\mathbf{z}^{\pm}(\mathbf{z})$, respectively. It is also assumed that $\lambda(\mathbf{z})$ is much smaller than the surface Σ_0 size (see Figure 23.1).

It is proposed to consider a new boundary value problem outsize the screen Σ_0 for the approximate solution of the problem (23.1):

$$\Delta u = 0 \ \text{ in } \Omega_0, \ \left(grad\, u, \mathbf{n}^{\pm}\right) = f^{\pm} \ on \ \Sigma_0, \tag{23.2}$$

$f^{\pm}(\mathbf{z}) = f(\mathbf{z}^{\pm}(\mathbf{z}))$, $z \in \Sigma_0$, $\Omega_0 = R^3 \backslash \Sigma_0$, with conditions $grad\, u \in L_2^{loc}(\Omega_0)$, u is bounded in $R^3 \backslash \Sigma_0$, $u(\mathrm{x}) \to 0$ as $|\mathbf{x}| \to \infty$, $\exists u^{\pm}$ on Σ_0, and $\exists\, (grad\, u)^{\pm}$ on $\Sigma_0 \backslash \partial\Sigma$.

We use the representation of the solution of the problem (23.2) in the form of a sum of single and double layer potentials:

$$u = V[\Sigma_0, \mu] + U[\Sigma_0, g], \tag{23.3}$$

$$V[\Sigma, \mu](\mathbf{x}) = \int_{\Sigma} \mu(\mathbf{y}) F(\mathbf{x} - \mathbf{y}) d\sigma_y, \ U[\Sigma, g](\mathbf{x}) = \int_{\Sigma} g(\mathbf{y}) \frac{\partial F(\mathbf{x} - \mathbf{y})}{\partial n_y} d\sigma_y,$$
$$F(\mathbf{x} - \mathbf{y}) = \frac{1}{4\pi} \frac{1}{|\mathbf{x} - \mathbf{y}|} \tag{23.4}$$

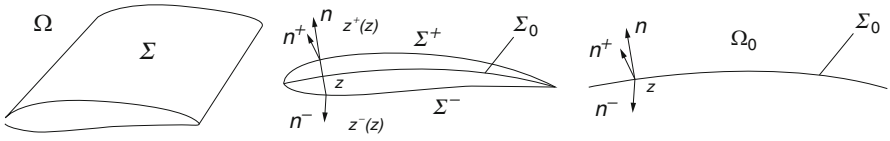


Fig. 23.1 Shifting the boundary condition

Let us recall some properties of the boundary values of potentials of simple and double layer [CoKr84]. Let $v = V[\Sigma, \mu],\ u = U[\Sigma, g]$.

If the functions μ and g are continuous on the surface Σ, then the functions v and u take the following boundary values on the surface Σ:

$$v^+ = v^- = v, \quad u^+ = u + \frac{g}{2}, \quad u^- = u - \frac{g}{2}. \tag{23.5}$$

If the function g is a Hölder continuous on the surface Σ, then the boundary values of the gradient of the function v on the surface Σ satisfy the relations :

$$(grad\ v)^{\pm} = grad\ v\ \mp \frac{\mu}{2}\mathbf{n},\ grad\, v(\mathbf{x}) = \int_{\Sigma} \mu(\mathbf{y}) grad_x F(\mathbf{x}-\mathbf{y}) dy, \tag{23.6}$$

where the integral is treated in the sense of the principal value. If the function g satisfies the condition $g(x) = 0$ on the edge of the surface Σ, and the surface gradient *Gradg* is Hölder continuous on the surface Σ, then the boundary values of the gradient of the function u on the surface Σ satisfy the relations:

$$(grad\, u)^{\pm} = \mathbf{n}\frac{\partial u}{\partial n} + \int_{\Sigma} g(\mathbf{y}) Grad_x \left(\frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y}\right) d\sigma_y \pm \frac{1}{2} Grad\, g, , \tag{23.7}$$

the integral is treated in the sense of the principal value. The boundary values of the normal derivative of the function u satisfy the relations [LiEtAl04]:

$$\left(\frac{\partial u}{\partial n}\right)^{\pm}(\mathbf{x}) = \int\limits_{\Sigma} g(\mathbf{y}) \frac{\partial^2 F(\mathbf{x}-\mathbf{y})}{\partial n_x \partial n_y} d\sigma_y \equiv \lim_{\varepsilon \to 0} \left[\int\limits_{\Sigma / U(\mathbf{x},\varepsilon)} g(\mathbf{y}) \frac{\partial^2 F(\mathbf{x}-\mathbf{y})}{\partial n_x \partial n_y} dy - \frac{g(\mathbf{x})}{2\varepsilon} \right]. \tag{23.8}$$

The last expression is the definition of the integral in the sense of the Hadamard finite value.

Formulas (23.5)–(23.8) are taken in the points of smoothness on the surface Σ, excluding the edges of the surface.

Let's return to the problem (23.2). Substituting the function u of the form (23.3) into the boundary condition and by using relations (23.5)–(23.8), we obtain the system of integral equations for the unknown functions μ and g

$$q^{\pm}(\mathbf{x}) \int_{\Sigma_0} g(\mathbf{y}) \frac{\partial^2 F(\mathbf{x}-\mathbf{y})}{\partial n_x \partial n_y} d\sigma_y + \int_{\Sigma_0} g(\mathbf{y}) \left(\boldsymbol{\tau}^{\pm}(\mathbf{x}), grad_x \frac{F(\mathbf{x}-\mathbf{y})}{\partial n_y}\right) d\sigma_y +$$
$$+ \int_{\Sigma_0} \mu(\mathbf{y}) \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_x^{\pm}} d\sigma_y \mp \frac{1}{2} q^{\pm}(\mathbf{x}) \mu(\mathbf{x}) \pm \frac{1}{2} \left(Grad\, g(\mathbf{x}), \mathbf{n}^{\pm}(\mathbf{x})\right) = f^{\pm}(\mathbf{x}),\ \mathbf{x} \in \Sigma_0, \tag{23.9}$$

$q^{\pm} = \left(\mathbf{n}, \mathbf{n}^{\pm}\right), \boldsymbol{\tau}^{\pm} = \mathbf{n}^{\pm} - q^{\pm}\mathbf{n}$.

More detailed proof of these equations is described in [Se16]. In this article has been proved the unique solvability of the boundary value problem (23.2) as well as of the system of integral Equations (23.9) in a case where the surface Σ_0 is plane

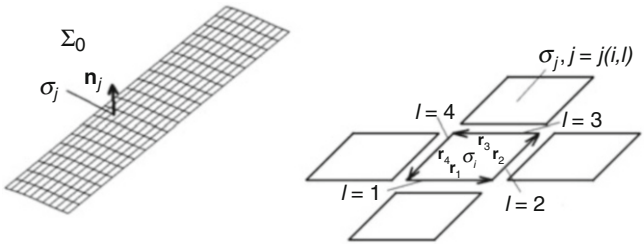


Fig. 23.2 Partition of the surface Σ_0

and under significant additional assumptions about the behavior of the vectors $\mathbf{n}^{\pm}$. Note that these additional restrictions are significant for the proof of solvability of the problem, but not for the derivation of the Equations (23.9).

Integral Equations (23.9) are solved numerically using the methods of piecewise constant approximations and collocation (see Figure 23.2). To construct the numerical scheme let's approximate surface Σ_0 with sets of cells σ_j, $j = 1, \ldots, n$. Let $\mathbf{x}_j \in \sigma_j$, $j = 1, \ldots, n$ – collocation points selected one for each cell, $\mathbf{n}_j = \mathbf{n}(\mathbf{x}_j)$. The cells in the form of quadrangles are used. The point $\mathbf{x}_j$ on cell σ_j selected as the crossing of the diagonals, the vector $\mathbf{n}_j$ constructed as the normal vector to these diagonals. Let also $\mathbf{n}_j^+ = \mathbf{n}^+(\mathbf{x}_j)$, $\mathbf{n}_j^- = \mathbf{n}^-(\mathbf{x}_j)$ - the normal vectors to the surfaces Σ^+ and Σ^- , respectively. Assuming that layers potentials density is constant on each cell. The solution u of problem (23.2) and its gradient can be approximated by expressions

$$u(\mathbf{x}) = \sum\nolimits_{j=1}^{n} g_j \int_{\sigma_j} \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y} d\sigma_y + \sum_{i=1}^{n} \mu_j \int_{\sigma_j} F(\mathbf{x}-\mathbf{y}) d\sigma_y,$$

$$grad\, u(\mathbf{x}) = \sum_{i=1}^{n} g_i \mathbf{V}_{\gamma,i}(\mathbf{x}) + \sum_{i=1}^{n} \mu_i \mathbf{V}_{q,i}(\mathbf{x}), \tag{23.10}$$

$$\mathbf{V}_{q,i}(\mathbf{x}) = \int_{\sigma_i} grad_x F(\mathbf{x}-\mathbf{y}) d\sigma_y \,,\ \mathbf{V}_{\gamma,j}(\mathbf{x}) = grad_x \int_{\sigma_j} \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y} d\sigma_y.$$

The velocity field $\mathbf{V}_{\gamma,j}(\mathbf{x})$ can be represented by Bio-Savart law and calculated analytically [LiEtAl04].

For the approximation of the vector $Grad\, g$ in the points $\mathbf{x}_j$, $j = 1, \ldots, n$, the following representation is used $Grad\, g = \mathbf{n} \times \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = Grad\, g \times \mathbf{n}$. Further formulas from the work [GuEtAl06] are used (see Figure 23.2):

$$Grad\, g(\mathbf{x}_i) = \mathbf{n}_i \times \gamma_i, \quad \gamma_i \approx -\frac{\Gamma_1^i + \Gamma_2^i + \Gamma_3^i + \Gamma_4^i}{s_i}, \tag{23.11}$$

$\Gamma_l^i = -\left(g_i - g_{j(l,i)}\right)\mathbf{r}_l/2$, $j(l,i)$ - cell number, which is bordered with the considered cell along the segment number l, s_i – area of the cell σ_i. We assume $g_{j(l,i)} = 0$ if there is no adjacent cell.

Writing the Equation (23.9) in the collocation points and using the representation (23.10), we obtain a system of linear algebraic equations for the unknowns g_j, μ_j, $j = 1, \ldots, n$

$$\sum_{j=1}^{n+m} a_{ij}^{\pm} g_j + \sum_{j=1}^{n} b_{ij}^{\pm} \mu_j + \frac{1}{2}[\boldsymbol{\gamma}_i \times \mathbf{n}_i]\mathbf{n}_i^{\pm} - \frac{1}{2}\mu_i \mathbf{n}_i \mathbf{n}_i^{\pm} = f_i^{\pm},\ i = 1, \ldots, n\ , \qquad (23.12)$$

$a_{ij}^{\pm} = \mathbf{V}_{\gamma,j}(\mathbf{x}_i)\mathbf{n}_i^{\pm}$ $b_{ij}^{\pm} = \mathbf{V}_{q,j}(\mathbf{x}_i)\mathbf{n}_i^{\pm}$ $f_i^{\pm} = -\mathbf{w}_\infty \mathbf{n}_i^{\pm}$, where $\boldsymbol{\gamma}_i$, $i = 1, \ldots, n$, are expressed in terms g_j using the formula (23.11), $m = 0$ for this problem.

23.3 Application to the Problem of the Flow Around a Wing in the Model of an Ideal Incompressible Fluid

Let's consider the flow around of a finite span a wing in the model of an ideal incompressible fluid (see Figure 23.3). The mathematical model described in the book [KaPl01] is used. It is assumed that the flow is potentially out of the wing and vortex wake. Vortex wake is approximated as potential discontinuity surface with a given shape in the plane. Note a wing surface as Σ, a surface that approximates the vortex wake as Σ_1, $L = \Sigma \cap \Sigma_1$ - separated line, Ω - the flow domain outside of the wing and vortex wake Σ_1 . The next boundary value problem for the velocity field $\mathbf{w} = \mathbf{w}(\mathbf{x})$, $\mathbf{x} \in \Omega$, is considered:

$$div\ \mathbf{w} = 0, \quad rot\,\mathbf{w} = 0 \text{ in } \Omega,\ \mathbf{wn} = 0 \text{ on } \Sigma,\ \mathbf{w}(\mathbf{x}) - \mathbf{w}_\infty \to 0 \text{ as } \rho(\mathbf{x}, \partial\Omega) \to \infty,$$

$$\mathbf{w}^+\mathbf{n} = \mathbf{w}^-\mathbf{n},\ p^+ = p^-\ on\ \Sigma_1,$$

where $\rho(\mathbf{x}, \partial\Omega)$ - the distance between the point and the set,

$$p = p_\infty + \rho\frac{\mathbf{w}_\infty^2}{2} - \rho\frac{\mathbf{w}^2}{2}$$

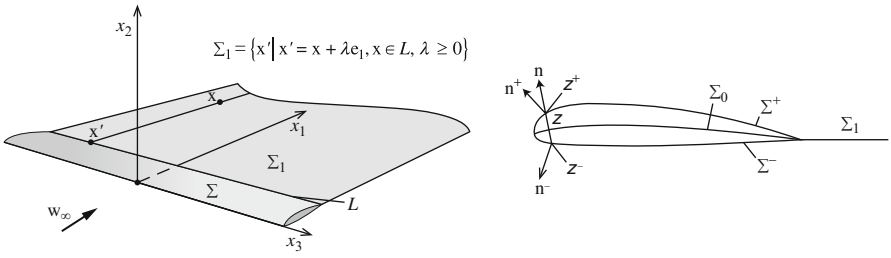


Fig. 23.3 Problem of the flow around a wing

- fluid pressure, $\mathbf{w}_\infty$ and p_∞ - given the velocity and the pressure at infinity. It is also assumed that $\mathbf{w}$ is bounded near the separation line L (except the ends).

Let us seek velocity field in the form $\mathbf{w}(x) = \mathbf{w}_\infty + grad\, u$. Then the boundary value problem for the potential appears:

$$\Delta u = 0 \text{ in } \Omega, \quad \frac{\partial u}{\partial n} = f \text{ on } \Sigma, \quad f = -\mathbf{w}_\infty \mathbf{n}, \tag{23.13}$$

$$\frac{\partial (u^+ - u^-)}{\partial x_1} = 0, \quad \left(\frac{\partial u}{\partial x_2}\right)^+ = \left(\frac{\partial u}{\partial x_2}\right)^- \text{ on } \Sigma_1, \; u(\mathbf{x}) \to 0 \text{ as } \rho(\mathbf{x}, \partial\Omega) \to \infty. \tag{23.14}$$

Assume that the surface of the wing Σ_0 has the same structure as the surface in the problem (23.1). Further the described method with shifting the boundary condition to the middle surface Σ_0 is used. Let $\Omega_0 = R^3 \backslash \Sigma_0 \backslash \Sigma_1$ be a domain outside the surfaces Σ_0 and Σ_1. New boundary problem in the domain Ω_0 is stated:

$$\Delta u = 0 \; in \; \Omega_0, \quad (grad\, u)^+ \mathbf{n}^+ = f^+, \; (grad\, u)^+ \mathbf{n}^+ = f^+ \; on \; \Sigma_0. \tag{23.15}$$

Conditions (23.14) are also required.

We seek a solution of the problem (23.14) in the form

$$u(\mathbf{x}) = U[\Sigma_0, g](\mathbf{x}) + V[\Sigma_0, \mu](\mathbf{x}) + U[\Sigma_1, g_1](\mathbf{x}),$$

where the potentials U and V are defined by the formulas (23.4). Then the following system of integral equations for the unknown potential density appears

$$\begin{gathered} \int\limits_{\Sigma_0} g(\mathbf{y}) \frac{\partial^2 F(\mathbf{x}-\mathbf{y})}{\partial n_x^{\pm} \partial n_y} dy + \int\limits_{\Sigma_1} g_1(\mathbf{y}) \frac{\partial^2 F(\mathbf{x}-\mathbf{y})}{\partial n_x^{\pm} \partial n_y} dy + \int\limits_{\Sigma_0} \mu(\mathbf{y}) \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_x^{\pm}} dy \mp \\ \mp \tfrac{1}{2} \mu(\mathbf{x}) (\mathbf{n}, \mathbf{n}^{\pm}) \pm \tfrac{1}{2} \left(Grad\, g, \mathbf{n}^{\pm}\right) = f^{\pm}, \; \mathbf{x} \in \Sigma_0 \end{gathered} \tag{23.16}$$

$$g_1(\mathbf{x} + \lambda \mathbf{e}_1) = g(\mathbf{x}) \; , \mathbf{x} \in L, \; \lambda > 0, \tag{23.17}$$

$\mathbf{n} = \mathbf{n}(\mathbf{x}), \quad \mathbf{n}^{\pm} = \mathbf{n}^{\pm}(\mathbf{x}), \quad \mathbf{x} \in \Sigma_0.$

For the numerical solution of Equations (23.16) – (23.17) the partition of the surface Σ_0 is performed, as well as for the Equations (23.9). The surface Σ_1 is approximated using the cells in the form of long strips (see Figure 23.4). We assume that the function g_1 is a constant on the each such cell in accordance with the condition (23.17). We use for the velocity field approximation

$$\mathbf{w}(\mathbf{x}) = \mathbf{w}_\infty + \sum_{i=1}^{n+m} g_i \mathbf{V}_{\gamma,i}(\mathbf{x}) + \sum_{i=1}^{n} \mu_i \mathbf{V}_{q,i}(\mathbf{x}),$$

where n is the number of cells on the surface Σ_0, m is the number of cells on the surface Σ_1. We obtain a system of linear algebraic Equations (23.13)–(23.14) for the

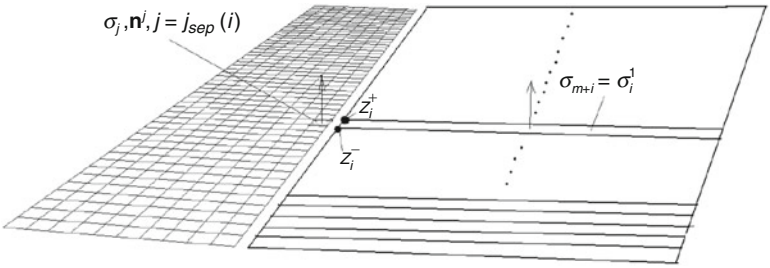


Fig. 23.4 Discretization the wing and vortex wake

unknowns g_j , $j = 1, \ldots, n+m$, μ_j , $j = 1, \ldots, n$. In addition, we write conditions

$$g_{n+i} = g_{j_{sep}(i)}, \ i = 1, \ldots, m.$$

These conditions follow from (23.17).

23.4 Numerical Results and Conclusions

The developed method is tested for solution of the problem of flow around of a rectangular wing (see Figure 23.5). For comparison we provided 3 variants of numerical solutions of the considered problem. The first variant uses a standard panel method for the original problem (23.13)–(23.14) - "volume wing" ([GuEtAl06, KaPl01]). Second one uses the developed method with shifting the boundary conditions to the middle surface (Equations (23.16)–(23.17) are solved numerically) - "thin wing with shifting the boundary condition." The third variant uses the standard panel method for solving the problem of the flow around the middle surface (without shifting the boundary conditions) - "thin wing" ([GuEtAl06, KaPl01]). Solutions using several variants of the partition of the wing surface were obtained for each of the 3 methods.

The obtained distribution of the pressure coefficients on the upper and lower surfaces of wings in the middle section are shown in Figure 23.6. The partition upper and lower wing surfaces on the $n_1 * n_2$ cells used for the variant "volume wing" (the first factor - the number of cells for the partition of the wing chord, the second - along the wing span). A similar partition of the middle surface used for variants "thin wing" and "thin wing with shifting the boundary condition." Pressure coefficient was introduced by the formula

$$C_p = \frac{p - p_\infty}{\rho \mathbf{w}_\infty^2/2} = 1 - \frac{1}{2}\frac{\mathbf{w}^2}{\mathbf{w}_\infty^2}.$$

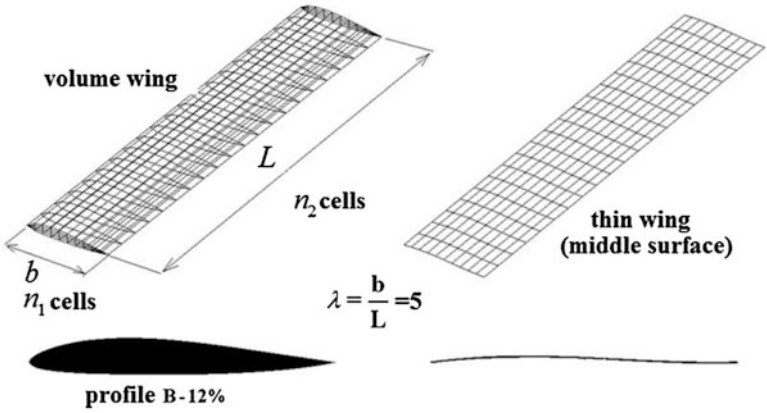


Fig. 23.5 The investigated wing

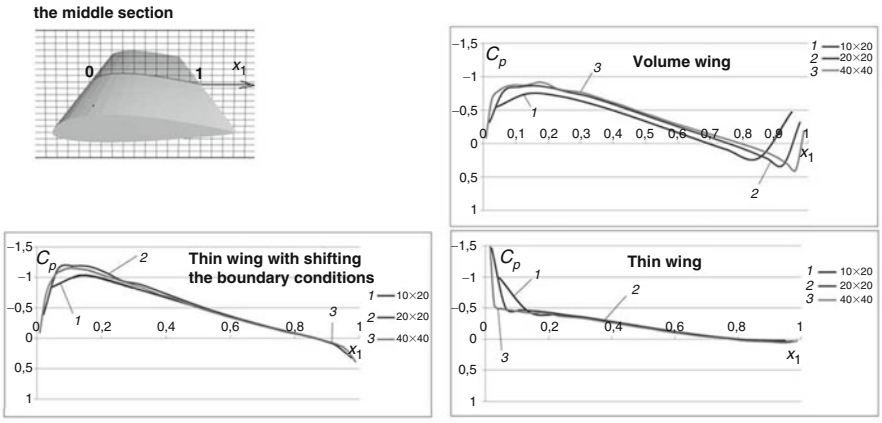
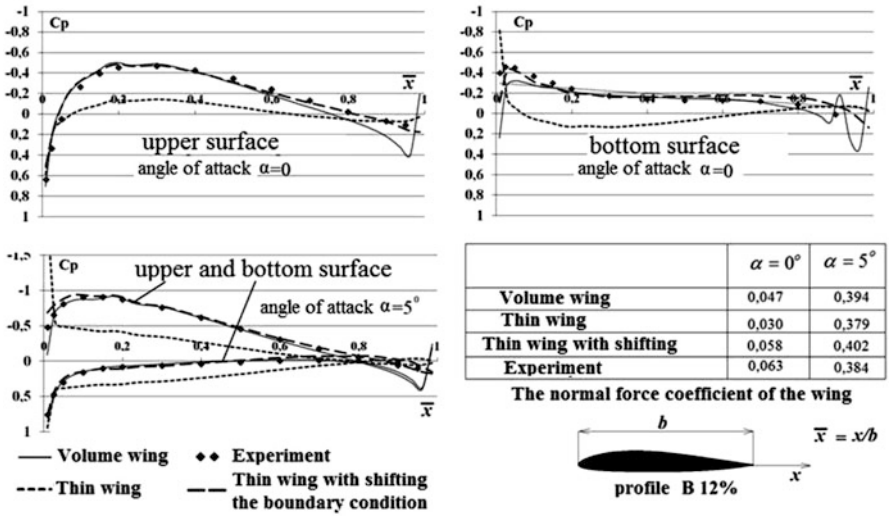


Fig. 23.6 The dependence of the solution of the number of cells

In variants "thin wing" and "thin wing with shifting the boundary condition," the boundary values of pressure on upper and lower sides of the middle surface have been considered as the values on upper and lower of the real wing surfaces. The boundary values of the velocity vector **w** calculated using formulas (23.5) – (23.8).

Figure 23.7 shows the pressure distributions on upper and lower wing surfaces in the middle section which were obtained using these three methods for comparison. Normal force coefficient values obtained in these calculations are shown in the table in Figure 23.7. Known experimental data are also shown (experiment TsAGI, Russia).

