

José António Tenreiro Machado
Dumitru Baleanu
Albert C. J. Luo

Nonlinear and Complex Dynamics

Applications in Physical, Biological,
and Financial Systems

Nonlinear and Complex Dynamics

José António Tenreiro Machado • Dumitru Baleanu
Albert C.J. Luo

Nonlinear and Complex Dynamics

Applications in Physical, Biological,
and Financial Systems

José António Tenreiro Machado
Department of Electrical Engineering
Institute of Engineering of Polytechnic
of Porto
Porto, Portugal
jtm@isep.ipp.pt

Dumitru Baleanu
Faculty of Art and Sciences
Department of Mathematics and Computer
Sciences
Cankaya University
Ankara, Turkey
dumitru@cankaya.edu.tr

Albert C.J. Luo
Department of Mechanical
and Industrial Engineering
Southern Illinois University
Edwardsville, IL, USA
aluo@siue.edu

ISBN 978-1-4614-0230-5 e-ISBN 978-1-4614-0231-2
DOI 10.1007/978-1-4614-0231-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011934799

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

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

Preface

Nonlinear Science and Complexity (NSC) presently plays a dynamic and active role in nonlinear physics and mathematics. In recent years, the significant achievement in such fields including nonlinear dynamics, chaos, discontinuous systems, fractional dynamics, economical, social and biological systems, and many other apparently unrelated fields, proved to be manifestations of the NSC paradigm.

Bearing these facts in mind, it was decided to organize a scientific event dedicated to these emerging concepts. This book presents recent developments, discoveries, and progresses in NSC. The aims of the selected papers are to present the fundamental and frontier theories and techniques for modern science and technology, and to stimulate more research interest for exploration of nonlinear science and complexity. The studies focus on fundamental theories and principles, analytical and symbolic approaches, and computational techniques in nonlinear physical science and nonlinear mathematics. This book, *Nonlinear Dynamics of Complex Systems: Applications in Physical, Biological and Financial Systems*, addresses specifically these areas.

The editors hope that this collection of chapters may be useful and fruitful for scholars, researchers, and advanced technical members of the industrial laboratory facilities for developing new tools and products.

The editors thank the Rector and the President of the Board of Trustee of Cankaya University as well as the Scientific and Technological Research Council of Turkey for the support needed to hold the discussions and debates and all colleagues for sharing their expertise and knowledge.

Porto, Portugal
Ankara, Turkey
Edwardsville, IL, USA

José António Tenreiro Machado
Dumitru Baleanu
Albert C.J. Luo

Contents

**Part I Nonlinear and Complex Dynamics: Applications
in Physical Systems**

On the Mechanisms of Natural Transport in the Solar System	3
Yuan Ren, Josep J. Masdemont, Gerard Gómez, and Elena Fantino	
A Method to Design Efficient Low-Energy, Low-Thrust Transfers to the Moon	15
Giorgio Mingotti, Francesco Topputo, and Franco Bernelli-Zazzera	
Low-Energy Earth-to-Halo Transfers in the Earth–Moon Scenario with Sun-Perturbation	39
Anna Zanzottera, Giorgio Mingotti, Roberto Castelli, and Michael Dellnitz	
On the Relation Between the Bicircular Model and the Coupled Circular Restricted Three-Body Problem Approximation	53
Roberto Castelli	
Adaptive Remeshing Applied to Reconfiguration of Spacecraft Formations	69
Laura Garcia-Taberner and Josep J. Masdemont	
A Cartographic Study of the Phase Space of the Elliptic Restricted Three Body Problem: Application to the Sun–Jupiter–Asteroid System	83
Cătălin Galeş	
Parameter Identification of the Langmuir Model for Adsorption and Desorption Kinetic Data	97
Dumitru Baleanu, Yeliz Yolcu Okur, Salih Okur, and Kasim Ocakoglu	

Effects of Suspended Sediment on the Structure of Turbulent Boundary Layer	107
H.P. Mazumdar, S. Bhattacharya, and B.C. Mandal	
A Renormalization-Group Study of the Potts Model with Competing Ternary and Binary Interactions	117
Nasir Ganikhodjaev, Seyit Temir, Selman Uğuz, and Hasan Akin	
The Mechanical Properties of CaX_6 ($\text{X} = \text{B}$ and C)	127
Sezgin Aydin and Mehmet Şimşek	
The System Design of an Autonomous Mobile Waste Sorter Robot	135
Ahmet Mavus, Sinem Gozde Defterli, and Erman Cagan Ozdemir	
Evidence of the Wave Phase Coherence for Freak Wave Events	147
Alexey Slunyaev	
Quantum Mechanical Treatment of the Lamb Shift Without Taken into Account the Electric Charge	159
Voicu Dolocan, Andrei Dolocan, and Voicu Octavian Dolocan	
Determining the Climate Zones of Turkey by Center-Based Clustering Methods	171
Fidan M. Fahmi, Elçin Kartal, Cem İyigün, Ceylan Yozgatligil, Vilda Purutcuoğlu, İnci Batmaz, Murat Türkeş, and Gülser Köksal	
The Determination of Rainy Clouds Motion Using Optical Flow	179
O. Raaf and A. Adane	
Hydrodynamic Modeling of Port Foster, Deception Island (Antarctica) ...	193
Juan Vidal, Manuel Berrocoso, and Bismarck Jigena	
Part II Nonlinear and Complex Dynamics: Applications in Biological Systems	
Localized Activity States for Neuronal Field Equations of Feature Selectivity in a Stimulus Space with Toroidal Topology	207
Evan C. Haskell and Vehbi E. Paksoy	
Intrinsic Fractal Dynamics in the Respiratory System by Means of Pressure–Volume Loops	217
Clara M. Ionescu and J. Tenreiro Machado	
Part III Nonlinear and Complex Dynamics: Applications in Financial Systems	
Forecasting Project Costs by Using Fuzzy Logic	231
M. Bouabaz, M. Belachia, M. Mordjaoui, and B. Boudjema	

Why You Should Consider Nature-Inspired Optimization Methods in Financial Mathematics	241
A. Egemen Yilmaz and Gerhard-Wilhelm Weber	
Desirable Role in a Revenue-Maximizing Tariff Model with Uncertainty	257
Fernanda A. Ferreira and Flávio Ferreira	
Can Term Structure of Interest Rate Predict Inflation and Real Economic Activity: Nonlinear Evidence from Turkey?	269
Tolga Omay	
Licensing in an International Competition with Differentiated Goods	295
Fernanda A. Ferreira	
Multidimensional Scaling Analysis of Stock Market Indexes	307
Gonçalo M. Duarte, J. Tenreiro Machado, and Fernando B. Duarte	
Index	323

Part I
Nonlinear and Complex Dynamics:
Applications in Physical Systems

On the Mechanisms of Natural Transport in the Solar System

Yuan Ren, Josep J. Masdemont, Gerard Gómez, and Elena Fantino

1 Introduction

Natural transport is a phenomenon in which particles of natural material transfer from their original orbit to very distant locations in the solar system. It is considered a crucial issue in the material exchange between terrestrial planets. The PCR3BP is a chaotic system, as the motion of the third body is highly sensitive to initial conditions and parameter values. In this system, the natural transport shows very different behaviors from those observed in the two-body model due to the long-term perturbations caused by the second primary and the effect of close passages. The natural transport has been investigated by many authors, in the general framework as well as relative to specific cases: for the Earth–Moon transport the reader is referred to [1, 2, 8], for the Mars–Earth transport to [4–6], and in general to [3]. In this paper, two types of natural transport in the PCR3BP are investigated: the short-time natural transport within two coplanar coupled PCR3BPs, i.e. direct connections between the manifolds of a pair of PCR3BPs, and the long-time natural transport in the PCR3BP, based on the analysis of the chaotic motion in this model.

For a definition of the adopted dynamical models (i.e. the PCR3BP, and the coupled PCR3BPs), the reader is referred to [7, 9, 10]. The analysis of the short-time natural transport between planets in solar system is presented in Sect. 2), whereas in Sect. 3 the process to determine the long-time natural transport is described. The natural transport from Mars to the Earth is used as an example throughout the paper. However, the method can be extended to other pairs of planets or other three-body systems.

Y. Ren (✉)

IEEC & Departament de Matemàtica Aplicada I, ETSEIB, Universitat Politècnica de Catalunya,
Diagonal 647, 08028 Barcelona, Spain
e-mail: yuan.ren@upc.edu

2 Short-Time Natural Transport

The short-time natural transport is based on the existence of heteroclinic connections between libration point orbits (LPOs), and in particular Lyapunov orbits, of a pair of consecutive Sun–planet PC3BPs, i.e. such that the two involved planets are on consecutive orbits in the solar system. Initially moving on the unstable manifold of an LPO of the departure PC3BP, the third body passes onto the stable manifold of an LPO of the arrival PC3BP and eventually reaches the vicinity of the corresponding planet. The dynamical model switch happens in the region where the two given invariant manifolds intersect in phase space. The existence of such intersection is the necessary condition for the specific short-time transport to take place. Figure 1 illustrates the outer branches of the stable and unstable manifolds of a Sun–Earth L_2 Lyapunov orbit and the inner branches of the stable and unstable manifolds of a Sun–Mars L_1 Lyapunov orbit. For convenience, each invariant manifold has been drawn in the respective synodic reference frame, a simplification that in the present case does not affect the values of the distances of the two objects from the Sun. Figure 1 shows that the two sets of objects stay well separated from each other. Computations extended over much longer times indicate that the intersection between a Sun–Mars and a Sun–Earth invariant manifold never occurs.

The manifolds of other types of LPOs look similar. In this respect, they can be represented and replaced by the invariant manifolds of the corresponding libration points. Figure 2 shows the invariant manifolds of the L_1 and L_2 points of a generic PCR3BP. Their minimum and maximum distances from the center of mass of the

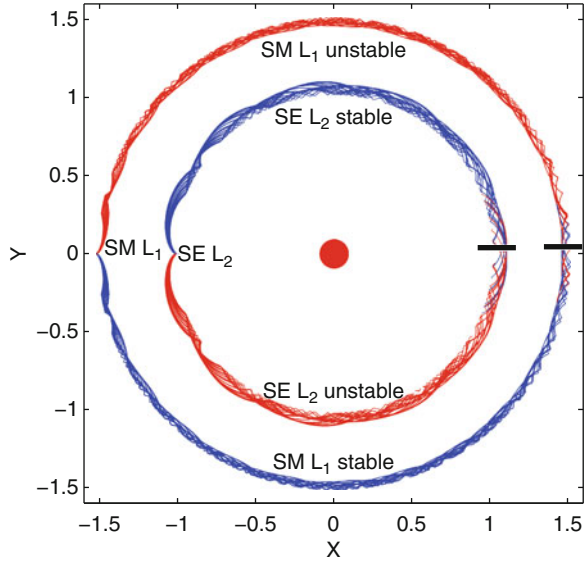


Fig. 1 Branches of stable and unstable invariant manifolds, respectively of periodic orbits around L_1 in the Sun–Mars PCR3BP and L_2 in the Sun–Earth PCR3BP. This plot has been obtained by superimposing the respective synodic reference frames

Fig. 2 Branches of the stable and unstable invariant manifolds of the L_1 and L_2 libration points in a PCR3BP. The minimum and maximum distances from the center of mass of the system are labelled $R[W_{L_1}^u]_{\min}$ and $R[W_{L_2}^u]_{\max}$, respectively

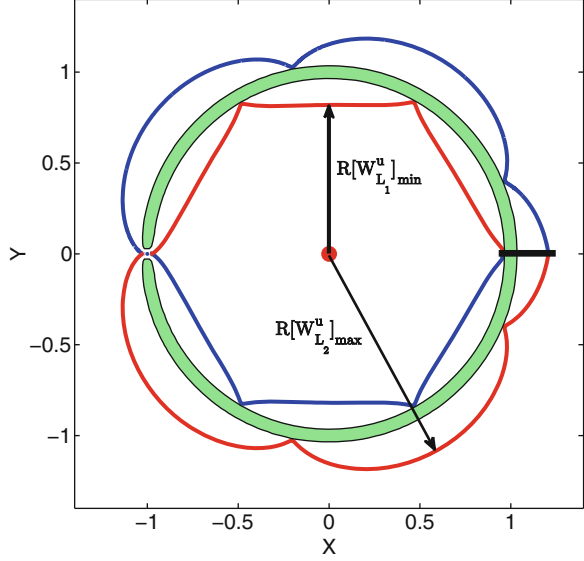


Table 1 The minimum and maximum heliocentric distances of the invariant manifolds of the L_1 and L_2 points of the Sun–planet PC3BPs. Quantities expressed in astronomical units. The superscript ‘a’ indicate the consecutive Sun–planet systems which satisfy the necessary condition for the short-time transfer

PCR3BP	$R[W_{L_1}^{u(s)}]$		$R[W_{L_2}^{u(s)}]$	
	Min	Max	Min	Max
Sun–Mercury	0.37547	0.38562	0.38858	0.39918
Sun–Venus	0.67152	0.71658	0.73011	0.78024
Sun–Earth	0.92328	0.98998	1.01008	1.08488
Sun–Mars	1.46684	1.51644	1.53094	1.58331
Sun–Jupiter	3.02493	4.85550	5.56589	9.46402 ^a
Sun–Saturn	6.63467 ^a	9.12494	9.99818	14.04464
Sun–Uranus	15.87518	18.75222	19.69233	23.49574 ^a
Sun–Neptune	23.35734 ^a	29.33805	30.89615	37.25573

system are labelled $R[W_{L_1}^u]_{\min}$ and $R[W_{L_2}^u]_{\max}$, respectively. When referred to a Sun–planet PCR3BP, such quantities constitute a good approximation of the minimum and maximum distances from the Sun, given that the mass ratio of any Sun–planet system is always a very small number. Table 1 reports the values of these quantities for all the Sun–planet PCr3BPs (except Sun–Pluto). The necessary condition for the short-time transfer can then be translated into the following inequality:

$$R_{\max}[W_{L_2}^{u(s)}]_{\text{inner}} \geq R_{\min}[W_{L_1}^{u(s)}]_{\text{outer}}. \quad (1)$$

It states that the short-time transport between consecutive Sun–planet systems is possible only if the maximum heliocentric distance of the inner problem is larger or equal to the minimum heliocentric distance of the outer problem, thus expressing in a quantitative way the intersection condition. According to Table 1, only two short-time connections are possible in the solar system, i.e. Jupiter–Saturn and Uranus–Neptune (denoted by asterisks). We conclude that not only is the short-time transport between Mars and the Earth not possible but also that this is the most common situation in our planetary system.

3 Long-Time Natural Transport

The short-time transport concept cannot comprehensively explain the exchange of natural material throughout our solar system, and it completely fails in the case of the terrestrial planets. This imposes to evaluate alternative mechanisms. The scenario proposed in this contribution is the long-time transport based on the study of the chaoticity within a PCR3BP, and in particular the analysis of transport between fixed points on the Poincaré maps or lobe dynamics.

The Mars-to-Earth transport is used as an example. Apart from the Sun, the gravity of the Earth, Mars and Jupiter are the dominant forces in the region between the orbits of Mars and the Earth (Fig. 3). For this reason, the long-time natural transport here discussed is based on the Sun–Jupiter PCR3BP. The chaoticity of the system is analysed by considering Poincaré maps drawn at several energy levels.

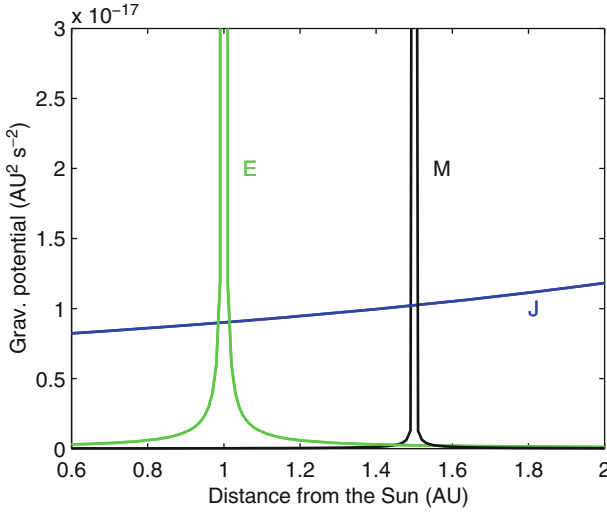
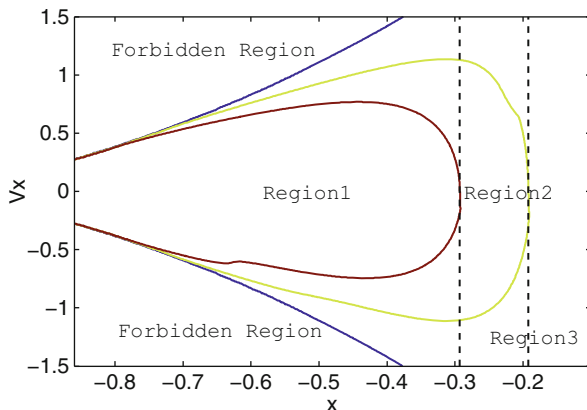