The analysis of these results allows the following conclusions. We have a faster convergence of numerical solutions for the scheme with shifting the boundary conditions (as well as for the problem of flow around a thin surface) in comparison



	$\alpha = 0°$	$\alpha = 5°$
Volume wing	0,047	0,394
Thin wing	0,030	0,379
Thin wing with shifting	0,058	0,402
Experiment	0,063	0,384

The normal force coefficient of the wing

Fig. 23.7 Comparison of the numerical solutions

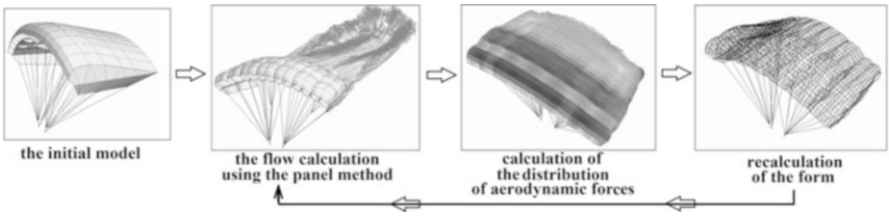


Fig. 23.8 The scheme of solving the problem of flow around a gliding parachute and the calculation of its shape

with the case of the volume wing. Replacement of the volume wing to thin surface allows to calculate the total force, but not the pressure distribution. The model with shifting the boundary conditions allows to calculate correctly the pressure distribution over the entire surface of the wing. In solving the original problem of the flow around the volume wing we have significant errors near the trailing edge.

The proposed method is applied to the problem of flow around a gliding parachute and the calculation of its shape. We used previously developed (jointly with Aparinov V.A., Morozov V.I., and Kiryakin V.Yu.) method for this problem based on panel methods. At first, we construct the initial form of the parachute. Then the calculation of the flow around this initial form using panel method is performed. Pressure difference distribution on the surface of the canopy of the parachute is calculated. Next we solve the problem of elasticity theory for determining the shape of a parachute under the load, which was calculated. Then aerodynamic calculation is performed again. The iterations are repeated (see Figure 23.8).

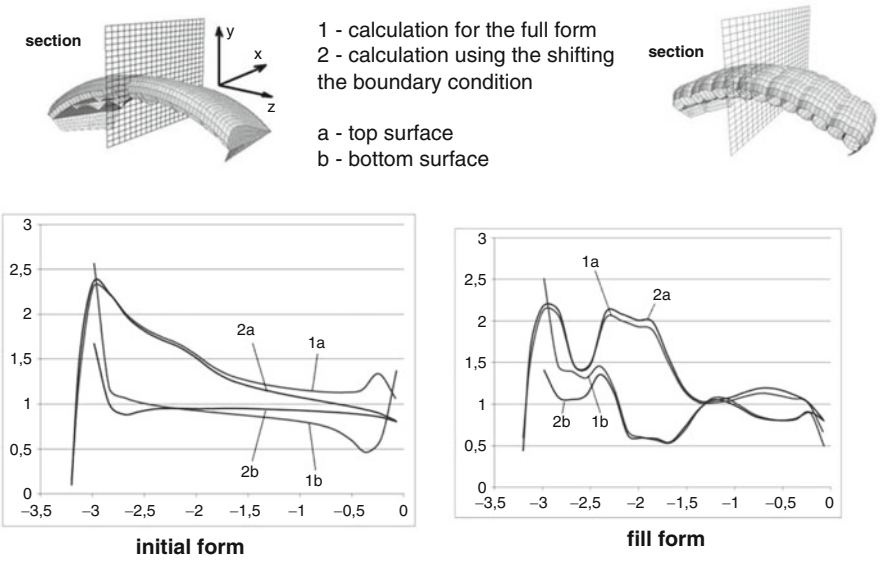


Fig. 23.9 The distribution of differential pressure coefficient in the median section

Within this model during solving the problem flow around the parachute the method of shifting the boundary condition to the middle surface was used . The Figure 23.9 (at the left) shows the initial form of a parachute and distribution of difference pressure coefficient in middle section, obtained by use of the classical panel method and of the new method. Classic panel method leads to a large error in the neighborhood of the trailing edge. The surface of the canopy would be a wrong bend, if these results will be used in solving the problem of calculating form.

The final canopy shape obtained by using the new method is shown in Figure 23.9 (at the right). This figure shows the distribution of differential pressure coefficient obtained by a new method. Also, for this final form we solve the problem of fluid flow past parachute with classical panel method. Author provide differential pressure coefficient distribution obtained in this calculation for the comparison. The aerodynamic loads obtained by the two methods are in good agreement. Thus, the new method allows us to produce more stable results for the initial form of parachute and for beginning the iterations.

References

[CoKr84] Colton, D., Kress, R.: Integral Equation Method in Scattering Theory. Wiley, New York (1984)

[GuEtAl06] Gutnikov, V.A., Lifanov, I.K., Setukha, A.V.: Simulation of the aerodynamics of buildings and structures by means of the closed vortex loop method. Fluid Dyn. **41**, 555–567 (2006)

[KaPl01] Katz, J., Plotcin, A.: Low-speed aerodynamics. Cambridge Aerospace Series (No. 13), 2nd edn. Cambridge University Press, New York (2001)

[LiEtAl04] Lifanov, I.K., Poltavskiĭ, L.N., Vaĭnikko G.M.: Hypersingular Integral Equations and Their Applications. Chapman & Hall/CRC, Boca Raton (2004)

[LiEtAl92] Lifanov, I.K., Matveev, A.F., Molyakov, N.M.: Flow around permeable and thick airfoils and numerical solution of singular integral equations. Russ. J. Numer. Anal. Math. Modell. **4**, 109–144 (1992)

[Se16] Setukha, A.V., Yukhman, D.A.: On the solvability of a boundary value problem for the Laplace equation on a screen with a boundary condition for a directional derivative. Differ. Equ. **52**, 1188–1198 (2016)

Chapter 24
Performance Assessment of a New FFT Based High Impedance Fault Detection Scheme

A. Soheili and J. Sadeh

24.1 Introduction

Single-line to ground faults are of the most common problems the power system distribution faces on a daily basis. A substantial proportion of these faults are labeled with a high impedance nature. When the cable comes in contact with a high impedance object, i.e., tree branch, asphalt, etc., current magnitudes have shown sudden drop and erratic fluctuations, resulting in typical over-current (OC) relays failure to effectively detect faulty conditions. Asymmetry, nonlinearity, and randomness in nature are also additional features of HIFs, due to the presence of electrical arcs. Fire hazards and electrical shocks are the main concern regarding HIFs. It should be noted that HIF protection schemes are designed with a different goal in mind. Common short circuit faults produce high rated currents where power system equipment may face irreversible repercussions, hence, over-current relays are addressed to prevent this situation. On the other hand, HIFs result in current amplitudes below rated values and cannot harm the nearby installed equipment, therefore, HIF detection schemes are engineered for different purposes.

Researchers have focused on finding unique HIF identifiers based on its characteristics. Mechanical attempts on HIF detection involved the installation of low resistance cables under distribution feeders, designed to create high current magnitudes upon contact. Extreme financial costs and low detection rates were the main reasons for the downfall of this solution. Contrary to mechanical methods, electrical applications showed promising future, due to the development of technology and mathematical tools.

A. Soheili • J. Sadeh (✉)
Faculty of Engineering, Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: soheiliadel@stu.um.ac.ir; sadeh@um.ac.ir

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_24

Initial time-domain options employed the use of current and voltage amplitudes, rate of change, etc., as means of detecting abnormality in power system distribution. However, false detection against typical Switching Events (SEs) such as Capacitor Bank Switching (CBS) has led to the distribution operator's distrust in available schemes. Mathematical Morphology (MM) has been proposed by Gautam as a signal processing tool, being capable of distinguishing between HIF and SE occurrences [GaEtAl12]. Russell et al. used various controlled experiments and simulation to study the features of HIFs and accompanying electrical arcs [KiEtAl88]. Based on his published findings, HIFs produce higher than average odd harmonics with a noticeable elevation in the third harmonics, since they increase system's nonlinearity. Milioudis et al. have recently shown similarity between HIFs and fluorescent lamps [MiEtAl15]. Based on studied waveforms, electrical arcs have shown to have varying ignition and extinguishing times with an approximately quarter-cycle life span. The use of odd, even, third, second, and inter-harmonics has been proposed as methods of the frequency-domain detection schemes [MaEtAl15]. Fast Fourier Transform (FFT) and Short Time Fourier Transform (STFT) are the main mathematical tools in this branch. The unavailability of infinite data and semi-periodic conditions of waveforms raises some doubts. However, high detection rates, lower computational burdens, and easy implementation compensate for its losses.

The Wavelet Transform (WT) has been proposed as a suitable tool for detecting small deviations in current waveforms [MiEtAl06, SeEtAl05]. Among its known benefits, the immunity in regard to unbalanced systems and the ability to detect SE conditions have been emphasized. That being said, dependency towards the selection of mother wavelet and network topology has created some concerns. The Stackwell Transform (ST), as a superior time-frequency technique, has been proposed to mask WT's vulnerability towards noise [RoEtAl15]. Hybrid and combinational solutions have also been presented as recent attempts for HIF detection [BaEtAl16]. Furthermore, mathematical tools have been utilized as noise reduction, feature selection, and extraction techniques. Heuristic approaches, such as Artificial Neural Network (ANN) and Genetic Algorithm (GA), are used as classification instruments [LiEtAl16, EbEtAl90]. The need for large reliable data bank for learning and dependency towards network topology are the main setbacks of these solutions. That being said, they uphold the highest detection rates among presented schemes.

In this paper, the performance of a detection scheme, previously published by authors, has been thoroughly investigated. The techniques use the combinational behavior of even, odd, third, and second harmonics to successfully detect and distinguish between HIF, CBS, and Motor Switching (MS) events. For comparison purposes, representatives from time-domain, frequency-domain, and time-scale domain approaches have been chosen. The remainder of this paper is organized as follows. Section II briefly introduces the five considered detection schemes with appropriate mathematical expressions. Subsequently, section III introduces the designated scenarios, simulation results, and an introduction to the IEEE case study. The paper concludes in section IV with a brief summary of the highlighted results.

24.2 Introduction to HIF Detection Schemes

In order to fully assess the performance of the previously proposed STFT based technique, the appointed scheme along with 4 competitors has been chosen for comparison. In this section, a brief introduction to each method under surveillance has been presented. It should be noted that since heuristic approaches tend to be network specific and do not match the computation time-line of the proposed method, they have been neglected from comparison.

1. Ratio Ground (RG)
   The RG relay was designed as an electromechanical relay, where active and restrain springs move about depending on the positive and zero sequence current. According to the definition provided by [LeEtAl83], the ratio of the zero sequence to the positive sequence has been set as the main criteria in this approach. Researchers have reported a high detection rate and the most stable output in its time. A value of 20% has been reported normal for distribution systems with small levels of unbalance load distribution [LeEtAl83]. Therefore, the algorithm will announce HIF conditions, provided that this criteria has been breached.
2. Mathematical Morphology (MM)
   Feature extraction, feature selection, and noise cancelation are the most fancied applications of MM. The high sensitivity towards small and minute deviations in waveforms makes this approach appealing for the detection of HIFs [GaEtAl12]. Gautam et al. proposed the difference between opening and closing operators (CODO signal), as means of detecting HIFs (shown by (24.1)–(24.3)).

$$Y_o(n) = (f \circ g)(n) = ((f \ominus g) \oplus g)(n) \tag{24.1}$$

$$Y_c(n) = (f \bullet g)(n) = ((f \oplus g) \ominus g)(n) \tag{24.2}$$

$$Y_{CODO}(n) = Y_c(n) - Y_o(n) \tag{24.3}$$

   Due to the nonlinear and asymmetric behavior of HIF, the rapid ignition and extinguishment of electric arcs result in small deviations in voltage and current waveforms. Since measuring equipment is commonly installed at the beginning of the main branch, current waveforms often witness damping if the distance between fault and measuring equipment is noticeable or multiple CT and PTs are present. That being said, voltage waveforms have shown little effect in this matter; hence, the voltage waveform is used in this case. In the duration HIFs are present, rapid close spikes with variations in magnitude have been recorded at CODO output, whilst capacitor bank switching will only cause temporary spikes at the inception time and nothing afterwards. Figure 24.1 illustrates an exemplary CODO outputs for a HIF phenomenon.

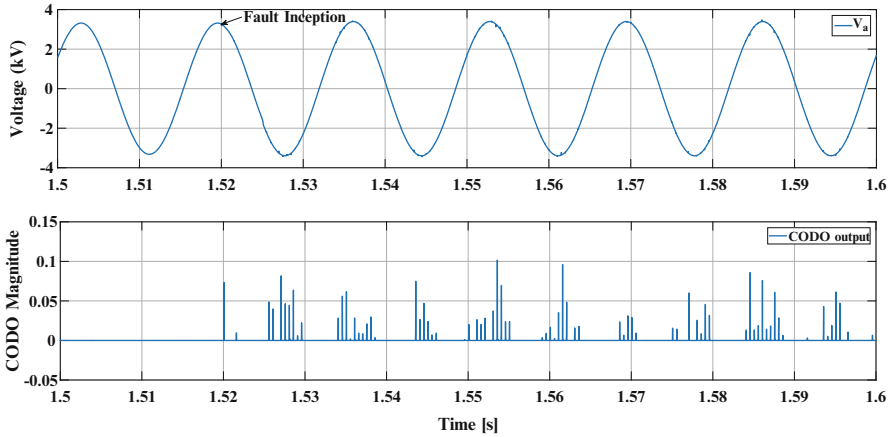


Fig. 24.1 Presence of HIF and corresponding CODO output signal

3. Combination of Odd and Even Harmonics (HA1)
   The third chosen method employs the frequency spectrum and the even and odd harmonics in specific to investigate the network condition in regard to the presence of HIFs. The method proposed by Torres et al. has been considered for this section [ToEtAl14]. According to the presented method, current THD is compared with average 3phase THD, provided that the third harmonic is higher than the sum of even harmonics and smaller than the odd harmonics for a specific period of time. Subsequently, if the phase THD is higher than the 3phase average THD, the algorithm will declare HIF as the feeder status. The main downside of this method is that SEs are not designed to be noticed. For future references, this method has been labeled as HA1.
4. Standard Deviation Criteria Wavelet Transform (WT)
   As mentioned before, WT is widely used for feature extraction applications. Chen et al. have shown that the order of magnitude changes for the standard deviation of specific WT detail layers [ChEtAl14]. Hence, the criteria shown by (24.4) was engineered for the purpose of detecting HIFs and SEs circumstances.

$$\delta_{D_i} = ln(std(D_i)) + 14 \tag{24.4}$$

   where i denotes the chosen WT detail layer. The natural logarithm of the standard deviation of the 2nd and 3rd detail layer has been set as suitable detection criteria. According to the presented paper, provided that both criteria are below 5, between 5 and 10, and higher than 10, the appointed scheme will announce Normal Conditions (NCs), SEs, and HIFs, respectively.

5. STFT Based Detection Scheme
   The STFT based method, proposed by authors [SoEtAl16], employs the unique relation of combining various harmonic behaviors. "Analyzer of Abnormal Conditions" and "HIF Signature Detector" are the two main sections of this method, which work based on the even, second and third harmonic. Simulations have shown immunity towards current magnitude and successful detection of HIF, CBS, and MS events. Consequently, depending whether a HIF or SE has occurred, the second or third harmonic is noticeably higher than the other, respectively. By doing so, the proposed scheme is immune to current magnitude oscillations and can effectively detect HIFs, MS, and CBS. Hence, the implementation of this method will undoubtedly result in invaluable virtues. Further information regarding detailed calculations are presented in [SoEtAl16]. In order to differentiate between the frequency-domain detection schemes, this method has been labeled STFT.

24.3 Simulation Results

The performance evaluation process has been carried out via several scenarios involving various fault conditions and switching events using PSCAD and MATLAB programming software. For comparison purposes, all scenarios have been maintained with identical conditions in regard to event parameters such as time of occurrence and position at feeder. It should be mentioned that simulations have been executed by a desktop computer powered by Intel©core i7 4570 with 16GB of RAM. The IEEE 13-Node distribution system is a 4.16kV short and heavily unbalanced standard test system, which offers suitable conditions for testing the proposed method's performance. The distribution grid contains various high current 3phase, low current 3phase extensions, and low current single phase extensions. Additionally, these feeders are both underground and overhead type feeders. Further information regarding the IEEE network can be found in [IEEEWeb]. Measuring equipment is located at the primary node (650), in order to maximize the likelihood to real world conditions. Figure 24.2 shows the IEEE 13-Node distribution test system layout.
For simulation purposes, the output of each method has been designated with a value in order to evaluate its performance in the considered duration of time. Therefore, for all the five abovementioned methods, the values 1, 0.75, and 0 have been assigned for statuses of HIF, SE, and NC, respectively. Also, the relays are designed to show values of 0.5 in the "pickup" process. However, some methods, such as the RG and WT, do not have this particular ability. Various HIF and switching events have been studied in scenario I and II, respectively.

A. Scenario I
   The first scenario focuses on the performance regarding the HIF detection. In this section, both 3phase and single phase feeders have been considered for

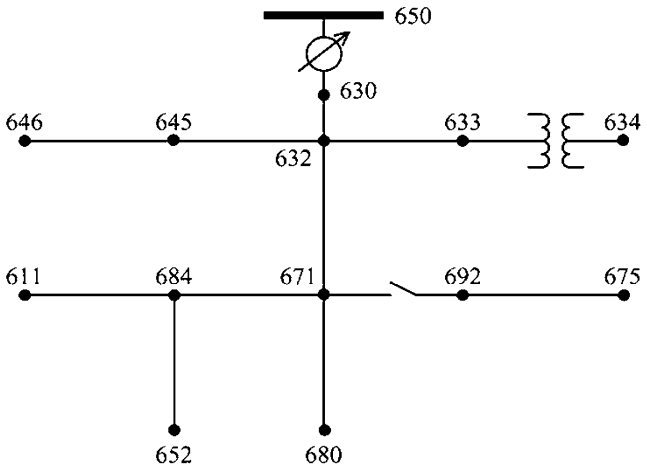


Fig. 24.2 IEEE 13-Node distribution system [IEEEWeb]

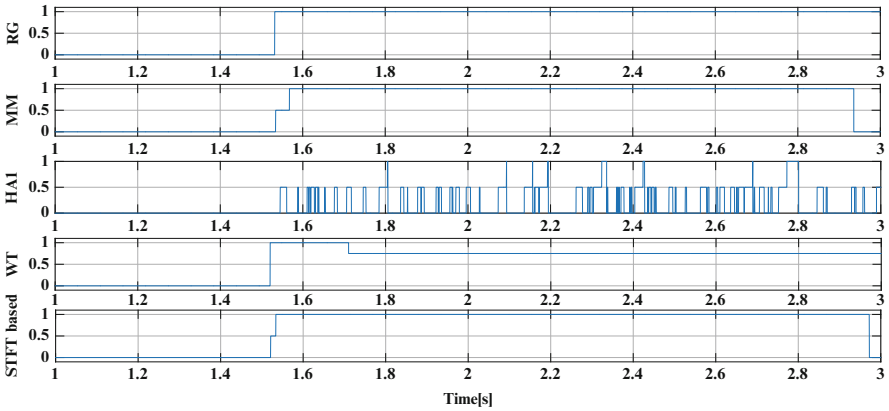


Fig. 24.3 Performance evaluation regarding HIF at feeder 650–632

performance evaluation. The feeders 650–632 and 632–633 represent the 3phase high current and low current extensions. Figures 24.3 and 24.4 illustrate the recorded response from the previously five mentioned HIF detection schemes. Putting aside the HA1 failed attempts and initial WT false detection, the remaining 3 chosen methods have successfully detected the HIF fault conditions. The 650–632 is a main branch feeder with rated current of about 400A, hence the presence of HIF would exponentially decrease the current. The RG method relies on this aspect which has led to a successful detection. The proposed method by authors, similarly to the MM, has been able to effectively detect HIF conditions.

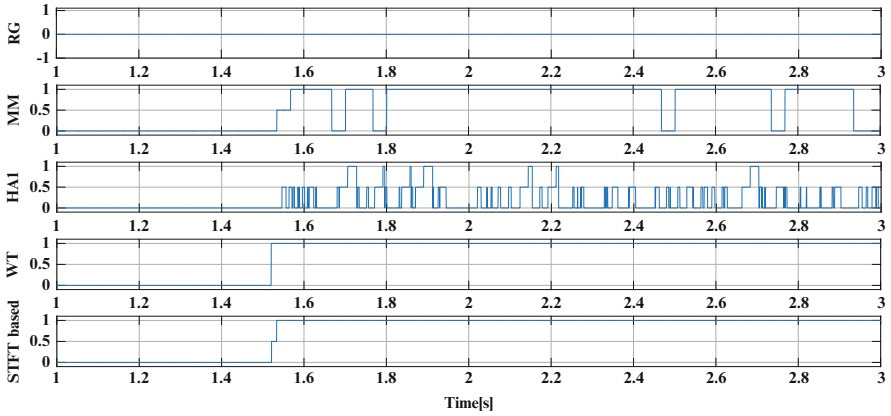


Fig. 24.4 Performance evaluation regarding HIF at feeder 632–633

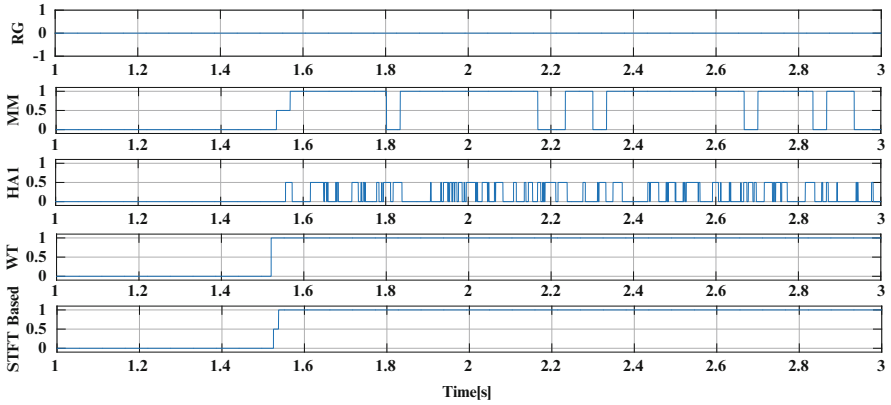


Fig. 24.5 Performance evaluation regarding HIF at feeder 684–652

Focusing on Figure 24.4, related to feeder 632–633 with rated current of 30A, only STFT and WT have been able to detect the HIF conditions without any doubts. On the contrary, since current amplitude has not decreased dramatically, the ratio of sequences has not varied significantly and hence, RG was not been able to detect any problems. Despite the short time resets, MM has been able to show HIF circumstances for most portions of time. The small gaps in-between is due to the relatively wider distance between spikes which in a window based movement, the MM detection algorithm mistakes it for fault clearance. The HA1 approach did acknowledge the presence of abnormal conditions by rapidly showing values of 0.5, however, declaring the HIF state has only briefly been shown. Generally, what concerns distribution operators is the occurrence of HIF on single phase extensions where the low current magnitude hinders the detection process. Figure 24.5 presents the detection results for feeder

684–652. What stands out from this figure is that STFT and WT have been able to effectively present definitive results of HIF conditions. RG and HA1 have shown results similar to the 3phase low current extensions case. In addition, MM has shown higher fluctuations between HIF and NC on feeders with lower magnitudes.

According to the presented data in this section, the RG relay is solely able to detect HIFs at the beginning of the distribution feeder and fails to show any change while the fault is located at lower current feeders. HA1 has completely failed to show acceptable detection rates. The MM algorithm presented acceptable detection rates in low current amplitudes, but higher oscillations have been witnessed at these states. Contrary to the MM, WT has shown higher detection rates at low current magnitudes. The presented STFT based algorithm has been able to successfully detect all HIF conditions, regardless of the current amplitude and fault position.

B. Scenario II

Generally, SEs have shown similar transients to HIF, where some detection schemes mistakenly react and hinder the distribution of power among consumers. Distribution operators do not fancy these "trigger happy" detection plans. Three of the chosen methods have been engineered to sense and distinguish SEs from HIFs, whereas RG and HA1 are only capable of detecting HIFs. Capacitor bank has been modeled using a star connected 250kVAr capacitor bank in order to rise power factor from 0.83 to 0.94, while a typical 700hp synchronous squirrel cage motor has been selected for motor switching events. Figure 24.6 demonstrates the performance of the selected five methods against CBS at node 632.

It can be seen that MM and STFT have been successful in labeling the abnormal conditions as SEs, whereas RG and HA1 failed to notice anything out of the ordinary. Also, further investigation shown that WT presented acceptable

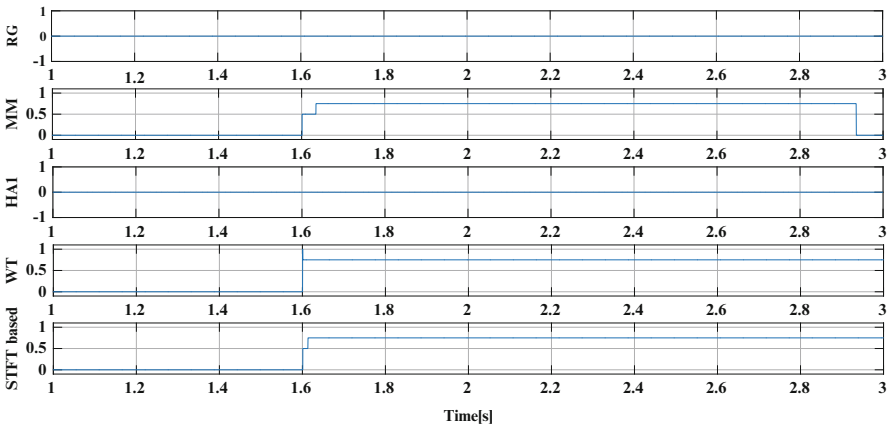


Fig. 24.6 Performance comparison due to capacitor bank switching at node 632

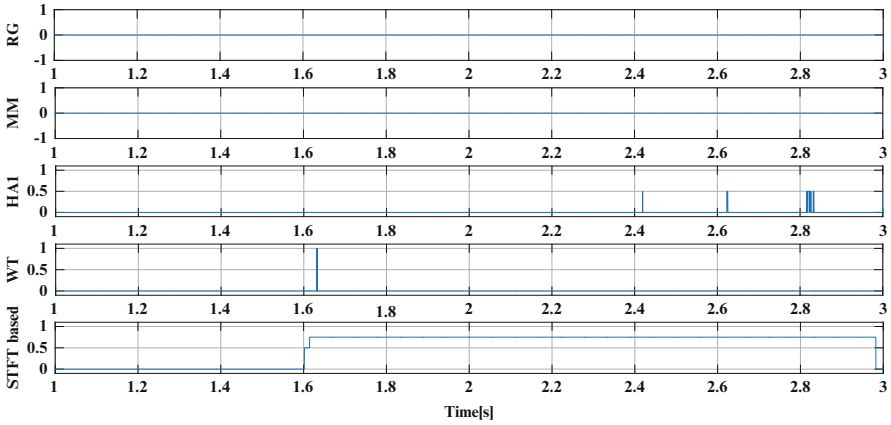


Fig. 24.7 Performance comparison due to motor switching at node 675

behavior during the transients of node 632, whilst at nodes at a farther distance, i.e., 675, 692, outputs have oscillated between CBS and NCs. The increased distance has led to moderate drop in the criterions and hence the output of NC and HIF was presented. The study on MS has also been carried out on nodes 675. Figure 24.7 illustrates the five responses gathered from the designated detection schemes. Despite the majority of chosen methods being capable of detecting switching event transients, it can be seen that only the proposed method is cable of distinguishing between NC and MS events. Further investigations on the presence of CBS and MS have shown the superiority of the proposed method by authors.

24.4 Conclusion

In the presented paper, an in-depth performance evaluation of a previously introduced STFT based HIF detection plan has been carried out via rigorous simulations. For comparison purposes, four time, frequency, and wavelet based approaches have been utilized in this process. Various HIF conditions including variation in fault position and feeder current have been considered. In order to fully cover the evaluation process, capacitor bank and motor switching have also been taken into account. Simulation results have shown the highest detection rate for both HIF and SE detection for the proposed STFT based technique. Dependability towards current amplitude and fault position in regard to the measuring equipment are some of the drawbacks among the four competition schemes. However, the proposed method has shown immunity towards these conditions.

References

[BaEtAl16] Batista, O.E., Flauzino, R.A., De Araujo, M.A., De Moraes L.A., Da Silva, I.N.: Methodology for information extraction from oscillograms and it's application for high-impedance fault analysis. Int. J. Electrical Power Energy Syst. **76**, 23–34 (2016)

[ChEtAl14] Chen, J.C., Phung, B.T., Wu, H.W., Zhang, D.M., Blackburn, T.: Detection of high impedance faults using wavelet transform. Australasian Universities Power Engineering Conference (AUPEC 2014), pp. 1–6. Curtin University, Perth, Australia (2014)

[EbEtAl90] Ebron, S., Lubkeman, D.L., White, M.: A neural network approach to the detection of incipient fault on power distribution feeders. IEEE Trans. Power Deliv. **5**(2), 905–912 (1990)

[GaEtAl12] Gautam, S., Brahma, S.M.: Detection of high impedance fault in power distribution systems using mathematical morphology. IEEE Trans. Power Deliv. **28**(2), 1226–1234 (2012)

[IEEEWeb] IEEE 13-bus: IEEE Power and Energy Society, test feeders [online] (2000) available: http://ewh.ieee.org/soc/pes/dsacom/testfeeders/

[KiEtAl88] Kim, C.J., Russell, B.D.: Harmonic behavior during arcing faults on power distribution feeders. Electric Power Syst. Res. **14**(3), 219–225 (1988)

[LeEtAl83] Lee, R., Bishop, M.: Performance testing of the ratio ground relay on a four-wire distribution feeder. IEEE Trans. Power Apparatus Syst. **102**(9), 2943–2949 (1983)

[LiEtAl16] Li, Y., Meng , X., Song, X.: Application of signal processing and analysis in detecting single line-to-ground (SLG) fault location in high-impedance grounded distribution network. IET Gener. Transm. Distrib. **10**(2), 382–389 (2016)

[MaEtAl15] Macedo, J.R., Resende, J.W., Bissochi, C.A., Carvalho, D., Castro, F.C.: Proposition of an interharmonic-based methodology for high-impedance fault detection in distribution systems. IET Gener. Transm. Distrib. **9**(16), 2593–2601 (2015)

[MiEtAl06] Michalik, M., Rebizant, W., Lukowiez, M., Lee, S.J., Kang, S.H.: High-impedance fault detection in distribution networks with use of wavelet-based algorithm. IEEE Trans. Power Deliv. **21**(4), 1793–1802 (2006)

[MiEtAl15] Milioudis, A.N., Andreoua, G.T., Labridis, D.P.: Detection and location of high impedance faults in multiconductor overhead distribution lines using power line communication devices. IEEE Trans. Smart Grid **6**(2), 894–902 (2015)

[RoEtAl15] Routray, P., Mishra, M., Rout, P.K.: High impedance fault detection in radial distribution system using S-transform and neural network. IEEE Power Communication and Information Technology Conference (PCITC), Bhubaneswar, pp. 545–551 (2015)

[SeEtAl05] Sedighi, A.R., Haghifam, M.R., Malik, O.P.: Soft computing applications in high impedance fault detection in distribution systems. Electric Power Syst. Res. **76**, 136–144 (2005)

[SoEtAl16] Soheili, A., Sadeh, J., Lomei, H., Muttaqi, K.: A new high impedance fault detection scheme: Fourier based approach. 2016 IEEE International Conference on Power System Technology (POWERCON2016), NSW, Australia, pp.1–6 (2016)

[ToEtAl14] Torres, V., Guardado, J.L., Ruiz H.F., Maximov, S.: Modeling and detection of high impedance faults. Int. J. Electrical Power Energ. Syst. **61**, 163–172 (2014)

Chapter 25
$\mathcal{H}^2$ Matrix and Integral Equation for Electromagnetic Scattering by a Perfectly Conducting Object

S.L. Stavtsev

25.1 Introduction

Hypersingular integral equations are applied in various areas of mathematics and technology, such as aerodynamics, filtration, elasticity, diffraction, acoustics, and electromagnetic waves. [LiEtAl04, LiEtAl02, LiEtAl04, St06]. To solve complicated problems described by a large number of parameters using integral equations one should use fine meshes, and the initial problem reduces to the solution of a very large system of linear equations with a dense matrix. Such a matrix is likely to exceed the computer memory capacity. To solve such kind of problems one can use supercomputers with distributed memory, as well as special numerical methods for dense matrix approximations. For example, [ApEtAl10, ApEtAl13, St12] tackle the solution of aerodynamic problems using low-rank approximations of large matrices.

This paper presents parallel algorithms combined with low-rank approximations of dense matrices to solve problems of the electromagnetic wave diffraction on perfectly conducting objects with a complex shape. If the problem is solved on an object with a high wave size, this yields a large dense matrix. A complex shape of the object makes it impossible to apply high-order quadrature formulas, and since the solution of the integral equation for a high frequency diffraction problem has strong oscillations, very fine meshes should be applied to approximate the integral operator with piecewise constant functions.

Short reviews of the matrix approximation methods can be found, for example, in [YoEtAl16, TaEtAl13]. One of the most widely spread matrix approximation methods is the multipole method [CoEtAl93, SoEtAl95]. Apart from the multipole methods there are kernel-independent matrix approximation methods. Low-rank approximation methods were already applied to diffraction problems. For example,

S.L. Stavtsev (✉)
Institute of Numerical Mathematics Russian Academy of Sciences, Moscow, Russia
e-mail: sstass2000@mail.ru

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_25

see [CoEtAl93] for a multipole-based algorithm for the solution of a diffraction problem. Kernel-independent methods (mosaic-skeleton approximations) were also applied to this problem.

Low-rank approximation methods allow to approximate matrices of size $N \times N$ and compute the matrix-vector multiplication in $O(N \log(N))$ instead of $O(N^2)$ operations.

Iterative methods, for example, GMRES [SaEtAl86], are used routinely for the solution of linear systems of algebraic equations with large dense matrices. GMRES employs only multiplication of a low-rank matrix by a vector and does not involve any other operations with the low-rank matrix. Since the parallel matrix-vector multiplication with a matrix in the mosaic-skeleton format scales well, the whole algorithm of the linear system solution with a matrix in the mosaic-skeleton format is also well scalable.

However, the number of GMRES iterations increases rapidly with the wave size of the object. Therefore, for large wave sizes GMRES requires a lot of memory. Even more memory demanding is the problem of solving the linear system with many right-hand sides, which arises in the computation of the inverse Radar Cross Section (RCS) characteristic.

To reduce the number of GMRES iterations one can use preconditioners. For example, in [St15] an effective preconditioner is constructed, but its parallel version is poorly scalable. Moreover, experiments of applying this preconditioner to the electrodynamics problem have shown that the number of iterations can be reduced only if the inverse matrix is approximated very accurately. This means that direct solvers must be used for this problem. An example of a direct solver can be found in [CoEtAl15]. This method is well scalable, but unfortunately it cannot be applied to the diffraction problem.

This paper constructs a parallel direct solver for a low-rank matrix that arises from the electrodynamics problem. As a low-rank matrix format we use the $\mathcal{H}^2$ representation [Ha15], a kernel-independent MultiLevel Fast Multipole Algorithm (MLFMA). Unfortunately, contrarily to the Fast Multipole Method (FMM), parallel MLFMA and $\mathcal{H}^2$ matrix construction algorithms have poor scalability. In this paper we develop well scalable parallel algorithms for calculating the $\mathcal{H}^2$ matrix and solving the linear system with the $\mathcal{H}^2$ matrix by a direct parallel solver.

25.2 Electrodynamics Problem and Integral Equation

Let us consider the diffraction problem on a perfectly conducting surface Σ, which can be either closed or open.

A monochrome wave with a frequency ω satisfies the Maxwell equations,

$$\nabla \times \vec{E} = i\mu\mu_0\omega\vec{H}; \nabla \times \vec{H} = -i\varepsilon\varepsilon_0\omega\vec{E}.$$

On a perfectly conducting surface the following boundary condition holds:

$$\vec{n} \times (\vec{E}_0 + \vec{E}) = 0,$$

where $\vec{E}_0$ is a given function, defined by the incident wave (we assume that the incident wave is planar), and $\vec{n}$ is a normal vector to the surface.

To find a unique solution it is necessary to pose additional conditions

$$\vec{E} \in L_2^{\text{loc}}(\Omega)$$

and

$$\frac{d}{d\tau}\begin{pmatrix}\vec{E}\\ \vec{H}\end{pmatrix} - ik\begin{pmatrix}\vec{E}\\ \vec{H}\end{pmatrix} = o\left(\frac{1}{|\vec{x}|}\right),\ \tau = |\vec{x}|,\ \tau \to \infty.$$

In accordance with [CoEtAl83], the problem can be reduced to the electric field integral equation on the unknown $\vec{j}(y)$:

$$\vec{n} \times \iint\limits_{\Sigma} \vec{j}(y)\Big(\text{grad}\,\text{div}F(x-y) + k^2F(x-y)\Big)d\sigma_y = -\vec{n} \times \vec{E}_0(x),\ x \in \Sigma, \tag{25.1}$$

where $k = \omega\sqrt{\varepsilon\varepsilon_0\mu\mu_0}$ is the wave number, and

$$F(R) = \frac{\exp(ikR)}{R},\ R = |x-y|.$$

In Equation (25.1) the integral can be understood in the sense of the Hadamard finite part.

For the numerical solution of the Equation (25.1) we use a numerical scheme presented in [LiEtAl02]. In this scheme the surface is uniformly divided into cells σ_i, $i = 1, n$, and for each cell an orthonormal basis $\vec{e}_{i1}$, $\vec{e}_{i2}$ is introduced. For each cell σ_i it is assumed that $\vec{j}_i = \vec{j}(x_i)$, where x_i is the center of mass of the cell. Each cell is considered to be planar. Discretization of the integral operator produces a matrix that consists of 2×2 blocks:

$$A_{ij} = \begin{pmatrix}\vec{E}_{1j}(x_i)\cdot\vec{e}_{1i} & \vec{E}_{2j}(x_i)\cdot\vec{e}_{1i}\\ \vec{E}_{1j}(x_i)\cdot\vec{e}_{2i} & \vec{E}_{2j}(x_i)\cdot\vec{e}_{2i}\end{pmatrix},$$

$$\vec{E}_{1j}(x_i) = \int\limits_{\partial\sigma_j} \vec{Q}(x_i)de_2 + k^2\vec{e}_{1j}\int\limits_{\sigma_j}\frac{\exp(ikR)}{R}d\sigma; \tag{25.2}$$

$$\vec{E}_{2j}(x_i) = -\int\limits_{\partial\sigma_j} \vec{Q}(x_i)de_1 + k^2\vec{e}_{2j}\int\limits_{\sigma_j}\frac{\exp(ikR)}{R}d\sigma,\ \vec{Q}(x) = \nabla_y\frac{\exp(ik|x-y|)}{|x-y|}.$$

In (25.2) the contour and surface integrals are calculated numerically.

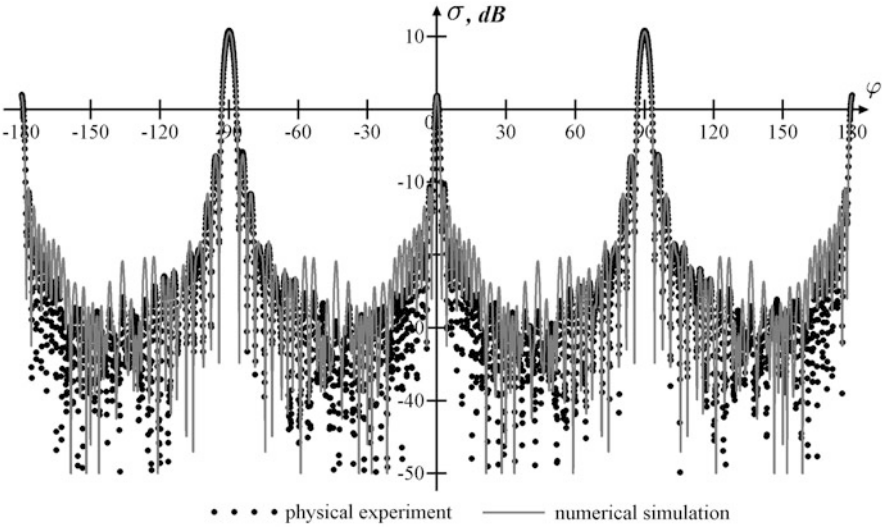


Fig. 25.1 RCS, 16GHz, vertical polarization, $n = 192156$

25.3 Mosaic-Skeleton Approximations

The problem reduces to the solution of the linear system of algebraic equations

$$A\xi = \beta \tag{25.3}$$

with a dense matrix A. To approximate the matrix we use the mosaic-skeleton method [Ty00, StEtAl09]. It partitions the matrix hierarchically into blocks, and the low-rank matrix blocks can be calculated independently using the incomplete cross approximation algorithm.

Let us investigate the approximation algorithm. In all examples below the surface Σ in Equation (25.1) is a round cylinder with the diameter 15cm and height 25cm.

In Figure 25.1 we present the inverse RCS for the frequency 16GHz. The σ value for different directions τ of the wave vectors of the incident wave is calculated as

$$\sigma(\tau) = \frac{4\pi}{|\vec{E}_0|^2} \left| \sum_{i=1}^{n} \left(\vec{j}_i - \tau \cdot (\tau \cdot \vec{j}_i) \right) k^2 \exp\left(-ik\tau \cdot x_i\right) \sigma_i \right|^2 . \tag{25.4}$$

Black points show the results of the experiment, the grey line shows the results of the numerical simulation.

In all calculations the number of cells is 192156, the number of right-hand sides is 2048, the approximation and solution accuracies are 10^{-3}.

Table 25.1 Number of iterations for various parameters of the electrodynamics problem

n	2GHz	4GHz	8GHz	16GHz
7872	1862	2355	4390	9410
21760	2821	4261	6025	11237
30400	3651	4791	7285	12990
45784	4262	5689	8269	21103

In Table 25.1 one can see the number of iterations needed to solve the system with 2048 right-hand sides up to the accuracy $5 \cdot 10^{-3}$ for different frequencies and numbers of cells n.

It can be seen from Table 25.1 that the number of iterations increases significantly with the frequency, and it requires a lot of memory and computational time.

So, let us apply the $\mathcal{H}^2$ matrix representation [Ha15] to solve the system.

25.4 Algorithm for Calculation of a $\mathcal{H}^2$ Matrix

The mosaic-skeleton approximation algorithm is well scalable on multiprocessor computing systems. The $\mathcal{H}^2$ matrix approach results in a better matrix compression rate, but the algorithms for constructing an $\mathcal{H}^2$ matrix, presented in [Ha15, MiEtAl16], have poorer scalability. We develop a parallel algorithm based on the software package [Mi16].

Following the notation from [CoEtAl15], an $\mathcal{H}^2$ matrix can be represented as follows:

$$A = D_0 + L_1\Big(D_1 + L_2(D_2 + \cdots)R_2\Big)R_1, \tag{25.5}$$

where matrices L_k and R_k are

$$L_k = \begin{pmatrix} L_{k1} & 0 & 0 & \dots & 0 \\ 0 & L_{k2} & 0 & \dots & 0 \\ 0 & 0 & L_{k3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & L_{km} \end{pmatrix}; \; R_k = \begin{pmatrix} R_{k1} & 0 & 0 & \dots & 0 \\ 0 & R_{k2} & 0 & \dots & 0 \\ 0 & 0 & R_{k3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & R_{kn} \end{pmatrix}, \tag{25.6}$$

where the blocks L_{kp} are defined by the row factors of a node p at the level k of the tree of splitting of the calculation points into blocks. In turn, R_{kq} are defined by the column factors of q at the level k of the tree, defining the splitting of the array of cells into blocks.

Example of an $\mathcal{H}^2$ matrix split into blocks (Figure 25.2) is presented in Figure 25.3. Due to the low-rank approximation, the matrices D_0, D_k, L_k, R_k, $k = 1, 2, \dots$, are represented by much less data than the original matrix A.

Fig. 25.2 Matrix blocks

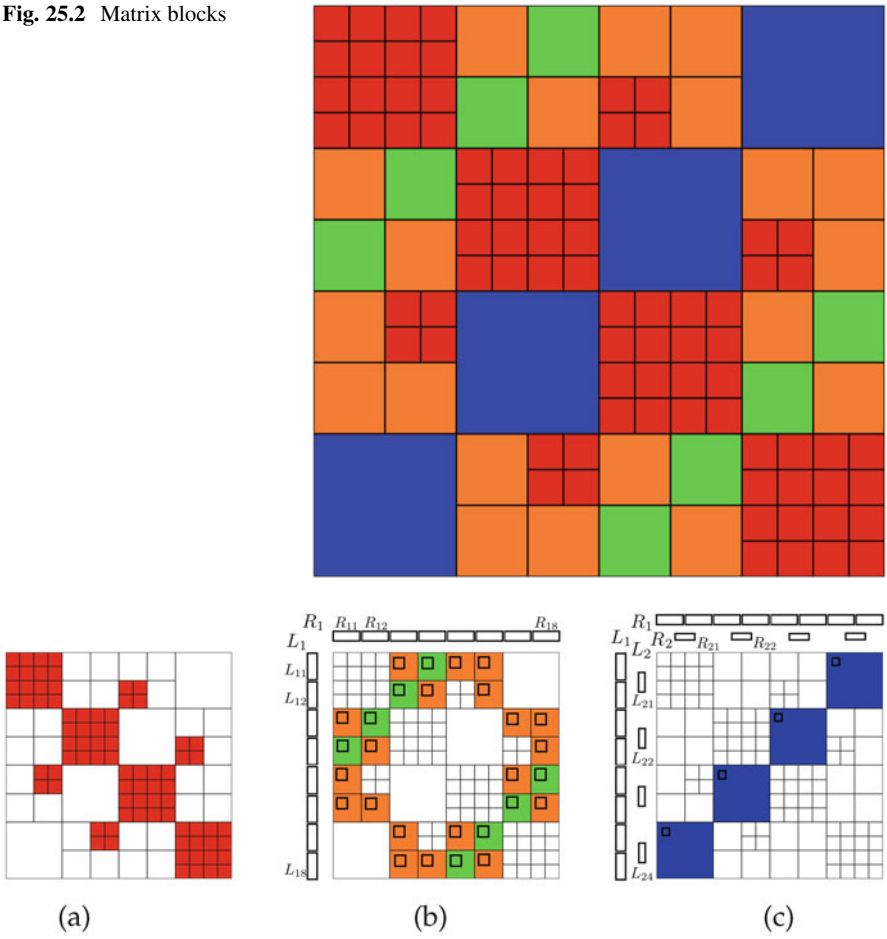


Fig. 25.3 **a**) matrix D_0, **b**) matrices L_1, R_1, D_1, **c**) matrices L_2, R_2, D_2

We have modified the $\mathcal{H}^2$ matrix construction algorithm from [Mi16] as follows. First, we construct all row basis factors, i.e., matrices L_k, $k = 1, 2, \ldots$ from decomposition (25.5). Then we carry out interprocessor data communications using MPI and calculate the column basis factors R_k, $k = 1, 2, \ldots$ and matrices D_k, $k = 1, 2, \ldots$. Second, if the number of processors exceeds significantly the number of blocks on the root node of the tree, defining the splitting of the matrix into blocks, then the blocks of the upper level are additionally partitioned into subblocks. For example, in Figure 25.4 the number of blocks on the upper level is 4 for both rows and columns (they marked with blue color). If there are more than 4 processors, for example, 8, then each block is partitioned into 4 subblocks, as shown in Figure 25.4. This kind of additional splitting reduces the number of interprocessor communications in the course of calculating the matrices L_k and R_k, $k = 1, 2, \ldots$.

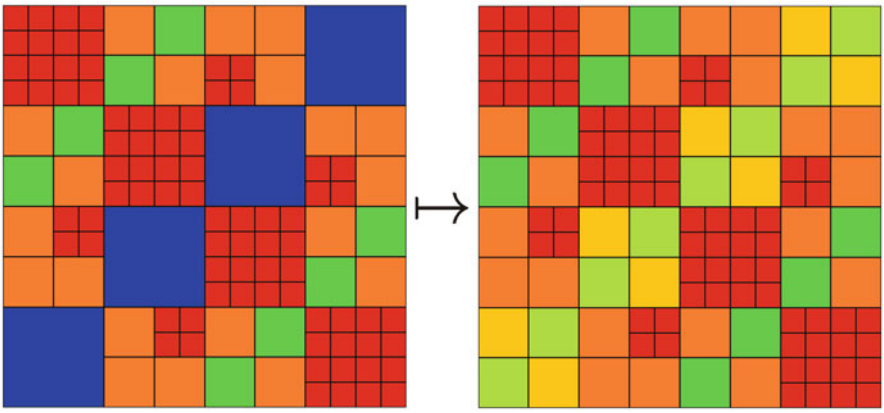


Fig. 25.4 Additional partitioning of the blocks

Table 25.2 Scalability of different algorithms for low-rank approximation construction

n_p	1	2	4	8	16	32	64	128
MS	1	1.91	3.61	7.24	13.38	24.51	41.31	51.84
$\mathcal{H}^2$	1	1.88	3.51	7.15	10.26	16.27	23.17	32.86

The scalability results for the electrodynamics problem on a cylinder with 125594 cells, the frequency 8GHz, and the approximation accuracy 10^{-2} are shown in Table 25.2. Here, n_p is the number of processors. Table 25.2 shows speedups of calculation of a low-rank matrix in the mosaic-skeleton (*MS*) and $\mathcal{H}^2$ matrix formats for various numbers of processors. All calculations have been run on Intel Xeon E5-2670v3 2.30Ghz CPUs at the INM RAS (http://cluster2.inm.ras.ru/) cluster. The code was compiled with the Intel Fortran Compiler 9.0 for Linux (9.0.033).