Fig. 3 The relative importance among the gravity of the Earth, Mars and Jupiter in the heliocentric region including the orbits of the Earth and Mars

Fig. 4 Subdivision of the initial conditions distributed on the Poincaré section for $C = 3.03$ into regions describing the extension (perihelion, aphelion) of the corresponding heliocentric inertial orbits relative to the orbits of the Earth and Mars



The adopted Poincaré section is the $x\dot{x}$ plane in the Sun–Jupiter PC3BP synodic barycentric reference frame, supplemented by the relations $\tan^{-1}(\dot{y}/x) = 180^\circ$ and $\dot{y} < 0$. In this way, a point on the Poincaré section corresponds to a well defined state, given that $y = 0$, and \dot{y} is straightforwardly determined by the previous conditions, once the Jacobi constant C is given. By forward propagating a given initial state over a short time (i.e. $< 2\pi$ in Sun–Jupiter adimensional time units), a short trajectory segment in synodic coordinates is obtained, corresponding to an almost complete elliptical orbit in inertial space. By propagating a dense set of initial conditions for $C = 3.03$, we have determined the perihelion and aphelion distances of the corresponding inertial orbits, based on which we have divided the Poincaré section into regions (Fig. 4): the two “forbidden” regions correspond to impossible motion at the given energy level; “region 1” contains initial states of inertial orbits that do not intersect the orbits of the Earth and Mars; the initial states in “region 2” produce orbits which only intersect the orbit of Mars; finally, the initial states in “region 3” correspond to motions which have intersections with the orbit of the Earth. Therefore, finding transport from “region 2” to “region 3” on the Poincaré section is equivalent to finding transport from Mars to the Earth in configuration space.

3.1 Poincaré Maps

Figure 5 shows four Poincaré maps for the Sun–Jupiter PCR3BP, characterized by decreasing value of C . Each of them has been obtained by setting a mesh grid on the Poincaré section, forward propagating the corresponding initial states and marking their successive intersections with the section. The four plots have been obtained from an initial mesh grid of 35 by 25 points such that $-0.86048 \leq x \leq -0.15$ and $-0.2 \leq \dot{x} \leq 0.2$, with $C = 3.14, 3.09, 3.06$ and 3.03 , respectively. They illustrate the first 10^3 returns. As such, they give an insight into the structure of the phase

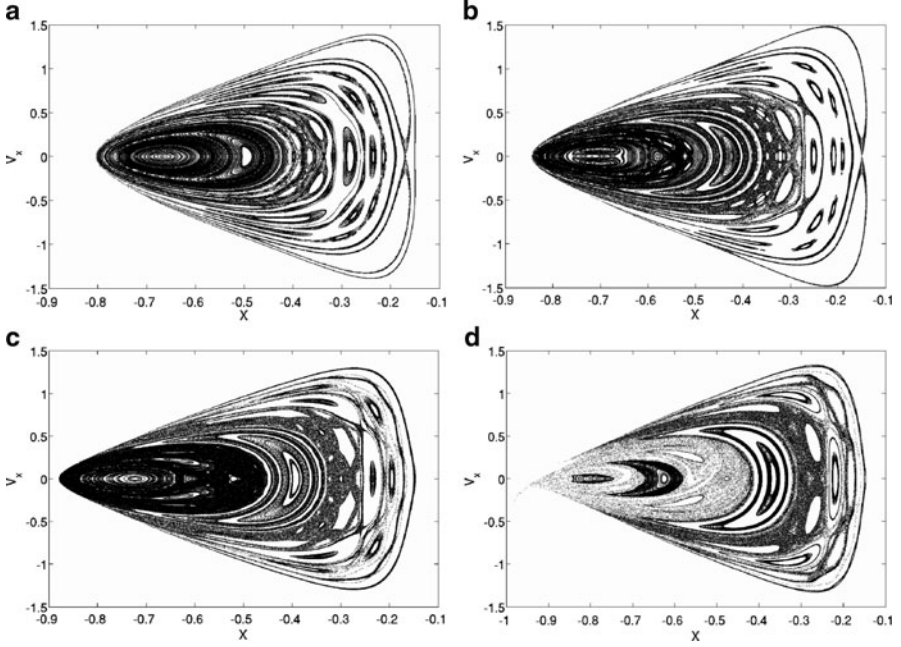


Fig. 5 Poincaré maps of the Sun–Jupiter PCR3BP corresponding to $C = 3.14$ (a), 3.09 (b), 3.06 (c), 3.03 (d)

space at varying energy. In particular, it is possible to detect islands (periodic orbits) and fixed points at their centers. When transformed into configuration space, the periodic orbits correspond to invariant tori and the fixed points to periodic orbits. The only possible transfer between different regions of the Poincaré map can only happen through the gaps in the chain of islands, i.e. through the chaotic sea. A close look at the four plots of Fig. 5 shows that the size of the chaotic sea increases as the Jacobi constant decreases (i.e. the energy increases), and in particular when $C = 3.03$, “region 2” and a part of “region 3” are in the chaotic sea, thus suggesting that at this energy level the long-time natural transport from Mars to the Earth may be possible.

3.2 Fixed Points and Their Manifolds

Fixed points can either be found at the center of the islands or in the chaotic sea, the latter case being associated to the transport. The most efficient way to identify the fixed points on the Poincaré map is by either the shooting method or the parallel shooting method, respectively when the order (periodicity, or number of distinct successive returns) of the fixed point is lower or greater than five, the reason being

the inability of the shooting method to converge to a high-order fixed point, with the tendency of detecting the nearest fixed point of order less than five.

The shooting method aims at zeroing the difference

$$\mathbf{F} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix}_{t_f} - \begin{bmatrix} x \\ \dot{x} \end{bmatrix}_{t_0} \quad (2)$$

between the final state (i.e. after a chosen number of returns) and the initial state on the Poincaré section. By setting $\mathbf{x} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix}_{t_0}$, the variation of \mathbf{F} can be expressed as

$$\frac{d\mathbf{F}}{d\mathbf{x}} = (\Phi - \mathbf{I}) \Delta \mathbf{x}, \quad (3)$$

where \mathbf{I} is the 2×2 identity matrix. Φ is a 2×2 matrix obtained as

$$\Phi_{2 \times 2} = \begin{bmatrix} \frac{\partial x_f}{\partial x_0} & \frac{\partial x_f}{\partial \dot{x}_0} & \frac{\partial x_f}{\partial \dot{y}_0} \\ \frac{\partial \dot{x}_f}{\partial x_0} & \frac{\partial \dot{x}_f}{\partial \dot{x}_0} & \frac{\partial \dot{x}_f}{\partial \dot{y}_0} \end{bmatrix}_{2 \times 3} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\partial \dot{y}_0}{\partial x_0} & \frac{\partial \dot{y}_0}{\partial \dot{x}_0} \end{bmatrix}_{3 \times 2}. \quad (4)$$

The 2×3 matrix in the above product represents the variation from the initial to the target Poincaré section. $\partial \dot{y}_0 / \partial x_0$ and $\partial \dot{y}_0 / \partial \dot{x}_0$ in the 3×2 matrix on the right account for the dependence of \dot{y}_0 on x_0 and \dot{x}_0 (through C). \mathbf{F} is driven to zero by Newton–Raphson. In practice, the process is divided into two steps: a first search is conducted with a large tolerance because different initial guesses tend to converge to the same fixed point, and the use of a strict tolerance would only increase the computation time. Then, the distance between the final points is checked: whenever two points are closer than twice the tolerance, one of them is simply the duplicate of the other and is eliminated. A successive application of the shooting method with a stricter tolerance allows to refine the estimation of the fixed points that have survived the first search.

The parallel shooting method introduces additional, intermediate Poincaré sections, as illustrated in Fig. 6: here the order of the fixed point is 7, and two intermediate sections are set (although for this order one section would be sufficient). Sections 2 and 3 represent respectively the second and the fourth crossing of the trajectory through the Poincaré section. The mesh grids are set on all the sections and the nodes on different sections are combined together to form a set of initial guesses. The side effect of this new method is a considerable increase in the computation time, but the strategy allows to detect fixed points to any order.