Let us now consider the parallel algorithm for solving a linear system with an $\mathcal{H}^2$ matrix.

25.5 Direct Solver for Systems with $\mathcal{H}^2$ Matrices

Let us consider the solution of the system (25.3) with the matrix (25.5).

To construct a sparse extended system we introduce additional variables

$$\begin{aligned} &\varphi_1 = R_1\xi,\ \varphi_k = R_k\varphi_{k-1}, k = 2,\dots,p, \\ &u_k = D_k\varphi_k + L_{k+1}u_{k+1}, k = 1,\dots,p-1, u_p = D_p\varphi_p, \end{aligned} \tag{25.7}$$

where p is the number of levels in (25.5).

From (25.5), (25.3), and (25.7) it follows that

$$D_0\xi + L_1 u_1 = \beta. \tag{25.8}$$

To solve the system (25.8) and find auxiliary variables u_k, φ_k we need to solve the following system:

$$\begin{pmatrix} D_0 & L_1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -R_1 & 0 & I & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -I & D_1 & L_2 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & -R_2 & 0 & I & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & -I & D_2 & L_3 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & D_{p-1} & L_p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & -R_p & 0 & I \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & -I & D_p \end{pmatrix} \begin{pmatrix} \xi \\ u_1 \\ \varphi_1 \\ u_2 \\ \varphi_2 \\ \vdots \\ \varphi_{p-1} \\ u_p \\ \varphi_p \end{pmatrix} = \begin{pmatrix} \beta \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{25.9}$$

where I is a square identity matrix. The matrix of the system (25.9) is block-tridiagonal. Therefore, we suggest to use the following algorithm to find ξ:

Let $\hat{A}_1 = A_0^{-1}$
DO $k = 1, \dots, p$
  $S_k := R_k \hat{A}_k L_k$
  $Q_k := (I + A_k S_k)^{-1}$, $k < p$
  $\hat{A}_{k+1} := S_k Q_k$, $k < p$
END DO
$P_p := (S_p A_p + I)^{-1}$, $\hat{H}_p := A_p P_p$
DO $k = p-1, \dots, 1$
  $P_k := (S_k A_k + I)^{-1}$
  $\hat{H}_k := A_k P_k + Q_k L_{k+1} \hat{H}_{k+1} R_{k+1} P_k$
  $\hat{A}_{k+1} := S_k Q_k$, $k < p$
END DO
The solution of the system can be found as

$$\xi = B\beta, \; B = \hat{A}_1 - \hat{A}_1 L_1 \hat{H}_1 R_1 \hat{A}_1. \tag{25.10}$$

As the diagonal blocks of the matrices L_k and R_k (25.6) are stored on different processors, it is easy to multiply L_k and R_k with other matrices. The main difficulty of the algorithm is a parallel computation of the inverse matrices. When p is large, the sizes of the matrices to be inverted are small. In this case we use ScaLapack to calculate them. When p is small, in particular, in the computation of $\hat{A}_1$, the matrices subject to the inversion are sparse. In this case we use MUMPS [MUMPS]. Both MUMPS and ScaLapack use the MPI library.

Table 25.3 shows the speedups of the parallel direct solver for various numbers of processors. The matrix is of size 251904, the frequency is 8 GHz.

Table 25.3 Scalability of the parallel direct solver

n_p	1	2	4	8	16	32	64
	1	1.52	2.42	3.58	6.50	9.15	11.45

To solve the system with a matrix of size 251904 with the GMRES method without a preconditioner required 157.8 GB of additional memory. This additional amount does not include the memory necessary to store a low-rank matrix in RAM. Instead, 157.8 GB is the memory that was needed to store the Krylov subspace basis in the GMRES method. As the preconditioner construction algorithm does not work on distributed memory computers, it was impossible to solve the same problem with the preconditioned GMRES. The memory necessary to store the preconditioner exceeds 64 GB, so it could hardly be done on a single processor. The direct solver with the $\mathcal{H}^2$ matrix needed only 13.42 GB of extra memory. The $\mathcal{H}^2$ matrix of the system occupies 3.28 GB of memory. So, the developed direct solver based on $\mathcal{H}^2$ matrices requires significantly less memory than other known methods for solving such kind of problems.

Acknowledgements This work was supported by the Russian Science Foundation, grant no. 14-11-00806.

References

[ApEtAl10] Aparinov, A.A., Setukha, A.V.: Application of mosaic-skeleton approximations in the simulation of three-dimensional vortex flow by vortex segments Comput. Math. Math. Phys. **50**(5), 890–899 (2010)

[ApEtAl13] Aparinov, A.A., Setukha, A.V.: Parallelization in the vortex method for solving aerodynamic problems Numer. Methods Program. **14**, 406–418 (2013) (in Russian)

[CoEtAl15] Corona, E., Martinsson, P.-G., Zorin, D.: An O(N) direct solver for integral equations on the plane. Appl. Comput. Harmon. Anal. **38**(2), 284–317 (2015)

[CoEtAl83] Colton, D., Kress, R.: Integral Methods in Scattering Theory. Willey, New York (1983)

[CoEtAl93] Coifman, R., Rokhlin, V., Wanzura, S.: The fast multipole method for the wave equation: a pedestrian prescription. IEEE Antennas Propag. Mag. **35**(3), 7–12 (1993)

[Ha15] Hackbusch, W.: Hierarchical Matrices: Algorithms and Analysis. Springer Series in Computational Mathematics. Springer, New York (2015)

[LiEtAl02] Lifanov, I.K., Stavtsev, S.L., Piven, V.F.: Mathematical modelling of the three-dimensional boundary value problem of the discharge of the well system in a homogeneous layer. Russ. J. Numer. Anal. Math. Model. **17**(1), 99–111 (2002)

[LiEtAl02] Lifanov, I.K., Petrov, D.Y.: Modifikatsiya metoda diskretnykh ramok k raschetu nekotorykh prostranstvennykh zadach difrakcii elektromagnitnykh voln. Elektromagnitnye Volny i Electronnie Systemy 7(7), 4–9 (2002) (in Russian)

[LiEtAl04] Lifanov, I.K., Poltavskii, L.N., Vainikko, M.G.M.: Hypersingular integral equations and their applications. Chapman & Hall/CRC 406, Boca Raton (2004)

[LiEtAl04] Lifanov, I.K., Stavtsev, S.L.: Integral equations and sound propagation in a shallow sea. Differ. Equ. **40**(9) 1330–1344 (2004)

[Mi16] https://bitbucket.org/muxas/h2tools

[MiEtAl16] Mikhalev, A.Y., Oseledets, I.V.: Iterative representing set selection for nested cross approximation. Numer. Linear Algebra Appl. **23**(2), 230–248 (2016)

[MUMPS] http://graal.ens-lyon.fr/MUMPS

[SaEtAl86] Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)

[SoEtAl95] Song, J.M., Chew, W.C.: Multilevel fast multipole algorithm for solving combined field integral equations of electromagnetic scattering. Microw. Opt. Technol. Lett. **10**, 14–19 (1995)

[St06] Stavtsev, S.L.: An iterative approach to the numerical solution of the system of integral equations for boundary value problems for the scalar Helmholtz equation. Differ. Equ. **42**(9), 1352–1360 (2006)

[St12] Stavtsev, S.L.: Application of the method of incomplete cross approximation to a nonstationary problem of vortex rings dynamics. Russ. J. Numer. Anal. Math. Model. **27**(3), 303–320 (2012)

[St15] Stavtsev, S.L.: Block LU preconditioner for the electric field integral equation. In Progress in Electromagnetics Research Symposium, pp. 1523–1527. Prague (2015)

[StEtAl09] Stavtsev, S.L., Tyrtyshnikov, E.E.: Application of mosaic-skeleton approximations for solving EFIE. In: Progress in Electromagnetics Research Symposium, pp. 1752–1755. Moscow (2009)

[TaEtAl13] Taboada, J.M., Araujo, M.G., Basteiro, F.O., Rodriguez, J.L., Landesa, L.: MLFMA-FFT parallel algorithm for the solution of extremely large problems in electromagnetics. Proc. IEEE **101**(2), 350–363 (2013)

[Ty00] Tyrtyshnikov, E.E.: Incomplete cross approximation in the mosaic skeleton method. Computing **64**(4), 367–380 (2000)

[YoEtAl16] Yokota, R., Ibeid, H., Keyes, D.: Fast multipole methods as a matrix-free hierarchical low-rank approximation (2016). https://arxiv.org/abs/1602.02244

Chapter 26
Fast Parameter Estimation for Cancer Cell Progression and Response to Therapy

P. Stpiczyński and B. Zubik-Kowal

26.1 Introduction

Many factors can alter cancer cell growth, including the response of the immune system. This continues to raise interest, as when stimulated, the immune system can attack cancer cells and be efficiently used to administer cancer-targeting therapy [AC15]. Consequently, mathematical modeling of tumor cell population growth and the competition between tumor cells and the immune system has emerged as an active area of research, and various mathematical models proposed throughout the literature have addressed open questions concerning the uncontrolled growth and spread of abnormal cells and its competition against the immune system. However, the growth of cancer cells involves many nonlinear intra- and extracellular phenomena that vary in time making it a complex multistep process [BeEtAl04, BeEtAl08, BeCh14, DrEtAl10]. Of particular relevance to the immune system, the competition between epithelial cells and immune cells that attempt to prevent cancer progression at an early stage has been detailed by a mathematical model proposed by Bellouquid and Delitala in the book [BeDe06]. The former of these cell populations refer to cells that have lost their differentiation and progress towards cancer competence [BeCh14]. The model is based on the mathematical kinetic theory for active particles developed by various authors, including, for example, the book [Be08] by Bellomo and the references found therein.

P. Stpiczyński
Institute of Mathematics, Maria Curie–Skłodowska University, Lublin, Poland
e-mail: przem@hektor.umcs.lublin.pl

B. Zubik-Kowal (✉)
Department of Mathematics, Boise State University, Boise, ID, USA
e-mail: zubik@math.boisestate.edu

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_26

Equations modeling the uncontrolled growth and spread of abnormal cells depend on a variety of biological parameters that need to be computed in light of available data, for example, through the use of global optimization techniques, which have been used in [AfBe14] for an integro-differential model of the response of the immune system against cancer growth. Various numerical studies validated against clinical and laboratory data have been performed in the literature, including a mouse model of breast cancer and laboratory data that have been used in [JoEtAl12] to identify biological parameters for two adenocarcinoma cell lines of the mammary gland in a female BALB/c mouse. Oncological data has been applied in [DrEtAl10] to estimate parameter values for a cancer-immune system dynamics model in the absence of treatment for a group of breast cancer patients, whereas the interaction of therapy against lung cancer progression has been examined and tested against clinical data for patients treated by chemotherapy and radiotherapy in [KoEtAl13]. The population kinetics of human tumor cells *in vitro* and their response to chemotherapy and/or radiotherapy have been examined in [BaEtAl03, JcEtAl09, Zu14]. The suppression of tumor growth by oxygen has been investigated in [JaEtAl09] through a mathematical model for brain cancer progression after therapy. Clinical data for a sample of brain cancer patients undergoing radiation treatment are compared in [NaZu15] to a macro-scale reaction-diffusion type model that accounts for large-dose stereotactic radiotherapy, providing good agreement with data.

Human tumor growth is associated with four phases of the cell cycle, involving DNA replication, mitosis, and cell division. As a result of variable phase-to-phase transition times, the length of the cell cycle is generally seen to be highly variable [BaEtAl03]. It has been reported that the median cycle time varies among individual cancer patients from as low as two days to up to several weeks [WiEtAl88]. With the underlying uncertainties associated with these highly variable quantities, it has become important to be able to accurately estimate the intrinsic physical parameters among individual patients and it is the goal of the present paper to provide a framework for fast computation for a model of the human tumor cell cycle by utilizing parallel computing environments. Conceptual aspects of the problem, not involving simulations in a parallel computing environment, have been considered in [Zu13, Zu14] for a simplified model.

Recently, multicore computer architectures have become very attractive for achieving high performance execution of scientific applications at low costs. Computer clusters are usually deployed to improve performance over that of a single computer. Unfortunately, the process of adapting existing software to such new architectures using OpenMP [ChEtAl01] and MPI (Message Passing Interface [Pa96]) can be difficult. The MATLAB Parallel Computing Toolbox readily allows to solve computationally and data-intensive problems using multicore processors and computer clusters. In this paper we also demonstrate that the MATLAB implementation of the proposed method can be easily and successfully adapted to clusters of multicore processors using the Parallel Computing Toolbox. Numerical experiments show that the parallel implementation achieves impressive speedup and good efficiency of the algorithm.

The paper is organized as follows. The model equations for the growth of human tumor cells are presented in Section 26.2. Sections 26.3 and 26.4 present the numerical algorithm, designed for the model as well as its extension, and results of numerical experiments. The paper finishes with concluding remarks sketched in Section 26.5.

26.2 Growth of Human Tumor Cells

The model equations developed in [BaEtAl03] are written in the form of delay partial differential equations

$$\frac{\partial G_1(x,t)}{\partial t} = 4b\,M(2x,t) \;-\; (k_1 + \mu_{G_1})\,G_1(x,t), \tag{26.1}$$

$$\begin{aligned}\frac{\partial S(x,t)}{\partial t} &= \varepsilon\,\frac{\partial^2 S(x,t)}{\partial x^2} \;-\; \mu_S\,S(x,t) \;-\; g\frac{\partial S(x,t)}{\partial x} \\ &\quad +\; k_1\,G_1(x,t) \;-\; I(x,t;T_S),\end{aligned} \tag{26.2}$$

$$\frac{\partial G_2(x,t)}{\partial t} = I(x,t;T_S) \;-\; (k_2 + \mu_{G_2})\,G_2(x,t), \tag{26.3}$$

$$\frac{\partial M(x,t)}{\partial t} = k_2\,G_2(x,t) \;-\; b\,M(x,t) \;-\; \mu_M\,M(x,t), \tag{26.4}$$

where the solutions $G_1(x,t)$, $S(x,t)$, $G_2(x,t)$, $M(x,t)$ represent the densities of cells in the G_1, S, G_2, and M-phases, respectively. The independent variable t represents time and x corresponds to the dimensionless relative DNA content used as a measure of cell size as the phase changes correspond to changes in DNA content. The delay term $I(x,t;T_S)$ is defined by

$$I(x,t;T_S) = \begin{cases} \displaystyle\int_0^\infty k_1\,G_1(y,t-T_S)\,\gamma(T_S,x,y)dy, & \text{for } t \geq T_S, \\ 0, & \text{for } t < T_S, \end{cases}$$

where

$$\gamma(\tau,x,y) = \exp(-\mu_S\tau)\Big[\rho((x-g\tau)-y) - \big(1+\nu(\tau,x,y)\big)\rho((x+g\tau)+y))\Big]$$

and

$$\rho(\xi) = \frac{1}{2\sqrt{\pi\varepsilon\tau}}\exp(-\xi^2/(4\epsilon\tau))$$

is the Gaussian distribution with variance $2\epsilon\tau$ and

$$\nu(\tau, x, y) = \frac{x+y}{g\tau}\Big(1 + O(\tau^{-1})\Big).$$

The parameter b represents the rate at which a cell in the M-phase divides into two daughter cells, ε represents the dispersion coefficient, and g is the average growth rate of DNA. Furthermore, k_1 is the transition rate from the G_1-phase to the S-phase, k_2 is the transition rate from the G_2-phase to the M-phase, and μ_{G_1}, μ_S, μ_{G_2}, μ_M are the death rates of cells in the G_1, S, G_2, and M-phases, respectively.

It is necessary to carefully estimate the parameters of the model separately for each subject's human melanoma cell line exposed to anti-cancer drugs. However, determining the parameters according to experimental data is a computationally heavy task that can easily become lengthy to run when doing so by means of sequential computations. On the other hand, it is important to keep required the computational time feasibly low in order to efficiently predict and simulate the growth of human tumor cells and their response to therapy. The goal of the paper is to develop efficient strategies for computing fast numerical solutions to (26.1)–(26.4) and an extension of it by invoking parallelization across independently working processors and to examine the computational gain attained with the use of parallel computing environments.

In the next section, we implement a time-domain decomposition that decouples the problem into a collection of independent subproblems and assigns the resulting separate tasks to independently working processors.

26.3 Parallelization Based on Time-Domain Decomposition

We apply pseudospectral differentiation matrices based on the Chebyshev-Gauss-Lobatto points

$$x_i = \frac{L}{2}\Big(1 - \cos\frac{i\pi}{I}\Big),$$

with $i = 0, 1, \ldots, I$, to discretize in $x \in [0, L]$ and approximate the first and second order spatial partial derivatives in (26.2). Pseudospectral semi-discretization leads to the following discretized differential system:

$$\begin{cases} \dfrac{du}{dt}(t) = Mu(t) + Q(t),\ 0 < t \le T, \\ u(0) = s_0, \end{cases} \tag{26.5}$$

that we wish to investigate henceforth. Here, M is a square matrix, $u(t)$ is a vector function whose elements are approximations to $S(x_i, t)$, s_0 is an initial vector corresponding to the initial values $S(x_i, 0)$, and $Q(t)$ is a vector function including the delay term computed from (26.1), (26.3), (26.4).

We wish to apply parallelization in the time-domain for (26.5) with an arbitrary number P of processors by choosing a positive integer P (representing any number of available processors) and dividing the interval $[0, T]$ into P equal subintervals $[t_{j-1}, t_j]$, where $j = 1, \ldots, P$, with $t_j = j\Delta t$ and $\Delta t = T/P$. The j-th processor is assigned to solve the following initial value problem:

$$\begin{cases} \dfrac{dv}{dt}(t) = Mv(t) + Q(t), \; t_{j-1} \le t \le t_j, \\ v(t_{j-1}) = \mathbf{0}, \end{cases} \tag{26.6}$$

where $\mathbf{0}$ is the zero vector. Even though the original system (26.5) is strongly joint and the solution at one instant depends on the behavior of the solutions at previous time points, the systems of equations given by the problem (26.6) are fully independent across all $j = 1, \ldots, P$. As the solution $v_j(t)$ to (26.6) is being computed, each utilized processor works independently over its subinterval $[t_{j-1}, t_j]$ without communicating with the remaining processors. Even though all processors work over distinct subintervals, the length of each subinterval is uniformly Δt. After the P processors finish their separate tasks and all solutions $v_j(t)$ for $j = 1, \ldots, P$, are computed, they are collected and the following formula:

$$u(t_j) = v_j(t_j) + \exp(\Delta t M) u(t_{j-1}).$$

is applied to generate the solution to (26.5). Convergence properties are proved in [Zu13].

The algorithm has been implemented in MATLAB using the Parallel Computing Toolbox™. The main part of the algorithm has been simply parallelized using the `parfor` construct for running parallel tasks on multiple processors (or cores). Our Matlab program has been tested on a cluster of four computers with two Intel(R) Xeon(R) CPU E5-2670 v3 (12 cores each with hyper-threading, 2.30 GHz, 128 GB RAM), running under Linux with MATLAB version R2015b and MATLAB Distributed Computing Server™, which allows to scale up programs developed with the Parallel Computing Toolbox to multiple computers.

We have checked the efficiency of the algorithm by applying it with different numbers of cores and different interval widths in a parallel computing environment. The resulting execution times obtained using Matlab performance measurement formalism are presented in Table 26.1. The values of P indicate the number of MATLAB parallel workers. For $P = 1, 2, 4, 8, 16$, computations have been performed on a single node, while for $P = 32$ and $P = 64$ we have used two and four nodes, respectively. The last two columns of the table show *speedup* and *efficiency* of the algorithm.

We can observe significant improvement as P increases illustrating the impressive speedup and reasonable efficiency of the algorithm. We note that single-node numerical simulations for the model equations have been obtained previously in [Zu14]. The resulting numerical solutions are presented in Figures 26.1–26.2. The top panels of Figure 26.1 present the evolution of $S(x, t)$ and $G_1(x, t)$ as functions

Table 26.1 Results of numerical experiments

P	Interval width	Elapsed time in seconds	Speedup	Efficiency
1	80.0	479.76	1.00	1.00
2	40.0	272.73	1.76	0.88
4	20.0	147.15	3.26	0.82
8	10.0	74.32	6.46	0.81
16	5.0	37.78	12.70	0.79
32	2.5	20.27	23.67	0.74
64	1.25	11.60	41.36	0.65

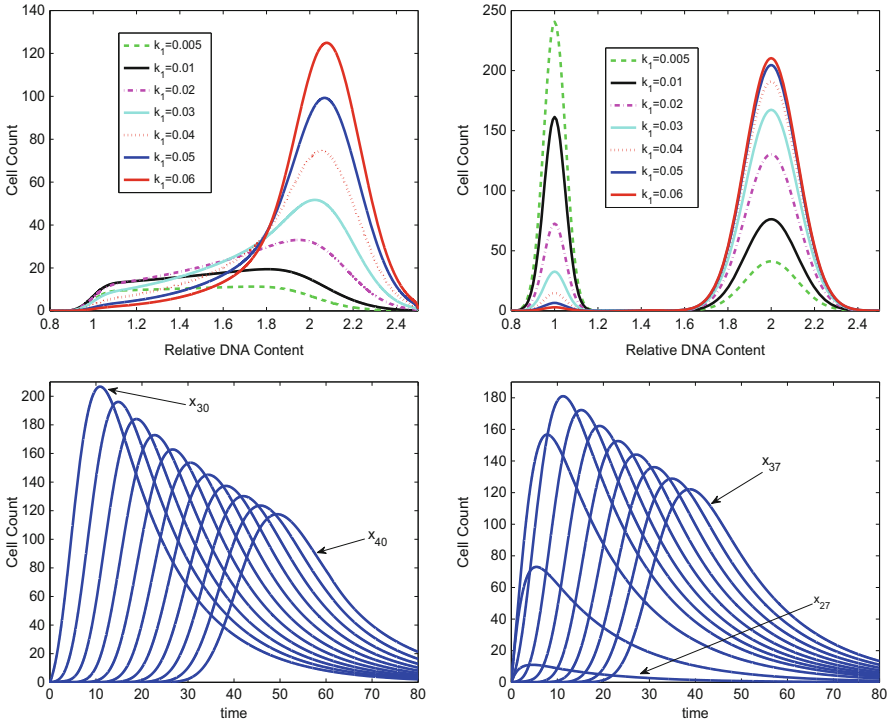


Fig. 26.1 $S(x,t)$ (top left) and $G_1(x,t)$ (top right) for varying transition rates k_1. $S(x,t)$ versus t at varying x locations for $k_1 = 0.06$ (bottom left) and $k_1 = 0.005$ (bottom right)

of x at $t = 80$ for a variety of rates k_1 of transition from the G_1-phase to the S-phase. The areas enclosed by the curves corresponding to larger values of k_1 are larger than the areas enclosed by the curves corresponding to smaller values of k_1, demonstrating an increase in the amount of cells in the S-phase with increasing k_1, as is naturally expected. The waveforms presented in the bottom panels of Figure 26.1 illustrate $S(x,t)$ as a function of time at varying values of the DNA content x ranging from x_{27} to x_{40} demonstrating a slower response in cell migration at higher DNA

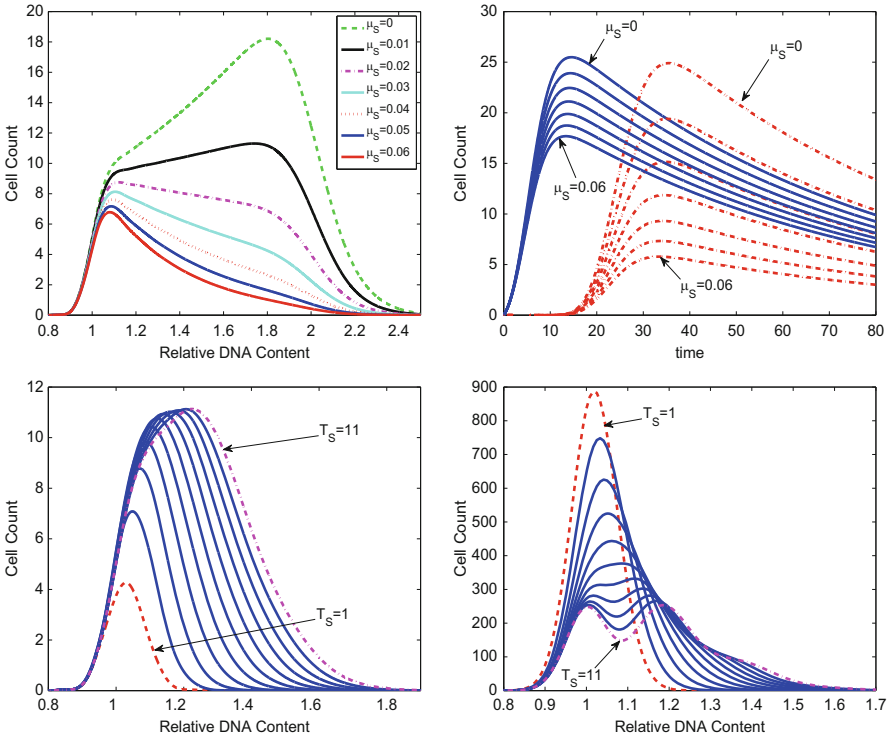


Fig. 26.2 $S(x,t)$ versus x (top left) and versus t (top right) for varying death rates μ_S. $S(x,t)$ (bottom left) and $G_1(x,t)$ (bottom right) for varying T_S

contents and low cell counts at low transition rates k_1. The top panels of Figure 26.2 present $S(x,t)$ for varying death rates of cells in the S-phase, ranging from $\mu_S = 0$ to $\mu_S = 0.06$ with an increment of 0.01. The top left-hand panel of Figure 26.2 presents $S(x,t)$ versus x at $t = 80$, in which there is a rapid decrease in cell count with increasing μ_S for intermediate DNA contents of approximately 1.8 and a lower rate of decrease for low DNA contents of approximately less than 1. The top right-hand panel of Figure 26.2 presents $S(x,t)$ versus t at x_{30} and x_{35}, illustrating the delay in the response of the cell count as DNA content varies. The bottom panels of Figure 26.2 present the evolution of $S(x,t)$ and $G_1(x,t)$ as functions of x at $t = 80$ for a variety of values of T_S ranging from $T_S = 1$ to $T_S = 11$ with the increment 1, depicting decreased migration from the S-phase to the G_1-phase with increasing values of the delay T_S and the formation of a bimodal cell count distribution in the G_1-phase above a threshold delay.

26.4 Parallelization for a Generalized Model of *in vivo* Tumor Growth

In this section, we test our algorithm for a generalized model developed in [BaEtAl03] that incorporates the effect of anti-cancer drug delivery and cell death *in vivo*. The generalized model accounts for apoptosis triggered by the delivery of anti-cancer drugs, which further induces cellular DNA loss and may provide understanding as to why some subjects fail to respond to therapy [BaEtAl03]. Increased cell deaths induced by apoptosis and the tracking of cells in the removal stage may aid in the analysis of drug-treated tumor development *in vivo*.

The generalized model is given in terms of more model parameters and a higher dimensional system, thus more intensive computations are required. The model involves G_1, S, G_2, M-phases together with an additional sub-population of cells in the removal stage (apoptosis) denoted by $R(x,t)$. The governing equation for the sub-population $R(x,t)$ introduced in [BaEtAl03] can be written in the following form:

$$\frac{\partial R}{\partial t}(x,t) = D_R \frac{\partial^2 R}{\partial x^2}(x,t) + \frac{\partial (g_R R)}{\partial x}(x,t) + \mu_{G_1} G_1(x,t) + \mu_S S(x,t) + \mu_{G_2} G_2(x,t) + \mu_M M(x,t), \tag{26.7}$$

where D_R is the dispersion coefficient and g_R is the average rate of decrease of DNA content per unit time.

We apply the principles of the parallel algorithm of Section 26.3 to the generalized model including (26.7) and have tested its efficiency by applying it in a series of numerical experiments involving different numbers of cores and different interval widths in a parallel computing environment. Table 26.2 lists the resulting execution times obtained using Matlab performance measurement functions and Figure 26.3 presents the solutions in the removal stage for varying dispersion coefficients and average rates of decrease of DNA content. Computational speedup is seen to be attained with increasing numbers of processors used. Higher dispersion coefficients result in distributed cell populations and higher average rates of decrease of DNA content yield advective drift towards unimodal distributions, with removal stage concentrations commensurate with cell count distributions.

Table 26.2 Results of numerical experiments

P	Interval width	Elapsed time in seconds	Speedup	Efficiency
1	80.0	494.85	1.00	1.00
2	40.0	270.72	1.83	0.91
4	20.0	145.62	3.40	0.85
8	10.0	76.73	6.45	0.81
16	5.0	40.63	12.18	0.76
32	2.5	20.79	23.80	0.74
64	1.25	11.51	43.00	0.67

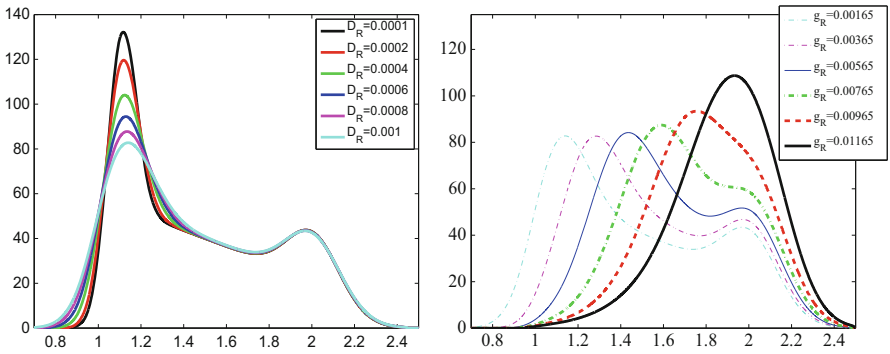


Fig. 26.3 Sub-population of cells in the removal stage for varying dispersion coefficients (left) and average rates of decrease of DNA content (right)

26.5 Conclusions and Future Work

This paper is devoted to fast simulations of two models of the human tumor cell cycle in a parallel computing environment and to the testing of the speedup gained. We have also extended our implementation to a generalized model of tumor development that tracks cells in the removal stage followed by apoptosis induced by the delivery of anti-cancer drugs *in vivo*.

We have shown that our MATLAB implementation can be easily and successfully adapted to clusters of computers with multicore processors using the Parallel Computing Toolbox.

It is clear that computationally intensive parts of the algorithm can be classified as data-parallel, thus we plan to implement the algorithm in GPU-accelerated computer architectures using CUDA [NV15] or OpenACC [Op13]. It should also be profitable to implement the algorithm in the Intel Many Integrated Core Architecture.

References

[AfBe14] Afraites, L., Bellouquid, A.: Global optimization approaches to parameters identification in an immune competition model. Commun. Appl. Ind. Math. **5**, e-466, 1–19 (2014)

[AC15] American Cancer Society: Cancer Facts & Figures 2015. American Cancer Society, Atlanta (2015)

[BaEtAl03] Basse, B., Baguley, B.C., Marshall, E.S., Joseph, W.R., van Brunt, B., Wake, G.C., Wall, D.J.N.: A mathematical model for analysis of the cell cycle in cell lines derived from human tumours. J. Math. Biol. **47**, 295–312 (2003)

[Be08] Bellomo, N.: Modeling complex living systems. A kinetic theory and stochastic game approach. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser, Inc., Boston (2008)

[BeEtAl04] Bellomo, N., Bellouquid, A., Delitala, M.: Mathematical topics on the modelling complex multicellular systems and tumor immune cells competition. Math. Models Methods Appl. Sci. **14**, 1683–1733 (2004)

[BeEtAl08] Bellomo, N., Li, N.K, Maini, P.K.: On the foundations of cancer modelling: selected topics, speculations, and perspectives. Math. Models Methods Appl. Sci. **18**, 593–646 (2008)

[BeCh14] Bellouquid, A., CH-Chaoui, M.: Asymptotic analysis of a nonlinear integro-differential system modeling the immune response. Comput. Math. Appl. **68**, 905–914 (2014)

[BeDe06] Bellouquid, A., Delitala, M.: Mathematical modeling of complex biological systems. A kinetic theory approach. With a preface by Nicola Bellomo. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser Boston, Inc., Boston (2006)

[ChEtAl01] Chandra, R., Dagum, L., Kohr, D., Maydan, D., McDonald, J., Menon, R.: Parallel Programming in OpenMP. Morgan Kaufmann Publishers, San Francisco (2001)

[DrEtAl10] Drucis, K., Kolev, M., Majda, W., Zubik-Kowal, B.: Nonlinear modeling with mammographic evidence of carcinoma. Nonlinear Anal. Real World Appl. **11**, 4326–4334 (2010)

[JaEtAl09] Jackiewicz, Z., Kuang, Y., Thalhauser, C., Zubik-Kowal, B.: Numerical solution of a model for brain cancer progression after therapy. Math. Model. Anal. **14**, 43–56 (2009)

[JcEtAl09] Jackiewicz, Z., Zubik-Kowal, B., Basse, B.: Finite-difference and pseudospectral methods for the numerical simulations of in vitro human tumor cell population kinetics. Math. Biosci. Eng. **6**, 561–572 (2009)

[JoEtAl12] Jorcyk, C.L., Kolev, M., Tawara, K., Zubik-Kowal, B.: Experimental versus numerical data for breast cancer progression. Nonlinear Anal. Real World Appl. **13**, 78–84 (2012)

[KoEtAl13] Kolev, M., Nawrocki, S., Zubik-Kowal, B.: Numerical simulations for tumor and cellular immune system interactions in lung cancer treatment. Commun. Nonlinear Sci. Numer. Simul. **18**, 1473–1480 (2013)

[NaZu15] Nawrocki, S., Zubik-Kowal, B.: Clinical study and numerical simulation of brain cancer dynamics under radiotherapy. Commun. Nonlinear Sci. Numer. Simul. **22**, 564–573 (2015)

[NV15] NVIDIA Corporation: CUDA Programming Guide (2015). NVIDIA Corporation available at http://www.nvidia.com/

[Op13] OpenACC: The OpenACC Application Programming Interface (2013). http://www.openacc.org

[Pa96] Pacheco, P.: Parallel Programming with MPI. Morgan Kaufmann, San Frncisco (1996)

[WiEtAl88] Wilson, G.D., McNally, N.J., Dische, S., Saunders, M.I., Des Rochers, C., Lewis, A.A., Bennett, M.H.: Measurement of cell kinetics in human tumours in vivo using bromo-deoxyuridine incorporation and flow cytometry. Br. J. Cancer **58**, 423–431 (1988)

[Zu13] Zubik-Kowal, B.: Numerical algorithm for the growth of human tumor cells and their responses to therapy. Appl. Math. Comput. **230**, 174–179 (2014)

[Zu14] Zubik-Kowal, B.: A fast parallel algorithm for delay partial differential equations modeling the cell cycle in cell lines derived from human tumors. In: Hartung, F., Pituk, M. (eds.) Recent Advances in Delay Differential and Difference Equations, vol. 94, pp. 251–260. Springer Proceedings in Mathematics & Statistics. Springer, Cham (2014)

Chapter 27
Development of a Poroelastic Model of Spinal Cord Cavities

J. Venton, P.J. Harris, and G. Phillips

27.1 Introduction

Syringomyelia is a rare medical condition characterised by large fluid filled cavities (syrinxes) in the spinal cord (Figure 27.1). How these syrinxes form is not fully understood, although it is thought to be influenced by pressure changes in the cerebrospinal fluid (CSF) surrounding the cord [El13]. CSF bathes the brain and spinal cord and actions such as coughing or bending along with physiological processes such as pulse cause harmless CSF movement and pressure changes.

Certain neurological disorders or traumatic spinal cord injuries can affect the size and shape of the CSF region around the cord, causing these fluid movements to become exaggerated. Over a period of months or years this can damage the spinal cord tissue, leading to a syrinx. It is impractical to observe this process in a patient, and mathematical modelling is increasingly seen to be a valuable tool for validating hypotheses of syrinx formation [El13].

To this end a mathematical model of the spinal cord tissue has been developed and solved using the finite element method. Elasticity, permeability and porosity parameters for spinal cord tissue have been obtained to improve the model's accuracy. Applied boundary conditions in the finite element model will simulate the exaggerated CSF pressures that occur following disorders or injuries that commonly precede syringomyelia. The results of these simulations will reveal the pressures

J. Venton (✉)
School of Computing, Engineering and Mathematics, University of Brighton, Brighton, UK
e-mail: j.venton2@brighton.ac.uk

P.J. Harris
University of Brighton, Brighton, UK
e-mail: p.j.harris@brighton.ac.uk

G. Philips
Brighton Centre for Regenerative Medicine, University of Brighton, Brighton, UK
e-mail: g.phillips@brighton.ac.uk

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_27

Fig. 27.1 A syrinx (∗) in the spinal cord in the neck (C Hardwidge, Hurstwood Park Neurological Unit)

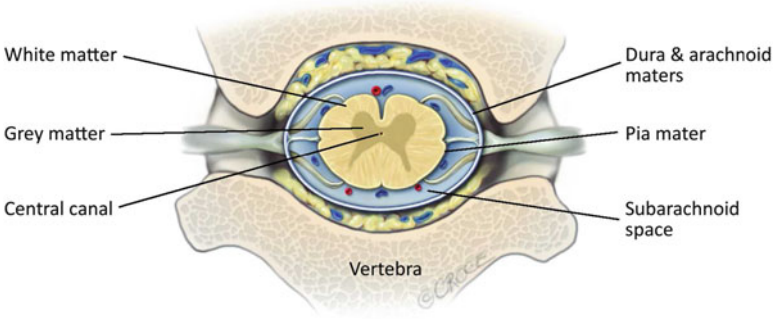


Fig. 27.2 Cross section of spinal cord anatomy [El13] (© 2016 Beth Croce)

and stresses that can occur within spinal cord tissue. This information can be used to determine whether the stresses and pressures present within the cord are high enough to cause damage and either start or worsen a syrinx.

27.2 Spinal Cord Model

The spinal cord is composed of grey and white matter and is enclosed by three separate layers known as the pia, arachnoid and dura maters (Figure 27.2). The subarachnoid space lies between the pia and arachnoid maters and contains the CSF surrounding the cord. Grey and white matter consist of cells such as axons and neurons, which are surrounded by a fluid similar to CSF known as the extracellular fluid. To fully capture the fluid/tissue nature of the spinal cord, a poroelastic model is used. Poroelasticity has the advantage of modelling both the solid part of a material and the fluid contained within it; subsequently in a spinal cord model both the tissue fibre stresses and the extracellular fluid pressures can be calculated when an external CSF pressure is applied.

27.2.1 Poroelastic model

A poroelastic material consists of a solid (known as the solid skeleton) containing many small interconnected fluid filled spaces (pores). Poroelastic spinal cord models have been used previously to study syringomyelia (see [Ha09, St16] for example), and are increasingly thought to be an appropriate model for several biological materials [Mo13]. A simplified poroelastic model of spinal cord tissue consists of a linear elastic solid skeleton and an incompressible pore fluid.

In a linear poroelastic model, the effective stress $\boldsymbol{\sigma}$ [Le98] is a combination of the stress from the solid skeleton $\boldsymbol{\sigma}_s$ and the stress due to internal pore fluid pressure $\boldsymbol{\sigma}_f$, that is

$$\boldsymbol{\sigma} = (1-\phi)\boldsymbol{\sigma}_s - \phi\boldsymbol{\sigma}_f \,.$$

The contribution of each of these stresses is determined by the porosity ϕ, a dimensionless parameter that describes what fraction of the material is occupied by pore fluid. Stress in the solid skeleton is related to strain $\boldsymbol{\varepsilon}$ by

$$\boldsymbol{\sigma}_s = D\boldsymbol{\varepsilon}$$

where D is the elasticity matrix, which includes elasticity properties of the solid skeleton such as Young's modulus E and Poisson's ratio ν [Bo10]. The momentum balance equation is given by

$$-\nabla\cdot\boldsymbol{\sigma} = ((1-\phi)\rho_s + \phi\rho_f)\frac{\partial^2\boldsymbol{u}}{\partial t^2} - \boldsymbol{F}_s \tag{27.1}$$

where the vector $\boldsymbol{u}$ represents displacements of the solid skeleton, $\boldsymbol{F}_s$ represents external forces and ρ_f and ρ_s are the pore fluid and solid skeleton densities, respectively.

Under the assumption that fluid movements are small [Le98], Darcy's law is used to approximate the movement of pore fluid

$$q = -\frac{\boldsymbol{\kappa}}{\mu}\nabla p$$

where q is the flux of fluid through a material, $\boldsymbol{\kappa}$ is the intrinsic permeability and μ is the dynamic viscosity of the pore fluid. Conservation of mass of pore fluid in a poroelastic material is given by

$$\frac{\partial p}{\partial t} = \frac{\kappa}{\mu}\nabla^2 p - \alpha\frac{\partial\boldsymbol{\varepsilon}}{\partial t} \tag{27.2}$$

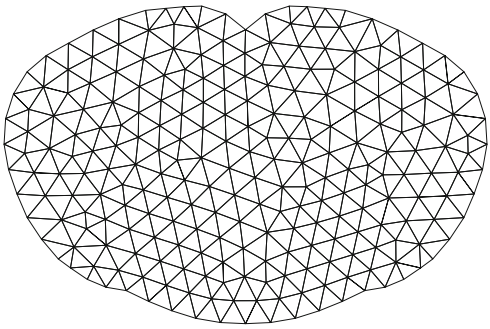


Fig. 27.3 A coarse finite element mesh of the spinal cord cross section taken from an image of the spinal cord and meshed using ABAQUS

which is essentially a linear diffusion equation with an extra term to describe how the solid skeleton displacements (strains) $\boldsymbol{\varepsilon}$ affect the pore pressure p [Le98]. The extent of this effect is determined by the Biot-Willis coefficient α.

27.2.2 Finite element simulations

Initially a two-dimensional plane strain model is being built, to represent a cross section of the spinal cord. This initial study allows the stresses and pressures across the cord cross section to be calculated when external CSF forces are applied. Equations (27.1) and (27.2) are solved using the finite element method, over a mesh derived from anatomical data (Figure 27.3). Applying the finite element method to (27.1) and (27.2) yields the following system of equations:

$$\begin{aligned} M\ddot{\boldsymbol{u}} + K\boldsymbol{u} - Q\boldsymbol{p} &= -\boldsymbol{F}_s \\ S\dot{\boldsymbol{p}} + Q^T\dot{\boldsymbol{u}} + H\boldsymbol{p} &= -\boldsymbol{F}_f \end{aligned} \tag{27.3}$$

where M and K are the solid mass and stiffness matrices, S and H are the fluid mass and permeability matrices, $\boldsymbol{F}_s$ and $\boldsymbol{F}_f$ are external solid and fluid forces and Q is the coupling matrix. Vectors $\boldsymbol{u}$ and $\boldsymbol{p}$ represent solid skeleton displacements and pore fluid pressures, respectively. The coupling matrix Q describes how the solid skeleton displacements affect the pore fluid pressure (via the $Q^T\dot{\boldsymbol{u}}$ term) and how the pore fluid pressure displaces the solid skeleton (via the $Q\boldsymbol{p}$ term).

To rewrite the system in Equation (27.3) in a solvable format ($A\boldsymbol{x} = \boldsymbol{b}$), the substitution $\dot{\boldsymbol{u}} = \boldsymbol{v}$ can be made and the system rewritten as

$$\begin{pmatrix} M & 0 & 0 \\ 0 & I & 0 \\ 0 & Q^T & S \end{pmatrix} \begin{bmatrix} \dot{\boldsymbol{v}} \\ \dot{\boldsymbol{u}} \\ \dot{\boldsymbol{p}} \end{bmatrix} = \begin{pmatrix} 0 & -K & Q \\ I & 0 & 0 \\ 0 & 0 & -H \end{pmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} - \begin{bmatrix} \boldsymbol{F}_s \\ \boldsymbol{0} \\ \boldsymbol{F}_f \end{bmatrix}$$

or

$$A_0\dot{\boldsymbol{x}} = A_1\boldsymbol{x} + F \tag{27.4}$$

where

$$A_0 = \begin{pmatrix} M & 0 & 0 \\ 0 & I & 0 \\ 0 & Q^T & S \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & -K & Q \\ I & 0 & 0 \\ 0 & 0 & -H \end{pmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix}.$$

This is now a linear equation, which can be solved using traditional methods. To solve the system dynamically, the Crank-Nicolson method is used to approximate the solution at the current ($\boldsymbol{x}_n$) and next ($\boldsymbol{x}_{n+1}$) time step using the following approximations:

$$\boldsymbol{x} \approx \frac{\boldsymbol{x}_{n+1} + \boldsymbol{x}_n}{2}, \quad \dot{\boldsymbol{x}} \approx \frac{\boldsymbol{x}_{n+1} - \boldsymbol{x}_n}{h} \tag{27.5}$$

where h is the length of the time step. Substituting (27.5) into (27.4) and rearranging yields:

$$(2A_0 - hA_1)\boldsymbol{x}_{n+1} = (2A_0 + hA_1)\boldsymbol{x}_n + 2hF$$

which can be solved to find the solid skeleton displacements and pore fluid pressures at the new time step.

27.3 Model Parameters

The properties of a linear poroelastic material such as that described by Equations (27.1) and (27.2) are defined by a set of material parameters, including Young's modulus, Poisson's ratio, permeability and porosity. Values of these parameters for spinal cord tissue are needed to increase the accuracy of the described model, as interpretation of finite element simulation results will be based on how the model behaves in particular disease and injury conditions. Certain parameters are being calculated in a program of in-house experimental work (permeability, porosity) whilst others are being taken from the literature (Young's modulus, Poisson's ratio).

27.3.1 Young's modulus and Poisson's ratio

Compared to permeability and porosity, more work has been undertaken to characterise the Young's modulus (YM) of spinal cord tissue. Several factors influence the measured value of YM including the type of testing technique used [Mc11] and the strain rate at which tissue was tested [Fr13]. As a consequence, values of YM for spinal cord in the literature range from 48 Pa [Ko15] to 1 400 000 Pa [Ma03]. In the present study, values obtained at strain rates similar to those exerted on the cord

by disturbed CSF have been chosen. In addition, for a poroelastic model YM of the solid skeleton (tissue fibres) will be different to that of the overall tissue. Studies measuring YM for individual tissue fibres are less common but produce values at the lower end of the scale [Ko15].

White matter is thought to be more anisotropic than grey matter and the direction in which the tissue is tested affects the measured YM [Ko15]. However, the initial model will represent both grey and white matter as isotropic with a single value of YM each. Furthermore, spinal cord tissue is viscoelastic nor linear elastic, but at lower strain rates a poroelastic model mimics spinal cord tissue behaviour well [Ch07]. The tissue is presumed to be almost incompressible, and Poisson's ratio is taken to be $\nu = 0.49$.

27.3.2 Permeability and porosity

A diffusion weighted MRI (DW-MRI) technique, neurite orientation and dispersion density indexing (NODDI) is being used to derive information regarding the permeability and porosity of spinal cord tissue. NODDI was developed [Zh12] to reveal the microstructure of central nervous system tissues, and DW-MRI data analysed using the NODDI model (details in [Gr15]) yields a set of structural parameters for the tissue. These parameters are used to derive permeability and porosity.

In the present work, the definition of porosity ϕ in spinal cord tissue is taken to be the fraction of spinal cord tissue not occupied by axons or neurons. Porosity of central nervous system tissue has been measured using tracers in tissue [Ni98] and DW-MRI [Gr15], values found are in the region of $\phi \approx 0.2$.

The definition of permeability in spinal cord tissue for the present work is the ease with which extracellular fluid can move through the extracellular space when the tissue is subjected to pressure gradients. In the literature, spinal cord permeability has been provisionally measured using three main techniques. Firstly, cord tissue compression data has been fitted to a poroviscoelastic finite element model and a permeability value derived in this way [Ch07]. Secondly DW-MRI has been used to obtain the direction of permeability, although the magnitude of permeability was not measured in this study [Sa06]. Finally the movement of tracers through central nervous system tissues has been measured producing a diffusion coefficient [Ni98] for spinal cord tissue. However, in unhealthy states such as syringomyelia, bulk movement (as opposed to diffusive movement) of fluid is thought to be present [Sy08].