The stability of a fixed point depends on the modulus of the eigenvalues of its monodromy matrix. Since the monodromy matrix is a 2×2 matrix, it has two eigenvalues. A saddle point is a fixed point for which the two eigenvalues have modulus greater and smaller than unity, respectively. There are stable and unstable

Fig. 6 Sketch of the parallel shooting method

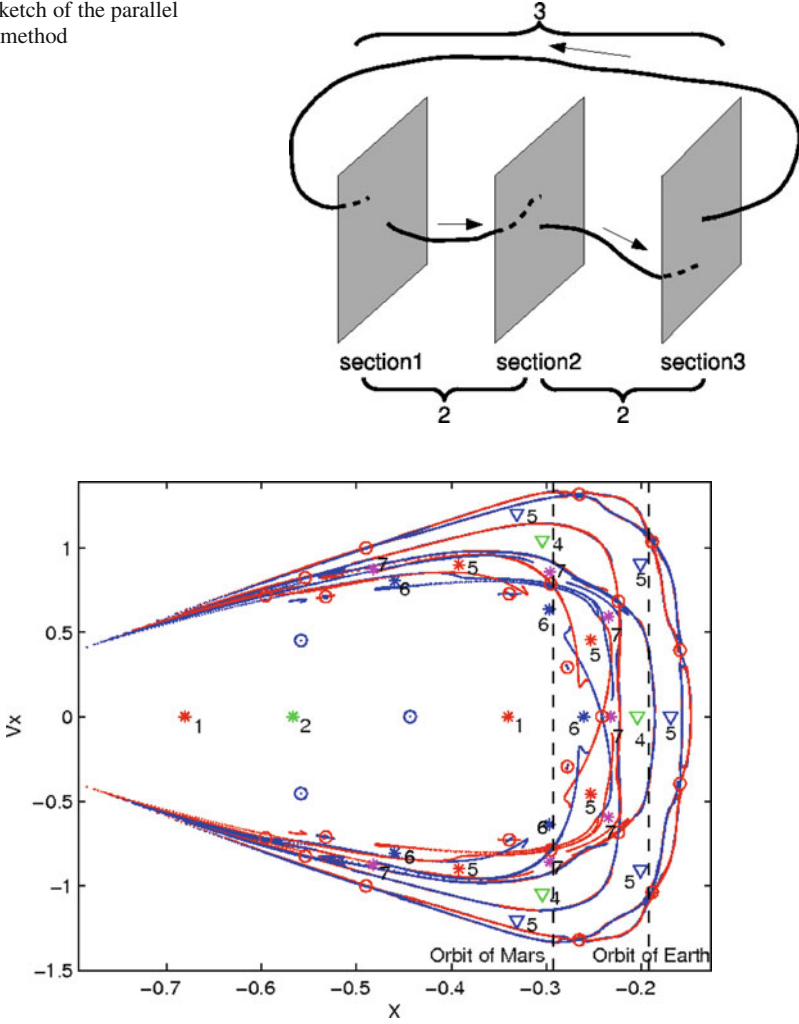


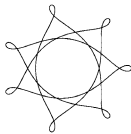
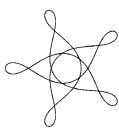
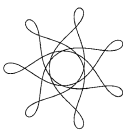
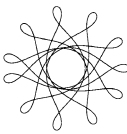
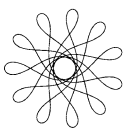
Fig. 7 Fixed points and their invariant manifolds in the Sun–Jupiter PCR3BP at $C = 3.0$: saddle points (○), stable fixed points (☆ and ▽) and their invariant manifolds

manifolds associated to a saddle point and the eigenvectors corresponding to the stable and unstable eigenvalues constitute the linear approximation of the stable and unstable manifolds, respectively.

Figure 7 shows all the fixed points identified in the Sun–Jupiter PCR3BP for $C = 3.0$. All the fixed points in the center of the islands are stable, whereas the saddle points appear at the gaps of the island chain.

Also the invariant manifolds of the fixed points are represented. There are five sets of saddle points. Their position on the Poincaré section, their eigenvalues and

Table 2 Position (x, \dot{x}) , stable and unstable eigenvalues (E^s and E^u) and period (in adimensional time units, $1 \text{ TU} \approx 12/2\pi \text{ years} = 1.9 \text{ years}$) for the detected saddle points in the Sun–Jupiter PCR3BP

Order	4	4	5	7	8
					
x	− 0.4432	− 0.2240	− 0.2412	− 0.3388	− 0.1591
\dot{x}	+0.0000	− 0.6839	+0.0000	+0.7269	+0.3938
E^u	− 1.3040	− 6.2870	+11.730	− 2.7930	+4.8820
E^s	− 0.7666	− 0.1590	+0.0852	− 0.3579	+0.2049
Period	25.1	12.6	18.9	31.5	25.2

the period of the corresponding periodic orbit in configuration space are reported in Table 2: an n -periodic orbit is identified by n fixed points, but for the sake of brevity, the table provides data only for one of them, the others being the result of iterating $n - 1$ times.

3.3 Lobe Dynamics

Figure 7 shows that the manifolds of one of the order-4 saddle points have intersections with the orbit of the Earth, the manifolds of the order-5 saddle point have intersections with the orbit of Mars, and these manifolds have intersections with each other. Thus, the natural transport can be realized with the aid of these manifolds: the particle leaves Mars following the unstable manifold of the order-5 orbit, then switches to the stable manifold of the order-4 orbit at the intersection of these two manifolds, passes through the gap in the chain of islands, simultaneously switches to the unstable manifold of the order-4 orbit and intersects the orbit of the Earth. The region enclosed by the stable and unstable manifolds is called “lobe.” The transport between regions of the phase space can be completely described by the dynamical evolution of the lobe. Hence, by choosing a set of initial states in the lobe, backward and forward propagation are performed until the transport from “region 2” to “region 3” (i.e. from the orbit of Mars to that of the Earth) is achieved. Figure 8 illustrates an example of lobe evolution: T represents the initial states near the intersection of the manifolds of the order-4 and order-5 saddle points. After 21 iterations in the forward direction, some of these states move from “region 2” to “region 3.” However, since such states were initially already very close to “region 3” (i.e. the perihelions of their initial orbits were very close to the orbit of the Earth), it means that the effect of the gravity of Jupiter is minor in this case.

By an additional backward propagation, an initial orbit with the perihelion far from the orbit of the Earth is sought, thus obtaining an overall transfer with a

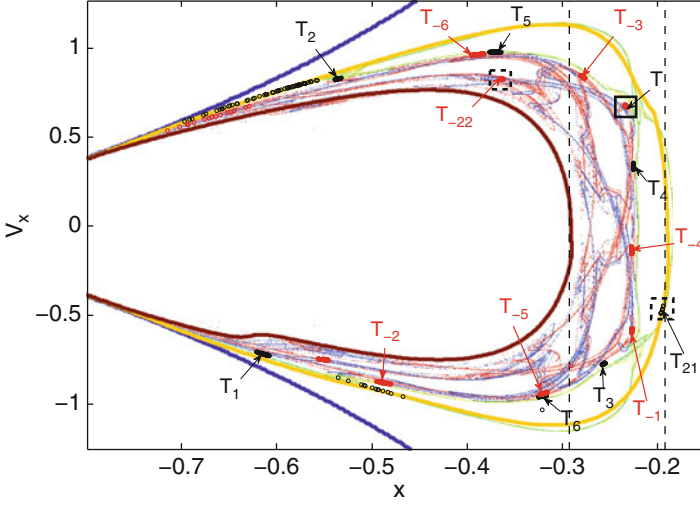
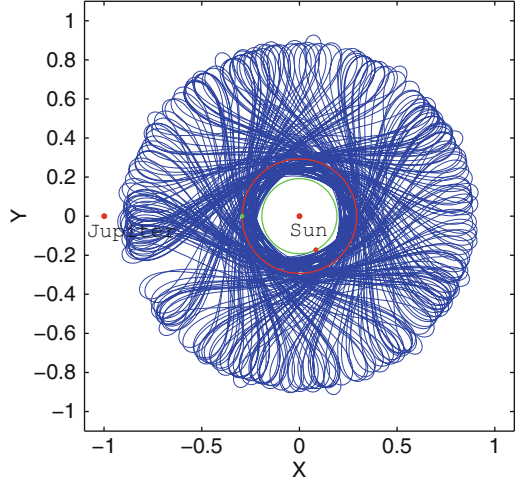