Consequently it is the intrinsic permeability κ, rather than a diffusion coefficient, that is needed for the present model. NODDI structural parameters used to derive permeability and porosity include the volume fraction ν_{in}, the orientation dispersion index ODI and the mean orientation vector $\boldsymbol{\mu}$.

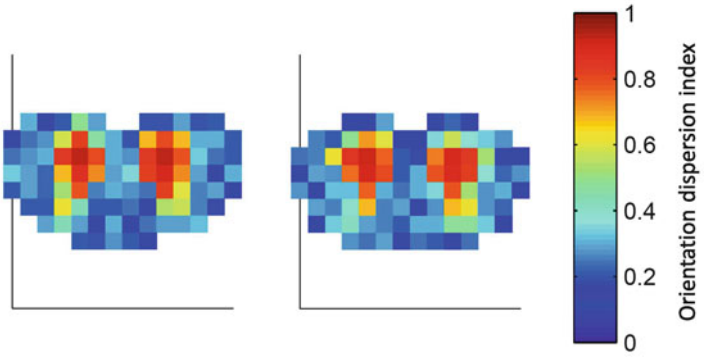


Fig. 27.4 Orientation dispersion index (*ODI*) values on example cross sections of a spinal cord scan. The difference between grey matter (red) and white matter (blue) is clear

The volume fraction ν_{in} is a dimensionless number describing the fraction of tissue occupied by tissue fibres such as axons and neurons. The porosity can be calculated from this in a fairly straightforward fashion as $\phi = 1 - \nu_{in}$.

The mean orientation $\boldsymbol{\mu}$ and orientation dispersion index *ODI* describe the direction of tissue fibres and how spread out the fibres are, respectively. Using $\boldsymbol{\mu}$ the directions of the spinal cord permeability tensor are calculated, and *ODI* is used to calculate the magnitude of the permeability. This is achieved using equations that calculate the permeability of a fibrous material with different fibre layouts. For full details of the permeability and porosity derivation see [Ve17].

Grey and white matter have different permeabilities - grey matter is thought to be almost isotropic whereas white matter is anisotropic due to the axon tracts that run its length. This is reflected in the *ODI* values, where values near 0 indicate nearly parallel fibres and values near 1 indicate randomly oriented fibres (see Figure 27.4). As a consequence the permeability tensors for grey and white matter will take different values.

27.4 Spinal Cord Simulations

Mathematical models of the spinal cord and surrounding tissue provide a useful technique for examining the processes preceding syrinx formation. The time scales over which a syrinx forms can be greatly accelerated and stresses within the spinal cord tissue can be calculated. As described in Section 27.1, boundary conditions applied to the model will simulate raised cerebrospinal fluid (CSF) pressures that result from neurological disorders or injuries [Cl13]. It is thought that a syrinx is often preceded by spinal cord tissue oedema [Fi00]; this hypothesis can be evaluated with the model by introducing regions of increased porosity.

A numerical simulation such as the one described is inevitably dependent on the accuracy of the parameters entered into the model. Whilst there is a reasonable amount of information in the literature regarding elastic properties of the spinal cord and healthy versus irregular pressures in the surrounding CSF, information relating to the permeability and porosity of spinal cord tissue is scarce. The preliminary data described in Section 27.3 will be used to update the model.

The poroelastic model allows both the stresses within the tissue fibres and the fluid pressure in the extracellular space to be calculated. If the internal spinal cord pressure and stresses induced by irregular CSF pressures are sufficiently high that they may damage the tissue, this indicates that irregular CSF pressures are at least partially responsible for syrinx formation and growth.

The results of the present work will contribute to the understanding of syrinx formation and growth via an improved spinal cord tissue model. This will be beneficial to clinicians as at present it can be difficult to determine the best treatment options when the exact cause of syrinx formation remains elusive.

References

[Bo10] Bower, A.: Applied Mechanics of Solids. Taylor & Francis/CRC, Boca Raton (2010)

[Ch07] Cheng, S., Bilston, L.E.: Unconfined compression of white matter. J. Biomech. **40**(1), 117–124 (2007)

[Cl13] Clarke, E., Fletcher, D., Stoodley, M., Bilston, L.: Computational fluid dynamics modelling of cerebrospinal fluid pressure in Chiari malformation and syringomyelia. J. Biomech. **46**(11), 1801–1809 (2013)

[El13] Elliott, N., Bertram, C., Martin, B., Brodbelt, A.R.: Syringomyelia: a review of the biomechanics. J. Fluid. Struct. **40**, 1–24 (2013)

[Fi00] Fischbein, N., Dillon, W., Cobbs, C., Weinstein, P.: The "presyrinx" state: is there a reversible myelopathic condition that may precede syringomyelia? Neurosurg. Focus **8**(3), 1–13 (2000)

[Fr13] Franze, K., Janmey, P., Guck, J.: Mechanics in neuronal development and repair. Ann. Rev. Biomed. Eng. **15**, 227–251 (2013)

[Gr15] Grussu, F., Schneider, T., Zhang, H., Alexander, D., Wheeler-Kingshott, C.: Neurite orientation dispersion and density imaging of the healthy cervical spinal cord in vivo. NeuroImage **111**, 590–601 (2015)

[Ha09] Harris, P., Hardwidge, C.: A porous finite element model of the motion of the spinal cord. In: Constanda, C., Pérez, M.E. (eds.) Integral Methods in Science and Engineering, vol. 2, pp.193–201. Birkhäuser, Boston (2009)

[Ko15] Koser, D., Moeendarbary, E., Hanne, J., Kuerten, S., Franze, K.: CNS cell distribution and axon orientation determine local spinal cord mechanical properties. Biophys. J. **108**(9), 2137–2147, (2015)

[Le98] Lewis, R.W., Schrefler, B.A.: The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media. Wiley, Chichester (1998)

[Ma03] Mazuchowski, E., Thibault, L: Biomechanical properties of the spinal cord and pia mater. In: Summer Bioengineering Conference, Key Biscayne, FL (2003)

[Mc11] McKee, C.T., Last, J.A., Russell, P., Murphy, C.J.: Indentation versus tensile measurements of Young's modulus for soft biological tissues. Tissue Eng. Pt. B-Rev. **17**(3), 155–164 (2011)

[Mo13] Moeendarbary, E., Valon, L., Fritzche, M., Harris, A., Moulding, D., Thrasher, A., Stride, E., Mahadevad, L., Charras, G.: The cytoplasm of living cells behaves as a poroelastic material. Nat. Mater. **12**(3), 253–261, (2013)

[Ni98] Nicholson C., Syková, E.: Extracellular space structure revealed by diffusion analysis. Trends Neurosci. **21**(5), 207–215 (1998)

[Sa06] Sarntinoranont, M., Chen, X., Zhao, J., Mareci, T.: Computational model of interstitial transport in the spinal cord using diffusion tensor imaging. Ann. Biomed. Eng. **34**(8), 1304–1321 (2006)

[St16] Støverud, K., Alnæs, M., Langtangen, H., Haughton, V., Mardal, K.: Poro-elastic modeling of syringomyelia. Comput. Methods Biomec. **19**(6), 686–698 (2016)

[Sy08] Syková, E., Nicholson, C.: Diffusion in brain extracellular space. Physiol. Rev. **88**, 1277–1340 (2008)

[Ve17] Venton, J., Bouyagoub, S., Harris, P.J., Phillips, G.: Deriving spinal cord permeability and porosity using diffusion-weighted MRI data. In Proceedings of the 6th Biot Conference on Poromechanics, Paris, France (2017). (Manuscript accepted, in publication)

[Zh12] Zhang, H., Schneider, T., Wheeler-Kingshott, C., Alexander, D.: NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage **61**(4), 1000–1016 (2012)

Chapter 28
A Semi-Analytical Solution for a Buildup Test for a Horizontal Well in an Anisotropic Gas Reservoir

B.J. Vicente, A.P. Pires, and A.M.M. Peres

28.1 Introduction

Well testing is a specialized area of petroleum reservoir engineering whose foundations come from groundwater theory and its applications. Well tests are very common field operations with the purpose of gathering dynamic data (i.e., bottom hole pressure and flowrate x time) at a single well during a relative short period of time. The data is subsequently analyzed using specialized techniques to provide local (hundreds of meters) reservoir parameters estimates (such as permeability, well impairment and distance to flow barriers near the well). Well test data is also used to constrain large-scale reservoir numerical models to improve geological description and parameters, which are built mostly from static data.

There are many well test types; the choice is driven by well testing objectives and economics. Buildup tests are the most popular. In a buildup test, the well is first brought to production at constant rate q for some time t_p (the flow period) and then the well is shut in (the buildup period) as shown in Figure 28.1. The wellbore pressure response to such rate schedule is also shown in this figure.

For most cases, published well testing analytical solutions consider a slightly compressible fluid only so that the governing flow equation in porous media is linear. In this case, conventional mathematical tools as integral transforms are applicable

B.J. Vicente (✉)
North Fluminense State University, Macaé, Brazil

Laboratório Nacional de Computação Científica, Petrópolis, Brazil
e-mail: brunojosevicente@gmail.com

A.P. Pires • A.M.M. Peres
North Fluminense State University, Macaé, RJ, Brazil
e-mail: adolfo.puime@gmail.com; alvarommperes@gmail.com

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_28

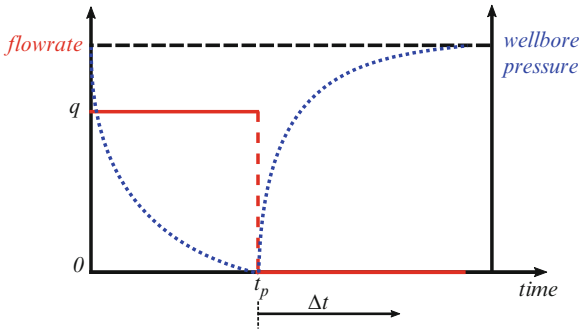


Fig. 28.1 Flowrate schedule and pressure behavior for a typical buildup well test

as well as the superposition principle; then solutions for a buildup test are easily obtained by superimposing two flow period solutions. There are relatively few attempts to approach nonlinear well testing problems by analytical means, most research and application papers consider numerical solutions only. In this text, we present an approximate analytical solution for nonlinear gas flow in porous media, which is constructed by the Green's Function technique. As an example application, the approximate solution for a horizontal well buildup test in an anisotropic gas reservoir is derived and compared to a finite-difference numerical solution.

28.2 Nonlinear Differential Equation Formulation

The diffusivity equation that governs the isothermal single phase flow of a real gas with constant composition through a homogeneous anisotropic porous media is given by

$$\nabla \cdot \left(\frac{p}{\mu(p)Z(p)} \mathbf{k} \nabla p \right) + \frac{p}{Z(p)} \tilde{q}\,(x, y, z, t) = \phi c_t\,(p)\, \frac{p}{Z(p)} \frac{\partial p}{\partial t}. \tag{28.1}$$

Equation 28.1 is a nonlinear partial differential equation for the dependent variable p, which assumes that Darcy's law is valid and gravity effects are negligible. The terms p, Z, $\mu(p)$, and $c_t\,(p)$ denote pressure, the gas compressibility factor, gas viscosity, and total compressibility (gas + pore compressibility), respectively. Also in Equation 28.1, ϕ and $\mathbf{k}$ represent porosity and the permeability tensor, whereas the source term $\tilde{q}$ represents the strength of a source/sink at coordinate (x,y,z) at a time t, with units of volumetric rate per infinitesimal source volume. This term is used to represent a single or several wells, either producer or injector, and handles several wellbore geometries such as vertical, horizontal, or fractured wells. A source (injector) well adds mass to the porous media, thus $\tilde{q} > 0$; consequently, for a producer gas well (sink) we have $\tilde{q} < 0$.

The nonlinear behavior of Equation 28.1 is reduced if the dependent variable p is replaced by the so-called pseudo pressure function $m(p)$ [Hu66] defined by

$$m(p) = 2\int_{p_{ref}}^{p} \frac{p'}{\mu(p')Z(p')}dp', \tag{28.2}$$

where p_{ref} is an arbitrary reference pressure, usually the atmospheric pressure. In heat conduction literature, Equation 28.2 is known as Kirchhoff transformation. This function depends on gas properties only and is specific for a given gas reservoir. It is also convenient to define a pseudo pressure change

$$\Delta m(p) = m(p_i) - m(p) = 2\int_{p}^{p_i} \frac{p'}{\mu(p')Z(p')}dp',$$

where p_i stands for initial reservoir pressure. Using these definitions, and assuming that Cartesian coordinates are aligned to permeability principal directions, one can rewrite Equation 28.1 for Cartesian coordinates as

$$k_x\frac{\partial^2 \triangle m\,(p)}{\partial x^2} + k_y\frac{\partial^2 \triangle m\,(p)}{\partial y^2} + k_z\frac{\partial^2 \triangle m\,(p)}{\partial z^2} = \phi\mu\,(\triangle m)\,c_t\,(\triangle m)\,\frac{\partial \triangle m\,(p)}{\partial t} + \frac{2p_{sc}T}{T_{sc}}\tilde{q}_{sc}(x,y,z,t),$$

where p_{sc} and T_{sc} denote pressure and temperature at standard conditions, k_x, k_y, and k_z represent permeability values in x, y, and z directions, T denote reservoir temperature, and $\tilde{q}_{sc}$ represents a variable-strength point source at standard conditions. Because $m(p)$ is a bijective function, we can write gas viscosity and total compressibility as a function of $m(p)$ or $\triangle m(p)$. Note that, by introducing the pseudo pressure function, the nonlinear partial differential equation becomes quasi-linear.

Dimensionless coordinates, time, pseudo pressure, and source are defined by

$$x_D = \frac{x}{l_c}\sqrt{\frac{k_{ref}}{k_x}}, \quad y_D = \frac{y}{l_c}\sqrt{\frac{k_{ref}}{k_y}}, \quad z_D = \frac{z}{l_c}\sqrt{\frac{k_{ref}}{k_z}};$$

$$t_D = \frac{k_{ref}t}{\phi\,(\mu c_t)_i\,l_c^2}; \quad m_D = \frac{\pi k_{ref}hT_{sc}\triangle m(p)}{q_{ref}p_{sc}T}$$

and

$$f_D\,(x_D, y_D, z_D, t_D) = \frac{2\pi hl_c^2}{q_{ref}}\tilde{q}_{sc}\,(x,y,z,t)\,.$$

In the above equations, k_{ref} and q_{ref} denote reference values for permeability and flowrate, $(\mu c_t)_i$ the viscosity-total compressibility product at initial reservoir pressure, and h the constant reservoir thickness. The parameter l_c represents a convenient characteristic length.

Then, the governing PDE in dimensionless variables is given by

$$\frac{\partial^2 m_D}{\partial x_D^2} + \frac{\partial^2 m_D}{\partial y_D^2} + \frac{\partial^2 m_D}{\partial z_D^2} - H_D \frac{\partial m_D}{\partial t_D} = f_D\left(x_D, y_D, z_D, t_D\right), \tag{28.3}$$

where H_D is a variable coefficient that represents the ratio of the viscosity compressibility product at a given pressure (or pseudo pressure) to its value at initial reservoir pressure, that is,

$$H_D = \frac{\mu\left(m_D\right) c_t\left(m_D\right)}{\left(\mu c_t\right)_i}.$$

For typical well testing applications, it is often enough to consider the reservoir being infinite in the x-y plane, with both top and bottom boundaries impermeable to flow. Therefore, for a gas reservoir initially at equilibrium, the initial and boundary conditions in dimensionless variables are given by

$$m_D\left(x_D, y_D, z_D, t_D = 0\right) = 0, \tag{28.4}$$

$$\begin{gathered} \lim_{x_D \to \pm\infty} m_D\left(x_D, y_D, z_D, t_D\right) = \lim_{y_D \to \pm\infty} m_D\left(x_D, y_D, z_D, t_D\right) = 0 \\ \left.\frac{\partial m_D\left(x_D, y_D, z_D, t_D\right)}{\partial z_D}\right|_{z_D=0} = \left.\frac{\partial m_D\left(x_D, y_D, z_D, t_D\right)}{\partial z_D}\right|_{z_D=h_D} = 0 \end{gathered}. \tag{28.5}$$

28.3 Reformulation as an Integral-Differential Equation

Introducing the auxiliary variable $\omega(m_D) = H_D - 1$, which represents the relative change of the μc_t product from its initial value, Equation 28.3 becomes

$$\frac{\partial^2 m_D}{\partial x_D^2} + \frac{\partial^2 m_D}{\partial y_D^2} + \frac{\partial^2 m_D}{\partial z_D^2} - \frac{\partial m_D}{\partial t_D} = f_D\left(x_D, y_D, z_D, t_D\right) - \omega\left(m_D\right) \frac{\partial m_D}{\partial t_D}. \tag{28.6}$$

Rewritten in this way, one may interpret that nonlinearity acts as a nonlinear source term defined by the last term on the right-hand side of Equation 28.6. [BaEtAl13] take this point of view to recast the pseudo pressure-diffusivity equation as an integro-differential equation. Following their procedure, one gets

$$m_D\left(\mathbf{x}_D, t_D\right) = \int_0^{t_D} \int_\Omega \left[f_D\left(\mathbf{x}'_D, t'_D\right) + \omega\left(m_D\right) \frac{\partial m_D\left(\mathbf{x}'_D, t'_D\right)}{\partial t'_D} \right] \times$$
$$G_D\left(\mathbf{x}_D, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D. \tag{28.7}$$

Equation 28.7 is a Volterra integro-differential equation where G_D denotes the dimensionless Green's function (GF) associated with the problem stated by Equations 28.3–28.5. Note that here $\mathbf{x}_D$ and $\mathbf{x}'_D$ are position vectors. Parameters $\mathbf{x}'_D$ and t'_D of G_D represent the position and time where and when the instantaneous impulse is applied, whereas $\mathbf{x}_D$ and t_D denote the position and time where and when the effect of such impulse is felt. The pertinent GFs associated with Equation 28.5 boundary conditions are shown in the Appendix. Detailed derivation of Equation 28.7 is shown by [MiEtAl16] and also by [Vi16].

We assume that the Integral-Differential Equation 28.7 is solvable by the following point iteration scheme:

$$m_D^{(k)}\left(\mathbf{x}_D, t_D\right) = -\int_0^{t_D} \int_\Omega f_D\left(\mathbf{x}'_D, t'_D\right) G_D\left(\mathbf{x}_D, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D$$
$$-\int_0^{t_D} \int_\Omega \omega\left(m_D^{(k-1)}\right) \frac{\partial m_D^{(k-1)}\left(\mathbf{x}'_D, t'_D\right)}{\partial t'_D} G_D\left(\mathbf{x}_D, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D,$$

where $m_D^{(0)}$ is the initial guess. The first integral on the right-hand side represents the linear component of the solution that is identical to the mathematical solution for a slightly compressible fluid (i.e., liquid). The second integral on the right accounts for the viscosity-compressibility variation with pressure and thus, it represents the deviation from the linear part of the solution. This term is referred here as the "corrective term."

We will take the result of the first iteration as an approximate solution for the dimensionless pseudo pressure, that is

$$m_D\left(\mathbf{x}_D, t_D\right) \simeq m_D^{(1)}\left(\mathbf{x}_D, t_D\right) = m_{0D}\left(\mathbf{x}_D, t_D\right) - m_{1D}\left(\mathbf{x}_D, t_D\right), \tag{28.8}$$

where

$$m_{0D}\left(\mathbf{x}_D, t_D\right) = -\int_0^{t_D} \int_\Omega f_D\left(\mathbf{x}'_D, t'_D\right) G_D\left(\mathbf{x}_D, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D \tag{28.9}$$

and

$$m_{1D}\left(\mathbf{x}_D, t_D\right) = \int_0^{t_D} \int_\Omega \omega\left(m_D^{(0)}\right) \frac{\partial m_D^{(0)}\left(\mathbf{x}'_D, t'_D\right)}{\partial t'_D} G_D\left(\mathbf{x}_D, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D. \tag{28.10}$$

Previous articles [BaEtAl13, SoEtAl15, MiEtAl16], that applied the same technique as in here, have shown that the above approximation is able to capture the essence of the nonlinear behavior of a gas reservoir, being sufficiently accurate for engineering applications to say the least.

28.4 Application: Buildup Test in a Horizontal Well

In this section, we apply the approximate solution presented before to a particular case of a gas field being produced by a single horizontal well. Figure 28.2 shows the horizontal well geometry and notation. The wellbore is represented by a segment of a line source with length equals to L_w, running parallel at distance z_w from the formation bottom.

28.4.1 Formulation and Solution at the Wellbore

We assume a uniform flux wellbore model in which the well flowrate at standard conditions q_{sc} is uniformly distributed over the length L_w, thus

$$\tilde{q}_{sc}(x,y,z,t) = \begin{cases} -\dfrac{q_{sc}(t)}{L_w}\delta(y-0)\,\delta(z-z_w), & 0 \le x \le L_w \\ 0, & x < 0,\ x > L_w \end{cases}, \tag{28.11}$$

where δ stands for the Dirac delta function. In order to conform to the well testing literature in which $q_{sc} > 0$ means production from a well, a negative sign appears in Equation 28.11. Setting the characteristic length l_c equal to L_w, the dimensionless source term becomes

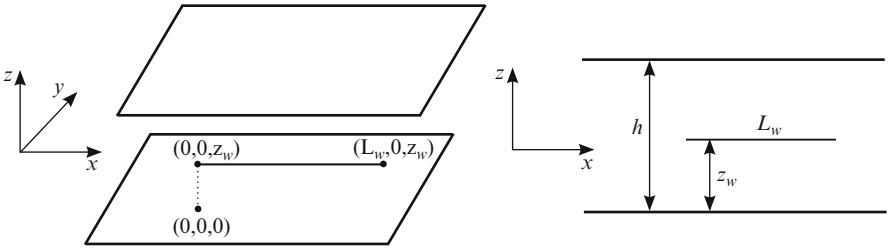


Fig. 28.2 Reservoir and horizontal wellbore geometries

$$f_D(x_D, y_D, z_D, t_D) = \begin{cases} -\alpha q_D(t_D)\,\delta(y_D - 0)\,\delta(z_D - z_{wD}), & 0 \le x_D \le L_{wD} \\ 0, & x_D < 0,\ x_D > L_{wD} \end{cases},$$

where $\alpha = 2\pi h_D \sqrt{k_{ref}/k_y}$, $L_{wD} = \sqrt{k_{ref}/k_x}$ and $q_D(t_D) = q_{sc}(t_D)/q_{ref}$.

For the buildup test flowrate schedule shown in Figure 28.1, we have $q_{sc}(t) = q_{sc}$ for $t < t_p$ and $q_{sc} = 0$ for $t > t_p$. Thus, by choosing $q_{ref} = q_{sc}$, the dimensionless rate schedule becomes

$$q_D(t_D) = \begin{cases} 1, & t_D < t_{pD} \\ 0, & t_D > t_{pD} \end{cases}. \tag{28.12}$$

In well testing applications, only the pressure at the wellbore radius r_w is relevant; therefore, we will only consider the application of the approximate solution at the wellbore. In here, the wellbore pressure (or pseudo pressure) is always evaluated at coordinates $\mathbf{x}_D = (L_{wD}/2, r_{wD}, z_{wD})$, which will be denoted by $\mathbf{x}_{wD}$. The y-coordinate r_{wD} is

$$r_{wD} = \frac{r_w}{2L_w}\left[\left(\frac{k_z}{k_y}\right)^{\frac{1}{4}} + \left(\frac{k_y}{k_z}\right)^{\frac{1}{4}}\right]\sqrt{\frac{k_{ref}}{k_y}}.$$

For dimensionless time such $t_D < t_{pD}$ (that is, during the flow period at constant flowrate), one obtains the dimensionless wellbore pseudo pressure solution m_{wD} by evaluating Equations 28.8–28.10 at $\mathbf{x}_{wD}$

$$m_{wD}(t_D) = m_{w0D}(t_D) - m_{w1D}(t_D). \tag{28.13}$$

The term m_{w0D} is the liquid-like part of the solution and is given by

$$m_{w0D}(t_D) = \alpha \int_0^{t_D} G_{yD}(r_{wD}, 0, \tau)\, G_{zD}(z_{wD}, z_{wD}, \tau)\, S_D(L_{wD}/2, \tau)\, d\tau, \tag{28.14}$$

where S_D represents a source function defined below

$$S_D(x_D, \tau) = \int_0^{L_{wD}} G_{xD}\left(x_D, x'_D, \tau\right) dx'_D = \\ = \frac{1}{2}\left[\operatorname{erf}\left(\frac{L_{wD} - x_D}{2\sqrt{\tau}}\right) + \operatorname{erf}\left(\frac{x_D}{2\sqrt{\tau}}\right)\right].$$

The "corrective term" evaluated at the wellbore is

$$m_{w1D}(t_D) = \int_0^{t_D}\int_\Omega \omega(m_{0D}) \frac{\partial m_{0D}\left(\mathbf{x}'_D, t'_D\right)}{\partial t'_D} G_D\left(\mathbf{x}_{wD}, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D, \tag{28.15}$$

where the m_{0D} terms are calculated by

$$m_{0D}\left(\mathbf{x}_D, t_D\right) = \alpha \int_0^{t_D} G_{yD}\left(y_D, 0, \tau\right) G_{zD}\left(z_D, z_{wD}, \tau\right) S_D\left(x_D, \tau\right) d\tau. \tag{28.16}$$

For the buildup period, that is, when $t_D > t_{pD}$, the dimensionless buildup wellbore pseudo pressure solution m_{wDV} is obtained from Equations 28.8–28.10 evaluated at $\mathbf{x}_{wD}$ for the flowrate schedule defined by Equation 28.12, which yields

$$m_{wDV}\left(t_D\right) \simeq m_{w0DC}\left(t_{pD} + \triangle t_D\right) - m_{w0DC}\left(\triangle t_D\right) - m_{w1DV}\left(t_{pD} + \triangle t_D\right), \tag{28.17}$$

where the dimensionless shut-in time is defined by $\Delta t_D = t_D - t_{pD}$.