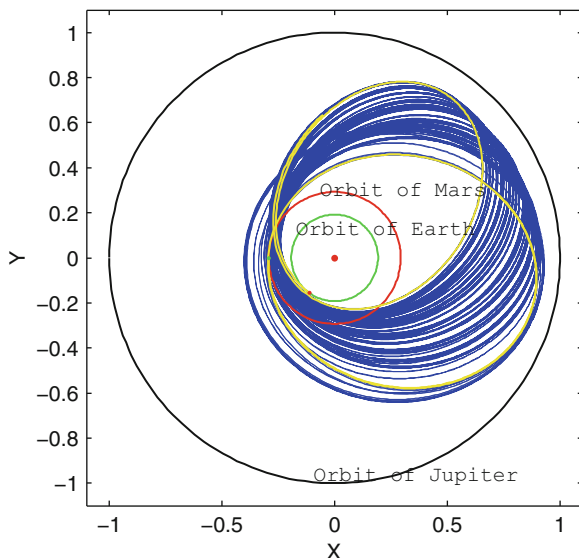
Fig. 8 Lobe evolution

Fig. 9 The Mars-to-Earth transfer orbit corresponding to the long-time natural transport described in the text for $C = 3.0$. In the synodic barycentric reference frame



significant effect of the perturbations of Jupiter. Figures 9 and 10 shows the orbit of a long-time natural transport from Mars to the Earth found by this procedure. The inertial frame view shows that the initial trajectory does not have any intersection with the orbit of the Earth. Then, as a result of the gravity of Jupiter, the orbit gradually evolves, eventually reaching the vicinity of the Earth's orbit. This specific transport involves 161 iterations corresponding to 671.32 time units in the Sun–Jupiter system, or 1282.6 years. We note, however, that the transfer orbit found by this method only guarantees a connection between the orbits of the Earth and

Fig. 10 The Mars-to-Earth transfer orbit corresponding to the long-time natural transport described in the text for $C = 3.0$. In the inertial heliocentric reference frame



Mars, not between the two planets. It certainly proves the possibility of the natural transport at this specific energy level, but the real transport from the surface of Mars to that of the Earth may require much longer times.

4 Conclusions

Two techniques for the analysis of the natural transport in the solar system have been presented. Each of them has a preferential framework of application. The first method achieves the short-time transport, with application to the design of spacecraft trajectories between Jupiter and Saturn and tours of Jupiter's icy moons. The second method is the long-time alternative for the cases in which the manifolds of the coupled PCR3BPs do not present direct intersection, such as the Mars-to-Earth case. Our numerical simulations prove that the second method is capable of describing the transfer of the martian ejecta with aphelion near the orbit of Jupiter to the region of the Earth's orbit. Nevertheless, the most appropriate framework for the study of the low-energy, natural transport between, e.g. Mars and the Earth is the n -body model (see [4, 6]), since the close passages at the Earth and Mars (not accounted for in the Sun-Jupiter PCR3BP) may be relevant. An attempt at describing the natural transport according to dynamical models of growing complexity (three-body, quasi-bicircular, n -body models) will be the subject of a subsequent publication.

Acknowledgements E. Fantino and Y. Ren have been supported by the Marie Curie Actions Research and Training Network AstroNet MCRTN-CT-2006-035151. G. Gómez and J.J. Masdemont have been partially supported by the grants MTM2006-05849/Consolider, MTM2009-06973 and 2009SGR859. The authors also acknowledge the use of EIXAM, the UPC Applied Math cluster system for research computing (see <http://www.mal.upc.edu/eixam/>), and in particular Pau Roldan for providing technical support in the use of the cluster.

References

1. Bollt, E.M., Meiss, J.D.: Controlling chaotic transport through recurrence. *Physica D: Nonlinear Phenomena*. **81**(3), 280–294 (1995)
2. Bollt, E.M., Meiss, J.D.: Targeting chaotic orbits to the Moon through recurrence. *Physics Letters A*. **204**(5/6), 373–378 (1995)
3. Dellnitz, M., Junge, O., Koon, W.S., Lekien, F., Lo, M.L., Marsden, J.E., Padberg, K., Preis, R., Ross, S.D., Thiere, B.: Transport in dynamical astronomy and multibody problem. *International Journal of Bifurcation and Chaos*. **15**(3), 699–727 (2005)
4. Gladman, B.: Destination Earth: Martian meteorite delivery. *Icarus*. **130**, 228–246 (1997)
5. Gladman, B.: The exchange of impact ejecta between terrestrial planets. *Science*. **217**, 1387–1392 (1996)
6. Gladman, B.: Delivery of planetary ejecta to Earth. Cornell University. (1996)
7. Gómez, G., Jorba, A., Llibre, J., Martinez, R., Masdemont, J., Simó, C.: Dynamics and Mission Design near Libration Points, Vol. I-IV. World Scientific Publishing Co., Singapore (2001)
8. Meiss, J.D., Ott, E.: Markov tree model of transport in area-preserving maps. *Physica D: Nonlinear Phenomena*. **20**(2/3), 387–402 (1986)
9. Parker, T.S., Chua, L.O.: Practical Numerical Algorithms for Chaotic Systems. Springer-Verlag, Berlin (1989)
10. Szebehely, V.: Theory of Orbits in the Restricted Problem of Three Bodies. Academic Press, New York (1967)

A Method to Design Efficient Low-Energy, Low-Thrust Transfers to the Moon

Giorgio Mingotti, Francesco Topputo, and Franco Bernelli-Zazzera

1 Introduction

Low-energy transfers to the Moon are being studied since the rescue of the Japanese spacecraft Hiten in 1991 [3]. In essence, a low-energy lunar transfer reduces the hyperbolic excess velocity upon Moon arrival, typical of a patched-conics approach. This process is called ballistic capture, and relies on a better exploitation of the gravitational nature ruling the transfer problem instead of the classic Keplerian decomposition of the solar system. The reduced speed relative to the Moon sets the trajectory to low energy levels, which in turn imply a reduced propellant mass needed to stabilize the spacecraft around the Moon.

It is known that the dynamical mechanism governing a class of exterior low energy transfers to the Moon is related to the structure of the invariant manifolds associated with the Lyapunov orbits about the collinear libration points [2]. In particular, a systematic method for the construction of low energy transfers using the knowledge of the phase space of the Sun–Earth and Earth–Moon systems is given in the works of the *Caltech and the Barcelona groups* [8, 9, 14].

It is worth mentioning that previous works have faced the combination of n -body dynamics with low-thrust propulsion. An interior ballistic capture state using low-thrust propulsion was found in [1]. This approach paved the way for the design of the trajectory for ESA’s SMART-1 mission [16]. Earth–Venus transfers have been obtained in [6] by combining invariant manifold dynamics and low-thrust, with set oriented methods. Moreover, there are also some examples of the integration of dynamical system theory and optimal control problems for the design of efficient low-energy, low-thrust transfers to the halo orbits [12, 13].

G. Mingotti (✉)

Institut für Industriemathematik, Universität Paderborn, Warburger Str. 100,
33098 Paderborn, Germany

e-mail: mingotti@math.uni-paderborn.de

In this paper, both efficient two-impulse transfers and the low-thrust version of the transfers described in [9] are presented, all of them starting from the same LEO with an impulsive maneuver given by the launcher. It is in fact possible to further reduce the propellant necessary to send a spacecraft to the Moon by exploiting both the simultaneous gravitational attractions of the Sun, the Earth, and the Moon, and the high specific impulse provided by the low-thrust engines (above 1,000 s). Nevertheless, including the low-thrust is not trivial and asks for a number of issues to be faced. It is of great importance to, for instance, overcome the loss of Jacobi integral, finding subsets of the phase space that lead to low-thrust ballistic capture (playing the separatrix-like role of the stable manifold associated with L_2 Lyapunov orbit of the Earth–Moon system), and summarize, using as few parameters as possible, all the reachable orbits that it is possible to target with the finite thrust magnitude available, like low lunar orbits (LLOs).

The purpose of this work is therefore to formulate a systematic approach for the design of efficient pure low energy as well as mixed invariant-manifold low-thrust transfers to low orbits around the Moon. Then, a comparison between the trajectories computed and some solutions found in literature is presented.

The remainder of the paper is organized as follows: first, a brief recall of the dynamics involved in the problem is given. Then, a description of the trajectory design strategy to low Moon orbits and the introduction of special *attainable sets* are given. Finally, the optimization problem is formulated and later the optimal solutions are discussed.

2 Design Strategy

With the *coupled restricted three-body problems approximation*, the four-body dynamics, characterizing the low-energy lunar transfers, is decomposed into two RTBPs, and the invariant manifolds of the Lyapunov orbits are computed. It is possible to show that, with a suitably chosen Poincaré section, the trajectory design is restricted to the selection of a single point on this section [9]. Flown backward, this initial condition gives rise to a trajectory close to the stable and unstable manifolds of the L_j , for $j = 1, 2$, Lyapunov orbits of the Sun–Earth system; integrated forward, a transit, lunar ballistic capture orbit (i.e. an orbit contained inside the stable manifold tube-like structure of the L_2 Lyapunov orbit of the Earth–Moon system) is achieved. A small Δv maneuver is eventually needed at this patching point, in order to match the energies of the two stages. With this approach, it is possible to find efficient Earth–Moon transfers like the ones described in [3].

The transfers studied in this work are defined as follows. The spacecraft is assumed to be initially on a circular parking orbit around the Earth at a height $h_E = 167$ km; then an impulsive maneuver, Δv_E , carried out by the launch vehicle, places the spacecraft on a translunar trajectory, performing a translunar insertion TLI. Two different typologies of mission are investigated, with respect to the propulsion adopted: (a) low-energy two-impulse transfers to LLOs: after the insertion, the

spacecraft flies ballistically under the dynamics of the problem until the Moon neighborhood, where a second impulsive maneuver inserts it on a stable low altitude orbit; (b) low-energy low-thrust transfers to LLOs: after the launch, the spacecraft can only rely on its low-thrust propulsion to reach a stable low-altitude orbit around the Moon.

As far as it concerns these transfers, the final orbit has moderate eccentricity, e , and periapsis/apoapsis, r_p/r_a , prescribed by the mission requirements. The transfer terminates when the spacecraft is at the periapsis of the final orbit around the Moon. While both e and r_p/r_a are assumed to be given, the orientation, i.e. the argument of periapsis, ω , of the final orbit around the Moon is not fixed.

In general, the two transfer families are achieved by optimizing, in a four-body scenario, a first guess derived in the coupled three-body problems approximation. From the prospective of this model, the transfer trajectory is conceived as made up of two distinct portions: the first, called Earth escape stage, is built in the Sun–Earth model, SE, whereas the second, called Moon capture stage, is defined in the Earth–Moon model, EM.

2.1 The Planar Circular Restricted Three-Body Problem

The motion of the spacecraft, m_3 , is studied in the gravitational field generated by the mutual circular motion of two primaries of masses m_1 , m_2 , respectively, about their common center of mass (see Fig. 3a). It is assumed that m_3 moves in the same plane of m_1 , m_2 under the following equations [19]:

$$\ddot{x} - 2\dot{y} = \frac{\partial \Omega}{\partial x}, \quad \ddot{y} + 2\dot{x} = \frac{\partial \Omega}{\partial y}, \quad (1)$$

where the auxiliary function is

$$\Omega(x, y, \mu) = \frac{1}{2}(x^2 + y^2) + \frac{1-\mu}{r_1} + \frac{\mu}{r_2} + \frac{1}{2}\mu(1-\mu), \quad (2)$$

and $\mu = m_2/(m_1 + m_2)$ is the mass parameter of the three-body problem. Equation (1) is written in a barycentric rotating frame with nondimensional units: the angular velocity of m_1 , m_2 , their distance, and the sum of their masses are all set to the unit value. It is easy to verify that the primary of mass $1 - \mu$, is located at $(-\mu, 0)$, whereas the smaller primary μ , is located at $(1 - \mu, 0)$; thus, the distances between m_3 and the primaries are:

$$r_1^2 = (x + \mu)^2 + y^2, \quad r_2^2 = (x + \mu - 1)^2 + y^2. \quad (3)$$

For fixed μ , the Jacobi integral reads

$$J(x, y, \dot{x}, \dot{y}) = 2\Omega(x, y, \mu) - (\dot{x}^2 + \dot{y}^2), \quad (4)$$

and, for a given energy C , it defines a three-dimensional manifold

$$F(C) = \{(x, y, \dot{x}, \dot{y}) \in \mathbb{R}^4 | J(x, y, \dot{x}, \dot{y}) - C = 0\}, \quad (5)$$

foliating the four-dimensional phase space. The projection of $F(C)$ on the configuration space (x, y) defines the Hill's curves bounding the allowed and forbidden regions associated with prescribed values of C . The vector field defined by (1) has five well-known equilibrium points, known as the Lagrange points, labeled Lj , $j = 1, \dots, 5$. This study deals with the portion of the phase space surrounding the two collinear points $L1$ and $L2$. In a linear analysis, these two points behave like the product *saddle* \times *center*. Thus, there exists a family of retrograde Lyapunov orbits and two-dimensional stable and unstable manifolds emanating from them [5, 10].

The system governed by (1) is used alternatively to describe the motion of the spacecraft either in the Sun–Earth (SE) or in the Earth–Moon (EM) system. The mass-parameter value assumed for these models are $\mu_{SE} = 3.0034 \times 10^{-6}$ and $\mu_{EM} = 1.2150 \times 10^{-2}$, respectively.

As for the SE model, the generic periodic orbit about Lj , $j = 1, 2$, is referred to as γ_j , whereas its stable and unstable manifolds are labeled $W^s(\gamma_j)$, $W^u(\gamma_j)$. Dealing with the EM model, the generic periodic orbit about Lj , $j = 1, 2$, is called λ_j , while its stable and unstable manifolds are named $W^s(\lambda_j)$, $W^u(\lambda_j)$.

3 Earth Escape Stage

If a value of Jacobi constant in the SE model, C_{SE} , is suitably chosen, there exists a unique Lyapunov orbit about both $L1$ and $L2$, labeled γ_1 and γ_2 , respectively. Assuming the energy values $C_{SE} \ll C_2$ such that both γ_1 and γ_2 exist, the Hill's regions are opened at both $L1$ and $L2$. Without any loss of generality, the Earth escape stage is constructed considering the dynamics around $L2$; using $L1$ instead of $L2$ is straightforward. The stable and unstable manifolds associated with γ_2 , $W^s(\gamma_2)$ and $W^u(\gamma_2)$, are computed starting from the Lyapunov orbit until a certain surface of section is reached.

Aiming at exploiting the structure of both $W^s(\gamma_2)$ and $W^u(\gamma_2)$, two surfaces of section are introduced to study their cuts at different stages. Section S_A , making an angle φ_A (clockwise) with the x -axis and passing through the Earth, is considered to cut $W^s(\gamma_2)$, whereas section S_B , inclined by φ_B (counterclockwise) on the x -axis and passing through the Earth, is assumed for $W^u(\gamma_2)$ (see Fig. 1a for the two manifolds topology, and see Fig. 2a for the two angles). The corresponding section curves, $\partial \Gamma_2^s$, $\partial \Gamma_2^u$, represented on the (r_2, \dot{r}_2) -plane, are diffeomorphic to circles (see Fig. 1b, $\partial \Gamma_2^s$, $\partial \Gamma_2^u$ are plotted on the (r_2, \dot{r}_2) -plane as $r_2 = y$, $\dot{r}_2 = \dot{y}$ for $x = 1 - \mu$, $\varphi_A = \varphi_B = \pi/2$).