Note that C subscripts indicate a mathematical solution obtained under constant unit rate $q_D = 1$, while solutions which carry a subscript V represent solutions under a variable-rate schedule. Thus, the first two terms in the right-hand side in Equation 28.17 are obtained by simply evaluating Equation 28.14 at $t_{pD} + \Delta t_D$ and Δt_D, respectively.

The corrective term at the wellbore for the buildup period is evaluated from

$$\begin{aligned} m_{w1DV}\left(t_D\right) = & \int_0^{t_{pD}} \int_\Omega \omega\left(m_{0DV}\right) \frac{\partial m_{0DV}\left(\mathbf{x}'_D, t'_D\right)}{\partial t'_D} G_D\left(\mathbf{x}_{wD}, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D \\ & + \int_{t_{pD}}^{t_D} \int_\Omega \omega\left(m_{0DV}\right) \frac{\partial m_{0DV}\left(\mathbf{x}'_D, t'_D\right)}{\partial t'_D} G_D\left(\mathbf{x}_{wD}, \mathbf{x}'_D, t_D, t'_D\right) d\mathbf{x}'_D dt'_D, \end{aligned} \tag{28.18}$$

where the terms m_{0DV} which appear in the integrands in Equation 28.18 are given by

$$m_{0DV}\left(\mathbf{x}_D, t_D\right) = m_{0DC}\left(\mathbf{x}_D, t_{pD} + \Delta t_D\right) - m_{0DC}\left(\mathbf{x}_D, \Delta t_D\right), \quad \text{when} \quad t_D > t_{pD};$$

otherwise become simply

$$m_{0DV}\left(\mathbf{x}_D, t_D\right) \equiv m_{0DC}\left(\mathbf{x}_D, t_D\right).$$

Note that m_{0DC} terms above are calculated directly from Equation 28.16 using the proper dimensionless time value.

By examining the ω parameter in Equation 28.18, one sees that the corrective buildup term m_{w1DV} has two components: the first integral in the right-hand side is associated with the variation of the gas properties that has occurred over the flow period; the second is due to gas properties variation that occurs as the buildup period progresses.

28.4.2 Fluid and Rock Data

All results presented in this text are based on a synthetic rock and fluid dataset. Gas viscosity and compressibility Z-factor vs. pressure, obtained from correlations, are shown in Figure 28.3 and are identical to the ones presented by [SoEtAl15]. Relevant rock, fluid, and some other data appear in Table 28.1. In this table, c_r, c_{gi}, and d_g denote, respectively, rock (pore) compressibility, gas compressibility at initial reservoir pressure (p_i), and gas gravity for this specific gas. The results are generated for cases where there is vertical anisotropy only, i.e., permeability field is isotropic in the x-y plane, so we have $k_x = k_y = k_H$. In this situation, it is convenient to set the reference permeability k_{ref} equal to k_H.

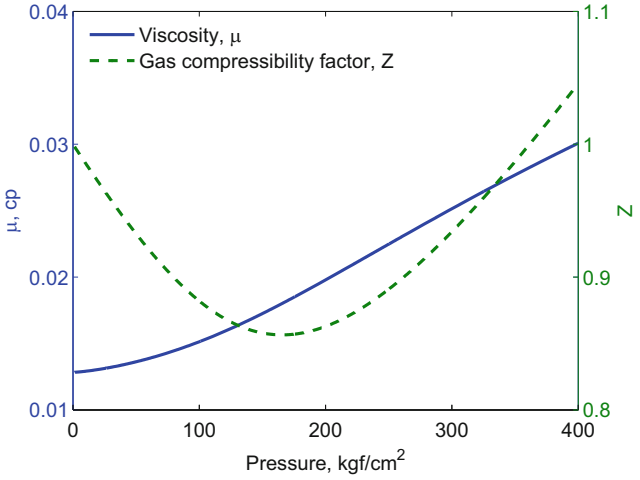


Fig. 28.3 Gas properties at reservoir temperature

Table 28.1 Basic rock and fluid data

property	value	unit
h	50	m
ϕ	0.1	–
c_r	50×10^{-6}	cm^2/kgf
p_i	400	kgf/cm^2
T	80	$°C$
c_{gi}	0.0013368	cm^2/kgf
μ_i	0.0300765	cP
d_g	0.7	–
L_w	400	m
r_w	0.1	m
q_{sc}	2×10^6	m^3/day

28.4.3 Comparison to Finite Difference

To assess the accuracy of the approximate solution at the wellbore presented in this work, results are compared with a commercial finite-difference reservoir simulator for the data shown in Table 28.1 and Figure 28.3 for both flow and buildup periods. Finite-difference gridding, local refinement, time step selection, as well as additional comparisons, can be found in [Vi16].

Evaluation of the wellbore pseudo pressure approximated solutions requires single-time integration for the liquid-like terms m_{0D} and m_{w0D}, whereas the corrective term m_{w1D} requires a quadruple numerical integration. These integrations are accomplished by the multidimensional numerical-integration package CUBA [Ha05]. The package presents several algorithms; see [Ha05] for details. Corrective term m_{w1D} values shown here are calculated by the Vegas algorithm from that package.

Figure 28.4a presents a log-log plot of the dimensionless wellbore pseudo pressure m_{wD} calculated from Equations 28.13–28.16 vs. the results obtained by a finite-difference numerical simulator. The results shown are for the flow period of an off-center horizontal well placed five meters distance from the formation bottom (i.e., $z_w = 5m$) with vertical anisotropy ($k_x = k_y = k_H = 2mD$ and $k_z = 0.2mD$). The finite-difference simulator output is pressure vs. time, thus, at the end of each run, the numerical results are first converted to pseudo pressure using the gas properties and Equation 28.2, and then the corresponding dimensionless pseudo pressure are calculated from the dimensionless-variable definitions given before in Section 28.2. In Figure 28.4a one sees that the approximate solution m_{wD} shows an excellent match to simulator results, except at very-short times. In all solutions proposed in this work, a segment of a line source represents the wellbore, whereas in the numerical simulator, the well has a finite wellbore radius equal to r_w, so the difference at short times observed in Figure 28.4a is expected. Anyhow, this difference does not have any impact in engineering applications. See [SoEtAl15] for

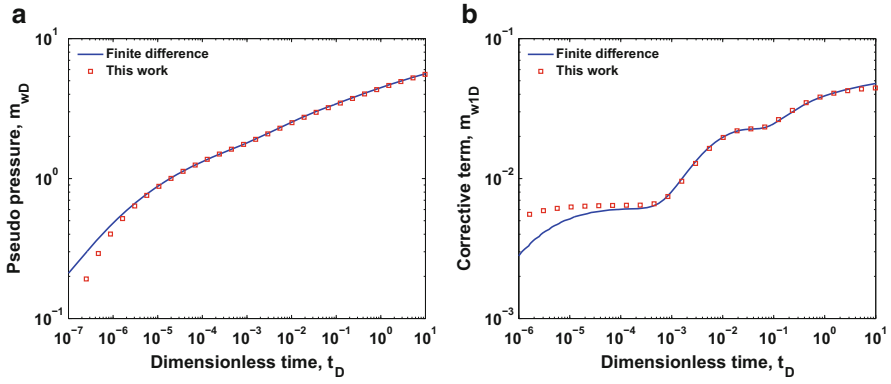


Fig. 28.4 Flow period comparison: **(a)** Pseudo pressure m_{wD} log-log plot and **(b)** Corrective term m_{w1D} log-log plot

the comparison of finite wellbore solution vs. line-source solution for vertical gas wells.

Figure 28.4b shows the corrective term at the wellbore (m_{w1D}) calculated from Equation 28.15 compared with the corrective term obtained by the finite-difference simulator. Simulator corrective-term results shown in this plot are obtained indirectly from the subtraction of two different runs keeping the grid and time steps unchanged. In the first run, the gas properties are allowed to vary according to gas properties shown in Figure 28.2, whereas in the second simulator run, gas properties are kept fixed at their values at initial reservoir pressure, so this run is equivalent to a slightly compressible-liquid response. Even though the magnitude of m_{w1D} is very small compared to m_{wD} values, Figure 28.4b shows an excellent agreement for $t_D > 10^{-4}$. For shorter times, finite difference and approximate-solution m_{w1D} curves do not match because the later represents the wellbore by a segment of a line source as explained previously. One should recognize that simulator values m_{w1D} represent the total deviation of the compressible-gas solution from the slightly compressible-liquid solution, whereas m_{w1D} from Equation 28.15 is just an approximation obtained by halting the successive substitution at the end of the first iteration. Thus, according to the results shown in Figure 28.4, one sees that the proposed solution given by Equation 28.13 is an accurate approximation. One also should note that, for gas production, all terms that appear in the integrand of Equation 28.15 are positive, and then we have $m_{w1D} > 0$. Therefore, it follows from Equation 28.13 that $m_{wD} < m_{w0D}$, that is, at a given time the pseudo pressure solution for a gas reservoir is always less than the equivalent liquid-like solution.

Similar comparison for a buildup test after 48 hours ($t_{pD} = 2.5 \times 10^{-2}$) of gas production at constant flowrate is shown next. Rock, fluid properties, vertical anisotropy, and horizontal well location are identical to those used in Figure 28.4.

In well testing practice, buildup data is always shown in terms of the difference between the pseudo pressure at given shut-in time Δt to the pseudo pressure value observed at end of the flow period (i.e., at t_p). In dimensionless variables, this difference is denoted by dimensionless pseudo pressure change (Δm_{wD}) and given by

$$\Delta m_{wD}\left(\Delta t_D\right) = m_{wD}\left(t_{pD}\right) - m_{wDV}\left(t_{pD} + \Delta t_D\right). \tag{28.19}$$

Using the dimensionless pseudo pressures defined by Equations 28.13 and 28.17 in the right-hand side of Equation 28.19, one gets

$$\Delta m_{wD}\left(\Delta t_D\right) = \Delta m_{w0D}\left(\Delta t_D\right) + m_{w1DV}\left(t_{pD} + \Delta t_D\right) - m_{w1D}\left(t_{pD}\right), \tag{28.20}$$

where Δm_{w0D} represents the buildup change difference of the related slightly compressible fluid solution, given by

$$\Delta m_{w0D}\left(\Delta t_D\right) = m_{w0DC}\left(t_{pD}\right) - m_{w0DC}\left(t_{pD} + \Delta t_D\right) + m_{w0DC}\left(\Delta t_D\right).$$

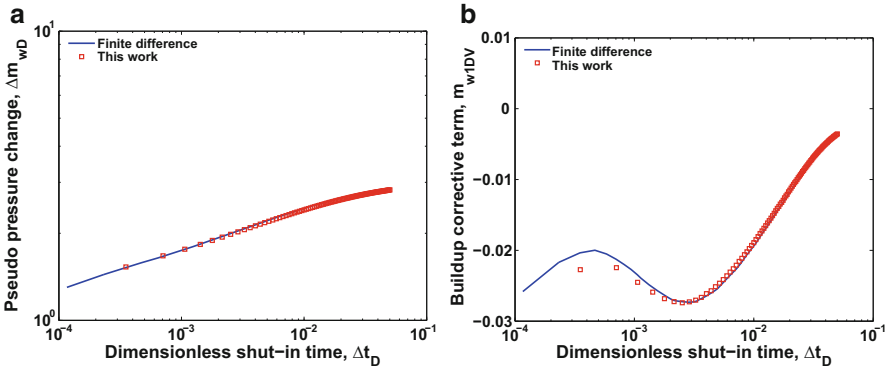


Fig. 28.5 Buildup period comparison: **(a)** Pseudo pressure change $\triangle m_{wD}$ log-log plot and **(b)** Buildup corrective term m_{w1DV} semilog plot

Comparison for the dimensionless pseudo pressure change $\triangle m_{wD}$ computed by Equation 28.20 to numerical simulator results is shown in Figure 28.5a. A very good agreement is achieved. The semilog plot in Figure 28.5b presents the corrective term m_{w1DV} calculated from Equation 28.18 compared to the finite-difference simulator. As explained before, simulator m_{w1DV} values shown in this figure come from two simulator runs. A good match is achieved at short shut-in times, getting increasingly better as shut-in time increases.

Two aspects of the buildup corrective term in Figure 28.5b deserve attention: first, the corrective term m_{w1DV} tends to zero as shut-in time goes to infinite; second, m_{w1DV} is negative. These two aspects appear in other cases considered by [Vi16] and seem to be a general behavior. Also note that being m_{w1DV} negative while m_{w1D} at t_{pD} is positive, Equation 28.20 indicates that we have $\triangle m_{wD}(\triangle t_D) < \triangle m_{w0D}(\triangle t_D)$. In other words, at a given shut-in time the pseudo pressure change for a gas well is always less than the equivalent slightly compressible-liquid solution. Thus, the pseudo pressure recovery towards the equilibrium during a buildup test is slower than for liquid-like behavior.

28.5 Conclusions

This article presents an analytical approach to the nonlinear flow of a real gas in an anisotropic porous media. The technique first employs a Kirchhoff-like transform to reach a quasi-linear partial differential equation. Then, the gas viscosity-compressibility product nonlinear behavior is written as a nonlinear source term so that the mathematical problem can be re-formed as a Volterra integral-differential equation in terms of Green's Functions. It is assumed that this integral equation can be solved by a point iteration scheme. Here, the results obtained at the end of the first iteration are taken as an approximate solution. The buildup test of a horizontal

well in an infinite gas reservoir serves here as an example application. Results for a synthetic dataset obtained by the approximate solution are shown to be very accurate by comparison to a finite-difference commercial reservoir simulator. The analytical approach presented here also provides useful insights about the solution behavior. It seems that the technique described in this article can be extended to problems in heat conduction in solids with temperature dependent physical properties.

Acknowledgements The authors wish to express gratitude for the financial support provided by the PRH-20/Agencia Nacional do Petroleo, Gas Natural e Biocombustiveis (ANP), Petroleo Brasileiro S.A. (Petrobras), and Universidade Estadual do Norte Fluminense (UENF).

Appendix

The 3-D anisotropic GF which appears in Equation 28.7 can be obtained from one-dimensional GF's by Newman's product rule [Ca59], that is

$$G_D(\mathbf{x}_D, \mathbf{x}'_D, t_D, t'_D) = G_{xD}(x_D, x'_D, t_D, t'_D) G_{yD}(y_D, y'_D, t_D, t'_D) G_{zD}(z_D, z'_D, t_D, t'_D).$$

Infinite space 1-D GF for the x and y directions is given by [CoEtAl11]

$$G_{lD}\left(l_D, l'_D, t_D, t'_D\right) = \frac{1}{\sqrt{4\pi\left(t_D - t'_D\right)}} \exp\left[-\frac{\left(l_D - l'_D\right)^2}{4\left(t_D - t'_D\right)}\right], \; t_D - t'_D > 0$$

with $l = x$ or y.

The z-direction GF is a slab with Newmann boundary conditions at both ends which has two equivalent expressions [CoEtAl11]

$$G_{zD}(z_D, z'_D, t_D, t'_D) = \frac{1}{\sqrt{4\pi(t_D - t'_D)}} \times \sum_{n=-\infty}^{\infty} \left\{ \exp\left[-\frac{(2nh_D + z_D - z'_D)^2}{4(t_D - t'_D)}\right] + \exp\left[-\frac{(2nh_D + z_D + z'_D)^2}{4(t_D - t'_D)}\right] \right\} \quad (28.21)$$

or

$$G_{zD}\left(z_D, z'_D, t_D, t'_D\right) = \frac{1}{h_D}\left\{1 + \sum_{n=1}^{\infty} \exp\left[-\frac{(n\pi)^2\left(t_D - t'_D\right)}{h_D^2}\right] \cos\left(\frac{n\pi z_D}{h_D}\right) \cos\left(\frac{n\pi z'_D}{h_D}\right)\right\}. \quad (28.22)$$

During calculations, it is important to observe that the series in Equation 28.21 converges faster for small values of $(t_D - t'_D)$, whereas Equation 28.22 evaluation is faster for large $(t_D - t'_D)$ values.

References

[BaEtAl13] Barreto, Jr. A.B., Peres, A.M.M., Pires, A.P.: A variable-rate solution to the nonlinear diffusivity gas equation by use of Green's-function method. SPE J. **18**, 57–68 (2013). SPE-145468-PA

[Ca59] Carslaw, H.S., Jaeger, J.C.: Conduction of Heat in Solids. Oxford University Press, London (1959)

[CoEtAl11] Cole, K.D., Haji-Sheikh, A., Beck, J.V., Litkouhi, B.: Heat Conduction Using Green's Functions. Taylor & Francis, Boca Raton (2011)

[Ha05] Hahn, T.B.: Cuba—a library for multidimensional numerical integration. Comput. Phys. Commun. **168**, 78–95 (2005)

[Hu66] Al-Hussainy, R., Ramey, H.J., Crawford, P.B.: The flow of real gases through porous media. J. Pet. Technol. **18**, 624–636 (1966). SPE-1243-A-PA

[MiEtAl16] Miranda, F.A., Barreto, Jr. A.B., Peres, A.M.M.: A novel uniform-flux solution based on the Green's function method for modeling the pressure-transient behavior of a restricted-entry well in anisotropic gas reservoirs. SPE J. **21**, 1870–1882 (2016). SPE-180919-PA

[SoEtAl15] Sousa, E.P.S., Barreto, Jr. A.B., Peres, A.M.M.: Finite-wellbore-radius solution for gas wells by Green's functions. SPE J. **20**, 842–855 (2015). SPE-169323-PA

[Vi16] Vicente, B.J.: Application of Green's functions for mathematical modeling of horizontal well pressure tests in gas reservoirs. D.Sc. Thesis (in Portuguese), Universidade Estadual do Norte Fluminense, Brazil (2016)

Chapter 29
Counter-Gradient Term Applied to the Turbulence Parameterization in the BRAMS

M.E.S. Welter, H.F. de Campos Velho, S.R. Freitas, and R.S.R. Ruiz

29.1 Introduction

The atmospheric dynamics is simulated by solving the Navier-Stokes equation, considering several physical phenomena. Some atmospheric processes are expressed by using parameterization: cloud formation, surface representation, turbulence, precipitation. The turbulence parameterization is closed under different orders. For zeroth order, the turbulent flux is represented by a function. In the first closure order, Reynolds tensors are approximated as a product between an eddy diffusivity and the gradient of the main property. The second order closure is parameterized with the third order tensor expressed as a parameter multiplying the second order tensor.

On the top of the Planetary Boundary Layer (PBL), under convective regime, a counter-gradient flux is verified from observations. Indeed, turbulent flow is a part of physics strongly supported by experimental efforts. We are not going to explain details on the structure of the convective PBL. But, we note that only the second order closure is able to represent the latter flux. However, modifying the first order approach by adding a counter-gradient term, it is possible to represent such flow. The first studies for representing the latter issue were carried out by Deardorff [De66, De72]. Here, the eddy diffusivity is formulated by Taylor's statistical theory of turbulence [DeEtAl00], and a new term is used – derived from the Large Eddy Simulation (LES) [CuEtAl98]. This new turbulence scheme is applied to BRAMS, a meso-scale meteorological model. The simulation is compared with experimental

M.E.S. Welter (✉) • H.F. de Campos Velho • R.S.R. Ruiz
National Institute for Space Research (INPE), São José dos Campos, SP, Brazil
e-mail: marowelter@gmail.com; haroldo@lac.inpe.br; rennaruiz@gmail.com

S.R. Freitas
National Aeronautics and Space Administration (NASA), Washington, DC, USA
e-mail: saulo.freitas@noaa.gov

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6_29

data from the Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) experiment (https://daac.ornl.gov/LBA/lba.shtml).

29.2 Turbulence Model

The turbulent flux for a property $\bar{\varphi}$ can be represented by a first order closure:

$$\langle v_i' \varphi' \rangle = K_{\alpha\alpha} \frac{\partial \langle \varphi \rangle}{\partial x} \tag{29.1}$$

where v_i' is the wind velocity components, $K_{\alpha\alpha}$ is the eddy diffusivity, $\alpha = x, y, z$, and the operator $\langle \varphi \rangle$ denotes time average. From Taylor's statistical theory on turbulence [Ta22, Ta35], the eddy diffusivity can be expressed as product between an average velocity and a characteristic length:

$$K_{\alpha\alpha} \sim \langle v_i(t) \rangle \langle x(t) \rangle \tag{29.2}$$

where the index-i indicates the wind components (u, v, w). The velocity is defined as $v(t) \equiv x(t)\, dt$. Substituting the velocity definition in the above equation and taking the average, the eddy diffusivity can be written as:

$$K_{\alpha\alpha} = \frac{d}{dt} \left[\langle x^2(t) \rangle \right] = 2 \langle v_i^2(t) \rangle \int_0^t \int_0^\tau \rho_{L_i}(\tau)\, dt' d\tau \tag{29.3}$$

where $v_i(t)$ is the i-th Lagrangian wind component of a *fluid particle*, and $x(t)$ is its displacement on the direction-i. The autocorrelation function is denoted by ρ_{L_i}, normalized by the Lagrangian velocity:

$$\rho_{L_i}(\tau) = \frac{\langle v_i(t+\tau) v_i(t) \rangle}{\langle v_i^2(t) \rangle} \,. \tag{29.4}$$

Eulerian formulation can be computed from Lagrangian quantities by using Gifford-Hay and Pasquill's assumption, where Lagrangian and Eulerian autocorrelations (or spectral) functions are the same, but shifted by a constant β [DeEtAl92, DeEtAl97]:

$$\rho_i(\tau) = \rho_{L_i}(\beta_i \tau) \,, \qquad \beta_i = \frac{\sigma_i}{U} \frac{\sqrt{\pi}}{4} \,, \qquad \sigma_i \equiv \sqrt{\langle v_i^2(t) \rangle} \,, \tag{29.5}$$

with U the wind intensity.