Both Poincaré sections represent two-dimensional maps for the flow of the RTBP. Indeed, any point on these sections uniquely defines an orbit. This property

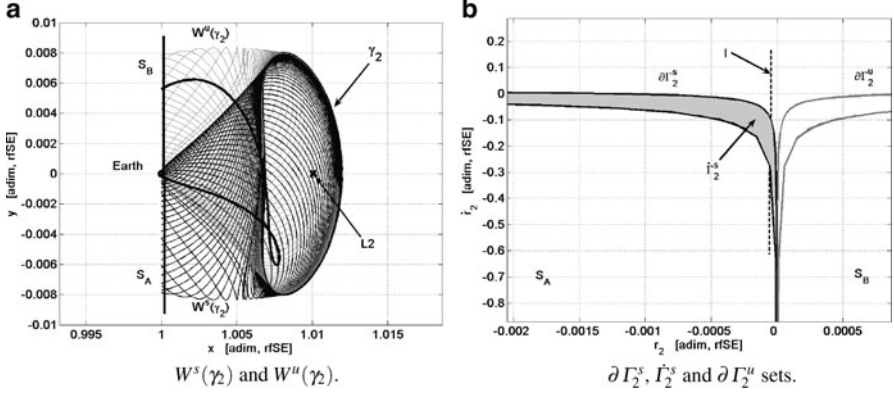


Fig. 1 Stable and unstable manifolds $W^s(\gamma_2)$, $W^u(\gamma_2)$ associated with the L_2 Lyapunov orbit γ_2 , and their section curves $\partial \Gamma_2^s$, $\partial \Gamma_2^u$, respectively. In (a), the **bold line** stands for a sample Earth escape trajectory. In (b), the set Γ_2^s (gray) is made up by the points of S_A that lie inside $\partial \Gamma_2^s$, whereas the line l (dashed) is the locus of points being at $h_E = 167$ km altitude above the Earth surface

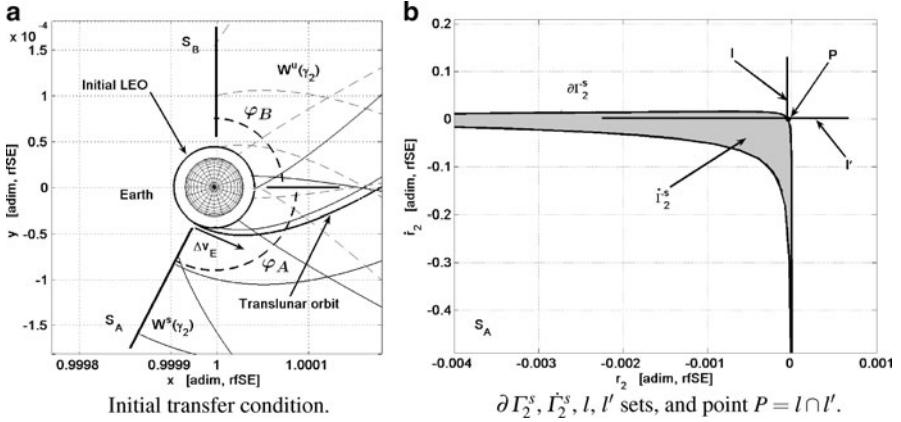


Fig. 2 Earth escape trajectory performed with a tangential Δv_E maneuver and its associated section point P

holds as both $F(C_{SE})$ and $S_{A,B}$ lower the dimension of the phase space to two. By definition, points on $\partial \Gamma_2^s$ generate orbits that asymptotically approach γ_2 in forward time. Points inside Γ_2^s give rise to transit orbits that pass from the Earth region to the exterior region, whereas points outside Γ_2^s correspond to nontransit orbits (the manifolds act as separatrices for the states of motion [5, 10]).

Candidate trajectories for Earth–Moon transfers are nontransit orbits close to both $W^s(\gamma_2)$ and $W^u(\gamma_2)$. This property is wanted, since the existence of $W^s(\gamma_2)$ and $W^u(\gamma_2)$ has to be exploited, although the transfer orbit does not exactly lie on any invariant subset. Let $\tilde{\Gamma}_2^s$ be the set of points in the (r_2, \dot{r}_2) -plane that are enclosed

by $\partial \Gamma_2^s$, and $\bar{\Gamma}_2^s$ the closed set made up of $\partial \Gamma_2^s \cup \bar{\Gamma}_2^s$ (see Fig. 1b). Points on $\bar{\Gamma}_2^s$ have to be avoided as they lead to either transit or asymptotic orbits. On the contrary, all the points that lie on

$$l = \{(r_2, \dot{r}_2) \in S_A, (r_2, \dot{r}_2) \notin \bar{\Gamma}_2^s | r_2 = R_E + h_E\} \quad (6)$$

are translunar candidate orbits as they intersect the initial parking orbit (R_E is the radius of the Earth). This intersection occurs in the configuration space only, as the initial parking orbit and the translunar trajectory have two different energy levels.

The pair $\{C_{SE}, \varphi_A\}$ uniquely defines the curve $\partial \gamma_2^s$ on S_A : C_{SE} stands for the orbit γ_2 , whereas φ_A defines the surface of section S_A to cut the first intersection of $W^s(\gamma_2)$. Thus, $\{C_{SE}, \varphi_A\}$ are used to define the first guess Earth escape stage. In order to obtain efficient transfer trajectories, the lowest possible initial instantaneous maneuver, Δv_E , is searched. It is necessary to define its components: a first contribution to the Δv_E amount is related to the radial term Δv_r , while the second tangential contribution Δv_t is needed to fill the gap ΔC between the energy of the initial parking orbit, C_E , and C_{SE} (i.e. $\Delta C = C_E - C_{SE}$). It is possible to show that $\Delta v(\Delta C, \varphi_A) = \Delta v_t(\Delta C) + \Delta v_r(\varphi_A)$, and it is even possible to lower Δv_r to zero by properly tuning φ_A . This approach leads to initial tangential maneuvers, i.e. the initial Δv_E is aligned with the velocity of the circular parking orbit around the Earth. The search is therefore restricted to the point $P \in S_A$ defined by $P = l \cap l'$, where l' is the set of points having zero radial velocity with respect to the Earth

$$l' = \{(r_2, \dot{r}_2) \in S_A, (r_2, \dot{r}_2) \notin \bar{\Gamma}_2^s | \dot{r}_2 = 0\}. \quad (7)$$

Point P does not exactly lie on the stable manifold (but outside), and can be found sufficiently close to $W^s(\gamma_2)$ by suitably tuning φ_A (see Fig. 2b). In practice, since at this stage a first guess solution is constructed to be later optimized, orbits sufficiently close to P can also be considered. In particular, points $P' \in S_A$ such that $\|P' - P\| \leq \varepsilon$ are also taken into account, where ε is a certain prescribed distance.

A number of P' points can be generated by tuning the angle φ_A . These points, flown forward, generate orbits that are close to $W^s(\gamma_2)$ until the region about γ_2 is reached. From this point on, the orbits get close to $W^u(\gamma_2)$, and their intersection with S_B is studied. The set labeled \mathcal{E}_{SE} , $\mathcal{E}_{SE} \in S_B$, stands for the set of orbits close to $W^u(\gamma_2)$ whose preimage \mathcal{E}_{SE}^{-1} , $\mathcal{E}_{SE}^{-1} \in S_A$, is made up by P' points. Earth escape trajectories defined on \mathcal{E}_{SE}^{-1} , \mathcal{E}_{SE} are taken into account, as the latter intersects a special subset leading to orbits at the Moon (see Fig. 4b, where \mathcal{E}_{SE} is reported).

It is worth noticing that the parking orbit is defined at a lower energy level than orbits on \mathcal{E}_{SE}^{-1} , therefore the instantaneous velocity change Δv_E is required to place the translunar trajectory on $F(C_{SE}) \cap \mathcal{E}_{SE}^{-1}$. In practice, this maneuver is provided by the launch vehicle once the spacecraft is on the Earth parking orbit. A mission profile including this parking orbit is in fact consistent with major architecture requirements, as summarized in [15].

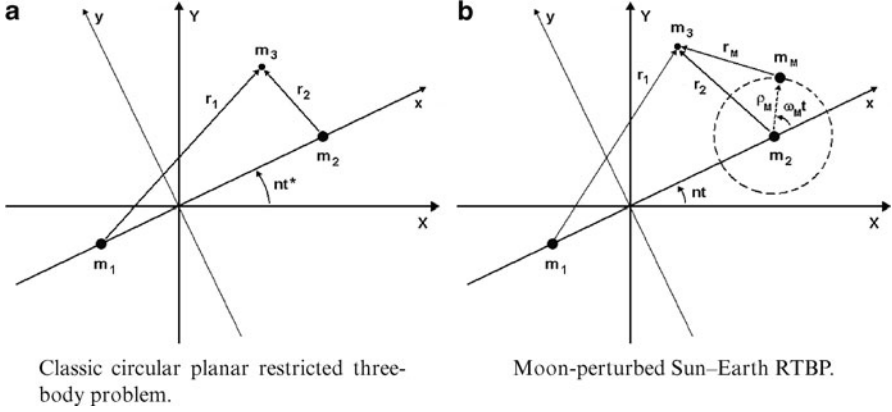


Fig. 3 Mathematical models to described the physics of the problem

The design of the SE phase is reduced in this way to the determination of just the two sets \mathcal{E}_{SE}^{-1} , \mathcal{E}_{SE} , and to the computation of Δv_E . This first phase of the transfer is constructed using (1). The SE trajectory does not make use of low-thrust propulsion, and the phase space structure of the ballistic PCRTBP is exploited.