Applying Fourier transform to Equation (29.3), and noting that $\rho_i(\tau)$ is an even function, Eulerian eddy diffusivity can be expressed by [DeEtAl92, DeEtAl97, Ca10]

$$K_{\alpha\alpha} = \frac{\sigma_i^2 \beta_i}{2\pi} \int_0^\infty \left[\frac{F_i^E(n) \sin(2\pi n t / \beta_i)}{n} \right] dn \tag{29.6}$$

where $F_i^E(n) \equiv S_i^E(n)/\sigma_i^2$, being $S_i^E(n)$ the turbulent kinetic energy on direction i, and n is a frequency. For large diffusion time ($t \to \infty$), an asymptotic expression for the $K_{\alpha\alpha}$ can be derived [DeEtAl92, DeEtAl97, Ca10] as

$$K_{\alpha\alpha} = \frac{\sigma_i^2 \beta_i F_i^E(0)}{4} \,. \tag{29.7}$$

An explicit formulation for eddy diffusivities $K_{\alpha\alpha}$ can be obtained by using empirical relations and Obukhov's similarity theory. A key issue is to derive an analytical formulation to the spectrum. Degrazia et al. [DeEtAl00] have derived a spectral formula for all atmospheric stability conditions:

$$\begin{aligned} nS_i^E(n) = & \frac{1.06 c_i f \, \psi_\epsilon^{2/3} \, (z/h)^{2/3} \, w_*^2}{[(f_m^*)_i^c]^{5/3} \left(1 + 1.5\,[f/(f_m^*)_i^c]\right)^{5/3}} \\ & + \frac{1.5 c_i f (\Phi_\epsilon)^{2/3} u_*^2}{[(f_m^*)_i^{n+es}]^{5/3} \left(1 + 1.5\,[f^{5/3}/(f_m^*)_i^{n+es}]\right)^{5/3}} \end{aligned} \tag{29.8}$$

where c_i are empirical constants, z is the level over the surface, h is the planetary boundary layer height, w^* is the velocity scale for the convective condition, u^* is the friction velocity, $f = nU/z$ is the frequency in Hertz, U is the wind velocity, ψ_ϵ and Φ_ϵ are non-dimensional dissipation functions for convective and stable/neutral conditions, $(f_m^*)_i^c$ and $(f_m^*)_i^{n+es}$ are the maximum frequencies for a convective and stable/neutral conditions, respectively. The wind variances can be calculated by integrating the spectra over all frequencies: $\sigma_i^2 = \int_0^\infty S_i^E(n)\, dn$.

29.2.1 Counter-Gradient Model

As already mentioned, the first order closure is not able to represent a counter-gradient flux. Therefore, a new term must be added in the parameterization scheme for describing such flow:

$$\overline{u'_\alpha \varphi'} = -K_{\alpha\alpha} \left[\frac{\partial \langle \varphi \rangle}{\partial x_\alpha} - \gamma_\varphi \right] \tag{29.9}$$

being γ_φ the counter-gradient. From experiments, Deardorff had estimated $\gamma_\theta \approx 6.5 \times 10^{-6}\,\mathrm{C\,cm^{-1}}$ [De66], and he did a new evaluation to $\gamma_\theta \approx 7 \times 10^{-6}\,\mathrm{K\,cm^{-1}}$ [De66].

The counter-gradient term is only applied under convective condition, i.e., it is not used for neutral and stable boundary layers. The γ_φ is employed for heat and mass transport, but it is not used to the momentum due to the pressure effect. The planetary boundary layer stability can be calculated from the Monin-Obukhov's

length L. In order to codify which parameterization should be applied for different stability conditions to the atmosphere, we adopted $|L| > 500$ to characterize the neutral boundary layer:

$$\begin{cases} -500 \leq L < 0 : \text{Convective} \\ |L| > 500 : \text{Neutral} \\ 0 < L \leq 500 : \text{Stable} \end{cases}$$

The Monin-Obukhov's length is expressed by

$$L = \frac{-u_*^3}{\kappa \left(g/\theta_{v_0}\right) \left(\overline{w'\theta'_v}\right)_0} \tag{29.10}$$

where κ is the von Kármán's constant, g is the gravity acceleration, θ_v virtual potential temperature, and $\left(\overline{w'\theta'_v}\right)_0$ is the heat flux from the surface. Cuijpers and Holtslag [CuEtAl98] have derived an expression to the counter-gradient from the LES results:

$$\gamma_\varphi = \beta_g \ell_w \frac{w_*^2}{\sigma_w} \frac{\varphi_*}{h}, \qquad \text{with:} \quad \varphi_* = \frac{1}{h w_*} \int_0^h \overline{w'\varphi'}\, dz\,. \tag{29.11}$$

where β_g is an experimental constant. The counter-gradient depends on the wind variance parameterization σ_w^2, the mixing length ℓ_w, and the quantity φ^*. The expressions for wind variance σ_i^2 and mixing length ($\ell_i = K_{\alpha\alpha}/\sigma_i$) are calculated from the Taylor's theory [Ca10]:

$$\sigma_i^2 = \frac{0.98 c_i}{(f_m)_i^{2/3}} \left(\frac{\psi_\epsilon}{q_i}\right)^{2/3} \left(\frac{z}{h}\right)^{2/3} w_*^2 \tag{29.12}$$

$$\ell_w = 0.2h \left[1 - \exp\left(-4\frac{z}{h}\right) - 0.003 \exp\left(8\frac{z}{h}\right)\right] \tag{29.13}$$

where $c_i = 0.3$ for u and 0.4 for (v, w), $f_m = 0.33$ is the frequency for the spectral peak, $q = (f_m)_i \, (f_m)_{n,i}^{-1}$ is a stability function. The dissipation function was derived by Campos Velho et al. [CaEtAl96]:

$$\psi_\epsilon = 3\left(1 - \frac{z}{h}\right)\left(\frac{z}{h}\right)\left[1 - \exp\left(4\frac{z}{h}\right) - 0.0003 \exp\left(8\frac{z}{h}\right)\right]^{-1}. \tag{29.14}$$

Different values are obtained in Equation (29.11) if different turbulence model is used. Therefore, the constant β_g is a parameter to be calibrated according to the parameterization applied. Cuijpers and Holtslag [CuEtAl98] have used $\beta_g = 1.5$, and Roberti and co-authors [RoEtAl04] employed $\beta_g = 0.07$. The value $\beta_g = 0.02$ provides the better results for our approach using Taylor's theory [We16].

29.3 Meso-Scale Atmospheric Model: BRAMS

BRAMS is employed for CPTEC (Portuguese acronym for Center for Weather Forecasting and Climate Studies), a division of INPE (National Institute for Space Research, Brazil), as the operational system for numerical weather forecasting over South America. The prediction system deals with 5 km of horizontal resolution, executed on Cray XE6 massively parallel computer. Operational forecasting uses 9600 processing cores[1]. BRAMS can be also configured as the operational environmental prediction system – the old version of the environmental system was called as CCATT-BRAMS. The development for the BRAMS is a permanent feature [FrEtAl17].

The model is coded with finite differences, where type-C Arakawa grid is employed for solving the fully compressible non-hydrostatic equations. Other interesting feature is the multiple grid nesting scheme, allowing the model equations to be solved simultaneously on any number of two-way interacting computational meshes of increasing spatial resolution. In the type-C grid, the variables temperature (T), pressure (p), and density (ρ) are defined in the center of a computational cell, and the wind components (u, v, w) are described in center of the cell edge. BRAMS features include an ensemble version of a deep and shallow cumulus scheme based on the mass flux approach. The surface model is the LEAF (Land Ecosystem Atmosphere Feedback) model, representing the surface-atmosphere interaction.

29.4 Simulation with BRAMS on the Amazon Region

The counter-gradient parameterization presented was codified in the BRAMS version 3.2, the same version used in Barbosa's studies [Ba07]. The simulation is compared with the measurements obtained in the LBA experiment.

The simulation domain embraces the North part of Brazil – see Figure 29.1. The LBA observations are collected inside the red box (left), and the experimental sites are marked with the yellow points (right): Biological reserve Rebio Jaru and Farm "Nossa senhora Aparecida" (Farm NSA) in the Rondonia state (Brazil).

Rebio Jaru (RJ): located at 100 km North to the Ji-Paraná city. This area is part of the rain forest. There is a tower 60 m high, installed at the end of the year 1998 placed at 10 °04′42″S and 61 °56′01″W.

Farm Nossa senhora Aparecida (RA) – site ABRACOS (Anglo Brazilian Amazonian Climate Observation Study): located at 50 km west direction from the Ji-Paraná city. The site characterizes a deforestation area, and from 1991 it has a pasture covering the surface. The farm has a tower placed at 10 °45″S and 62 °22″W.

[1]The Cray XE6 supercomputer installed CPTEC-INPE: 1280 processing nodes and 30,720 cores (2 processors per node and 12-cores for each processor).

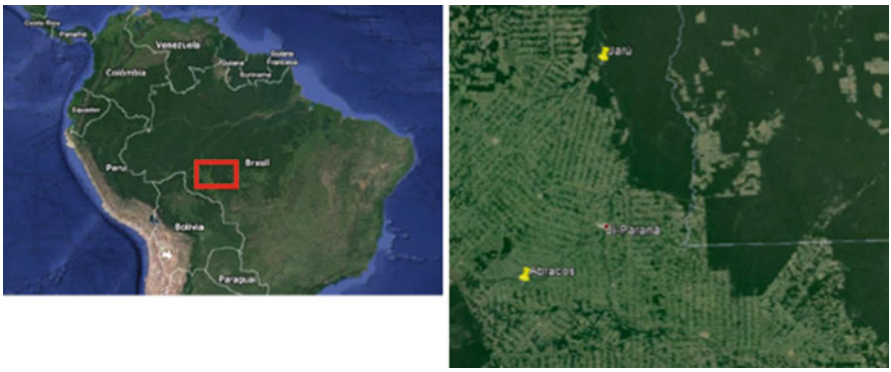


Fig. 29.1 Satellite images: the red box indicates the region of measurements, and yellow points indicate the location for LBA's stations (Rebio Jaru and Abracos)

BRAMS was initiated with data from ECMWF re-analysis. The meteorological variables in the latter dataset are: temperature, moisture, geopotential, zonal, and meridional winds, with space resolution of 2.5×2.5 degrees. The LBA data are also merged to the re-analysis for providing initial and boundary conditions. The simulation covers the period without rainfall. The first day for the simulation started at 00UTC February 10th up to 12th, 1999, performing 48 hours of simulation.

BRAMS was configured with 194 and 100 mesh points for Longitude and Latitude, respectively. The horizontal resolution is 20 km over a stereograph polar grid, with center at Latitude 10S and Longitude 61W. For vertical direction, 40 mesh points were defined, with finer resolution close to the surface. Time discretization $\Delta t = 30$ seconds.

The boundary layer height h is determined by using different approaches depending on the stability condition. For neutral/stable conditions, the formulation presented by Zilitinkevich is applied [Zi72]:

$$h = B_v \, u_*^{3/2} \tag{29.15}$$

where $B_v = 2.4 \times 10^3 \ \mathrm{m}^{-1/2} \, \mathrm{s}^{3/2}$. Under convective conditions, the approach suggested by Vogelezang and Holtslag [VoEtAl96] is employed:

$$Ri_g = \frac{(g/\theta_{v_s}) \, (\theta_{v_h} - \theta_{v_s}) \, (h - z_s)}{(u_h - u_s)^2 + (v_h - v_s)^2} \tag{29.16}$$

being $Ri_g = 0.4$ the critical Richardson number, and $z_s = 0.1h$ the reference value to express the values of the horizontal wind components and temperature.

Figure 29.2 shows the short wave radiation for the Taylor approach alone, the same simulation with counter-gradient term, and observations, considering two days of simulation for the Rebio Jaru experimental site. Figure 29.3 displaysthe vertical

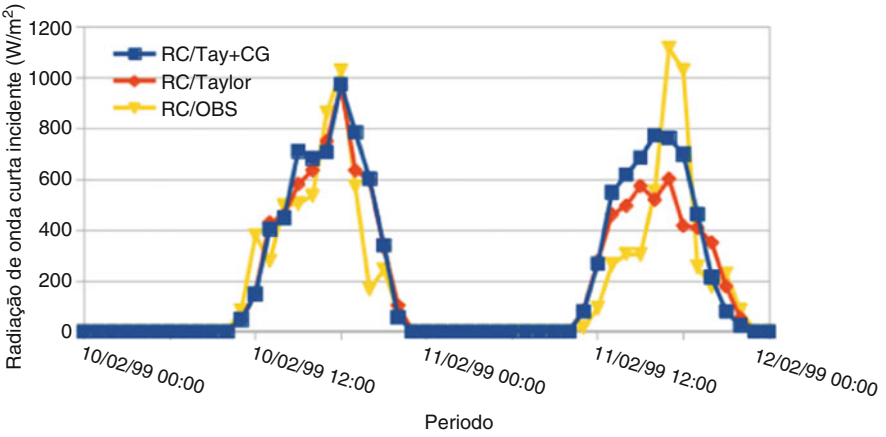


Fig. 29.2 Short wave radiation for Rebio Jaru station: Taylor's theory (Taylor), Taylor + Counter-Gradient (Taylor + CG), and observations

profiles for the potential temperature for different parameterizations for turbulence at the end of the simulation. The comparison with the observation shows similar results for all turbulent schemes.

Two snapshots for the wind field over the simulated region are shown in Figure 29.4. It is important to consider the wind influence, just to understand if the atmospheric dynamics response is due to the surface representation or an effect associated with the wind drag. During the day 10/Feb/1999 at 00 UTC (Figure 29.4a), sites RJ and RA are under forest influence. For the day 11/Feb/1999 at 00 UTC (Figure 29.4b), the sites have influence from the pasture covering.

29.5 Final Remarks

The paper describes a formulation for the counter-gradient term, where the Taylor's theory was applied. According to Figure 29.3, the simulation results were similar for all parameterizations used. However, the Smagorinsky's scheme requires the calculation of the vertical/horizontal deformation tensors for each grid point and the Brunt-Vaisäla frequency (depending on the temperature vertical gradient), and the Mellor-Yamada's method introduces more 12 new additional partial differential equations and parameterizations for the third order Reynolds tensors. Therefore, both latter approaches have a higher computational effort than Taylor's schemes.

With addition to the new term, the Taylor's parameterization can also simulate a counter-gradient flow, and the described parameterization is already codified to the BRAMS version 5.2 [FrEtAl17].

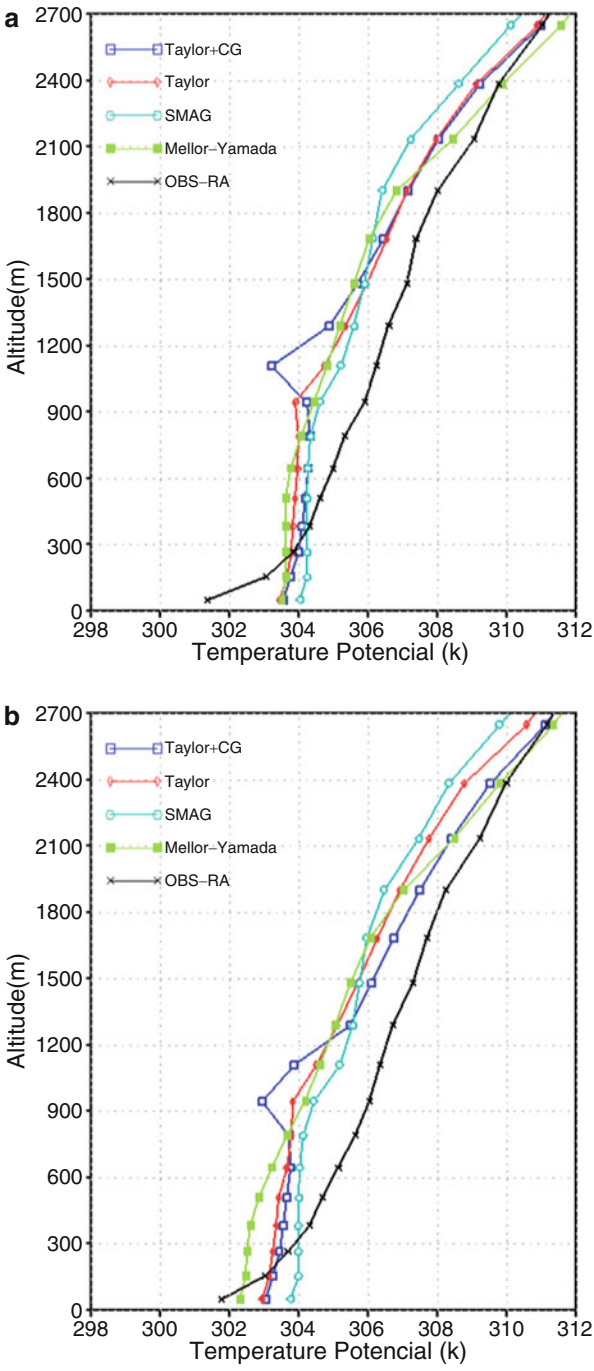


Fig. 29.3 Vertical potential temperature profiles for Taylor's theory (Taylor), Taylor + Counter-Gradient (Taylor + CG), Smagorinsky (SMAG), Mellor-Yamada, radiosonde data (OBS), day 12/Feb/1999 at 00UTC: (**a**) ABRACOS, (**b**) Rebio Jaru

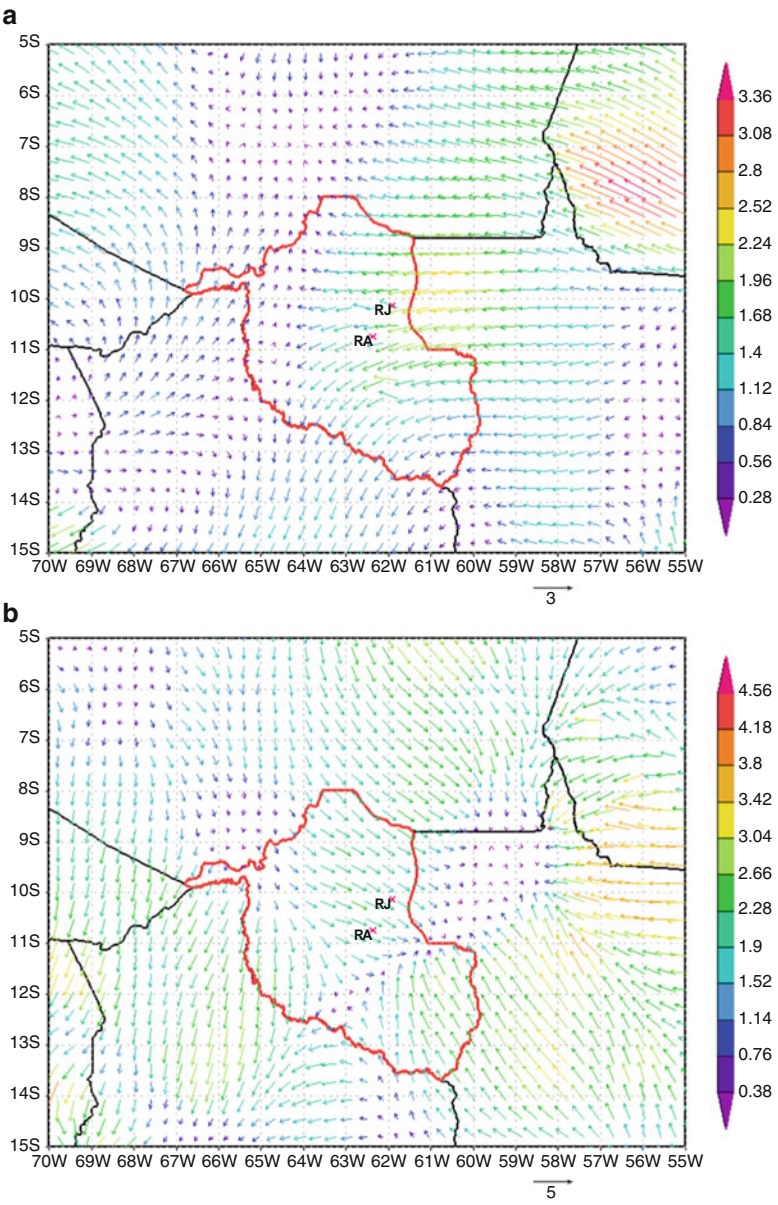


Fig. 29.4 Wind field from the BRAMS (Taylor + CG): (**a**) day 10/Feb/1999 at 00 UTC, (**b**) day 11/Feb/1999 at 00 UTC

Considering the radiation for long (not shown) and short waves the counter-gradient approach had a slightly better representation [We16]. But, more simulations are needed in order to have a definitive conclusion.

References

[Ba07] Barbosa, J.P.S.: New atmospheric turbulence parameterizations for the BRAMS. M.Sc. thesis in Applied Computing (INPE), São José dos Campos, SP, Brazil (2007) (in Portuguese)

[CaEtAl96] Campos Velho, H.F., Degrazia, G.A., Carvalho, J.C.: A new formulation for the dissipation function under strong convective regime. Brazilian Congress on Meteorology, vol. 3. Campos do Jordao, Brazil (1996)

[CaEtAl98] Campos Velho, H.F., Holtslag, A.M., Degrazia, G., Pielke, R.Sr.: New parameterizations in RAMS for vertical turbulent fluxes. Technical Report, Colorado State University, Fort Collins (CO), USA (1998)

[Ca10] Campos Velho, H.F.: Mathematical modeling in atmospheric turbulence – short-course. Braz. Soc. Compuat. Appl. Math. (2010). ISSN 2175-3385 (in Portuguese)

[CuEtAl98] Cuijpers, J., Holtslag, A.M.: Impact of skewness and nonlocal effects on scalar and boundary fluxes in convective boundary layers. J. Atmos. Sci. **51**, 151–162 (1998)

[De66] Deardorff, J.W.: The counter-gradient heat flux in the lower atmosphere and in the laboratory. J. Atmos. Sci. **23**, 503–506 (1966)

[De72] Deardorff, J.W.: Theoretical expression for the countergradient vertical heat flux. J. Geophys. Res. **77**, 5900–5904 (1972)

[DeEtAl00] Degrazia, G.A., Anfossi, D., Carvalho, J.C., Mangia, C., Tirabassi, T., Campos Velho, H.F.: Turbulence parameterisation for PBL dispersion models in all stability conditions. Atmos. Environ. **21**, 3575–3583 (2000)

[DeEtAl92] Degrazia, G.A., Moraes, O.L.L.: A model for eddy diffusivity in a stable boundary layer. Bound.-Layer Meteorol. **58**, 205–214 (1992)

[DeEtAl97] Degrazia, G.A., Campos Velho, H.F., Carvalho, J.C.: Nonlocal exchange coefficients for the convective boundary layer derived from spectral properties. Beiträge zur Phys. Atmosphäre **70**, 57–64 (1997)

[FrEtAl09] Freitas, S.R., Longo, K.M., et al.: The coupled aerosol and tracer transport model to the Brazilian developments on the regional atmospheric modeling system (CATT-BRAMS) – Part 1: model description and evaluation. Atmos. Chem. Phys. **9**, 2843–2861 (2009)

[FrEtAl17] Freitas, S.R., Panetta, J., Longo, K.M., et al.: The Brazilian developments on the regional atmospheric modeling system (BRAMS 5.2): an integrated environmental model tuned for tropical areas. Geophys. Model Dev. **130**, 1–55 (2017)

[LoEtAl10] Longo, K.M., Freitas, S.R., et al.: The coupled aerosol and tracer transport model to the Brazilian developments on the regional atmospheric modeling system (CATT-BRAMS) – Part 2: model sensitivity to the biomass burning inventories. Atmos. Chem. Phys. **10**, 2843–2861 (2010)

[MeEtAl82] Mellor, G. L., Yamada, T.: Development of a turbulence closure model for geophysical fluid problems. Rev. Geophys. **20**, 851–875 (1982)

[PiEtAl92] Pielke, Sr. R., Cotton, W.R., Walko, R.L., Tremback, C.J., Lyons, W.A., Grasso, L.D., Nicholls, M.E., Moran, M.D., Wesley, D.A., Lee, T.J., Copeland, J.: A comprehensive meteorological modeling system: RAMS. Meteorog. Atmos. Phys. **49**, 69–91 (1992)

[RoEtAl04] Roberti, D.R., Campos Velho, H.F., Degrazia, G.: Identifying counter-gradient term in atmospheric convective boundary layer. Inverse Prob. Eng. **12**, 329–339 (2004)

[Sm63] Smagorinsky, J.: General circulation experiments with the primitive equations: I. the basic experiment. Mon. Weather Rev. **91**, 99–164 (1963)

[Ta22] Taylor, G.I.: Diffusion by continuous movements. Proc. R. Soc. Lond. **20**, 196–212 (1922)

[Ta35] Taylor, G.I.: Statistical theory of turbulence. Proc. R. Soc. Lond. Ser. A Math. Phys. Sci. **151**, 421–444 (1935)

[VoEtAl96] Vogelezang, D.H.P., Holtslag, A.A.M.: Evaluation and model impacts of alternative boundary-layer height formulations. Bound. Layer Meteorol. **81**, 245–269 (1996)

[We16] Welter, M.E.S.: Counter-gradient term modeling for turbulent Parameterization in the BRAMS atmospheric model. M.Sc. Thesis on Applied Computing, INPE, São José dos Campos, SP, Brazil (2016) (In Portuguese)

[Zi72] Zilitinkevich, S.S.: On the determination of the height of the Ekman boundary layer. Bound. Layer Meteor. **3**, 141–145 (1972)

Index

C. Constanda et al. (eds.), *Integral Methods in Science and Engineering, Volume 2*,
DOI 10.1007/978-3-319-59387-6