3.1 The Moon-Perturbed Sun–Earth Restricted Three-Body Problem

When the gravitational attraction of the Moon is taken into account, (1) are augmented aiming at introducing the dynamics of the Moon in an autonomous fashion, leading to the formulation of the bicircular restricted four-body problem, BRFBP (see Fig. 3b). The dynamical system moves from the fourth order to the fifth one. Some assumptions are taken into account, recalling that the orbits of the primaries show low eccentricity values (≈ 0.01 , ≈ 0.04), and the Moon inclination with respect to the ecliptic plane is small ($\approx 5^\circ$): (a) the Sun and the Earth are revolving in circular orbits around their center of mass; (b) the Earth–Moon barycenter is moving in a circular orbit around the center of mass of the Sun–Earth–Moon system.

Assuming all the hypotheses written above, the planar equations of motion are:

$$\ddot{x} - 2\dot{y} = \frac{\partial \Omega_M}{\partial x}, \quad \ddot{y} + 2\dot{x} = \frac{\partial \Omega_M}{\partial y}, \quad \dot{\theta} = \omega_M \quad (8)$$

where the subscripts denote the partial derivative of the auxiliary function

$$\Omega_M(x, y, \theta) = \Omega(x, y, \mu_{SE}) + \frac{m_M}{r_M} - \frac{m_M}{\rho_M^2} (x \cos \theta + y \sin \theta). \quad (9)$$

The quantity $\Omega(x, y, \mu_{SE})$ stands for the classic CRTBP potential expressed by (2), while the remaining part represents the gravitational perturbation of the Moon.

The dimensionless physical constants introduced to describe the Moon influence are in agreement with those of the SE model. Thus, the distance between the Moon and the Earth is $\rho_M = 2.5721 \times 10^{-3}$, the mass of the Moon is $m_M = 3.6942 \times 10^{-8}$, and its angular velocity with respect to the SE rotating frame is $\omega_M = 1.2367 \times 10^1$. The location of the Moon is therefore at $(1 - \mu_{SE} + \rho_M \cos \theta, \rho_M \sin \theta)$, such that:

$$r_M^2 = (x - 1 + \mu_{SE} - \rho_M \cos \theta)^2 + (y - \rho_M \sin \theta)^2. \quad (10)$$

According to the differential equation (8), the system does not admit the existence of any libration point or integral of motion. Anyway, as the Moon can be considered as a small perturbation of the Sun–Earth model, a qualitative global analysis about the motion of the spacecraft is proposed, assuming the restricted four-body model as a perturbation of the invariant objects of the classic RTBP. If the points belonging to the escape set \mathcal{E}_{SE} are backwards integrated under the dynamics associated with (8), the topology of their trajectories in the configuration space is only slightly and negligible different. The main variations appear associated with the preimage set \mathcal{E}_{SE}^{-1} . If the trajectories pass near the Moon, the points $P' \in S_A$ show almost the same phase–space coordinates (r_2, \dot{r}_2) as before, while the tangential velocity reduces significantly, with respect to the classic Sun–Earth PCRTBP computation. In details, when a lunar flyby is explicitly taken into account while designing the Earth escape stage, the difference between the energy level of the parking orbit at a height $h_E = 167$ km and the one of the orbits on \mathcal{E}_{SE}^{-1} is lowered. This means that a reduced instantaneous velocity change Δv_E is now required to place the translunar trajectory on $F(C_{SE}) \cap \mathcal{E}_{SE}^{-1}$.

4 Low-Thrust Propulsion and Attainable Sets

Once the initial transfer stage is defined, the final one that leads to the Moon is required. A general approach useful to mathematically describe impulsive maneuvers, low-thrust arcs. The core of the formulation is based on a perturbed version of the classic PCRTBP. The perturbation is the low-thrust propulsion, and when its driving parameters are wisely tuned, different types of transfer arcs as well as several final conditions at the Moon can be investigated.

To model the *controlled* motion of m_3 under both the gravitational attractions of m_1 , m_2 , and the low-thrust propulsion, the following differential equations are considered:

$$\ddot{x} - 2\dot{y} = \frac{\partial \Omega}{\partial x} + \frac{T_x}{m}, \quad \ddot{y} + 2\dot{x} = \frac{\partial \Omega}{\partial y} + \frac{T_y}{m}, \quad \dot{m} = -\frac{T}{I_{sp} g_0}, \quad (11)$$

where $T = \sqrt{T_x^2 + T_y^2}$ is the thrust magnitude, I_{sp} the specific impulse of the engine and g_0 the gravitational acceleration at sea level. The ballistic motion (1) is represented by a fourth-order system, while the controlled motion (11) is described by a fifth-order system of differential equations. Continuous variations of the spacecraft mass, m , are taken into account when low-thrust propulsion is considered. This causes a singularity arising when $m \rightarrow 0$, beside the well-known singularities given by impacts of m_3 with m_1 or m_2 .

The thrust law $\mathbf{T}(t) = \{T_x(t), T_y(t)\}^\top$, $t \in [t_i, t_f]$, in (11) is not given, but rather in this approach it represents an unknown that is found when the optimal control problem is solved (t_i and t_f are the initial and final times, respectively). It is determined such that a certain state is targeted and, at the same time, a certain objective function is minimized. However, in order to build first guess solutions, at this stage the profile of \mathbf{T} over time is described. In particular, using tangential thrust, attainable sets can be defined in the same fashion as reachable sets are defined in [6].

Let \mathbf{y}_i be a vector representing a generic initial state, i.e. $\mathbf{y}_i = \{x_i, y_i, \dot{x}_i, \dot{y}_i, m_i\}^\top$, and let $\phi_{\mathbf{T}(\tau)}(\mathbf{y}_i, t_i; t)$ be the flow of system of (11) at time t , starting from (\mathbf{y}_i, t_i) and considering the thrust profile $\mathbf{T}(\tau)$, $\tau \in [t_i, t]$. The latter has to be taken within proper bounds that are typically given by technological constraints. This condition usually reads $T(t) \leq T_{\max}$, where T_{\max} is the maximum available thrust magnitude. With this notation, it is possible to define the generic point of a tangential low-thrust trajectory through

$$\mathbf{y}(t) = \phi_{\bar{\mathbf{T}}}(\mathbf{y}_i, t_i; t), \quad (12)$$

where $\bar{\mathbf{T}} = \bar{T}(\mathbf{v}/v)$, $v = \sqrt{\dot{x}^2 + \dot{y}^2}$, $\mathbf{v} = \{\dot{x}, \dot{y}\}^\top$. Equation (12) represents the flow of the differential system governed by (11), when constant tangential thrust of magnitude \bar{T} is considered. With given \bar{T} , tangential thrust maximizes the variation of Jacobi energy, which is the only property that has to be dealt with when designing trajectories in the RTBP. (In [11], a comparison between tangential thrust in either rotating or inertial frame is proposed). The low-thrust orbit, at time t , can be expressed as

$$\gamma_{\bar{\mathbf{T}}}(\mathbf{y}_i, t) = \{\phi_{\bar{\mathbf{T}}}(\mathbf{y}_i, t_i; \tau) | \tau < t\}, \quad (13)$$

where the dependence on the initial state \mathbf{y}_i is kept. The attainable set, at time t , can be defined as

$$\mathcal{A}_{\bar{\mathbf{T}}}(t) = \bigcup_{\mathbf{y}_i \in \mathcal{Y}} \gamma_{\bar{\mathbf{T}}}(\mathbf{y}_i, t), \quad (14)$$

where \mathcal{Y} is a domain of admissible initial conditions. Attainable set in (14) is associated with a generic \mathcal{Y} ; this set can be suitably defined for the three different types of transfers at hand. Thanks to the definition of $\mathcal{A}_{\bar{\mathbf{T}}}(t)$, low-thrust propulsion can be incorporated in a three-body frame, using the same methodology developed for the invariant manifolds [9]. More specifically, invariant manifolds are replaced by attainable sets which are manipulated (i.e. intersected) to find a transfer point on a suitable surface of section. The idea is to mimic the role that invariant

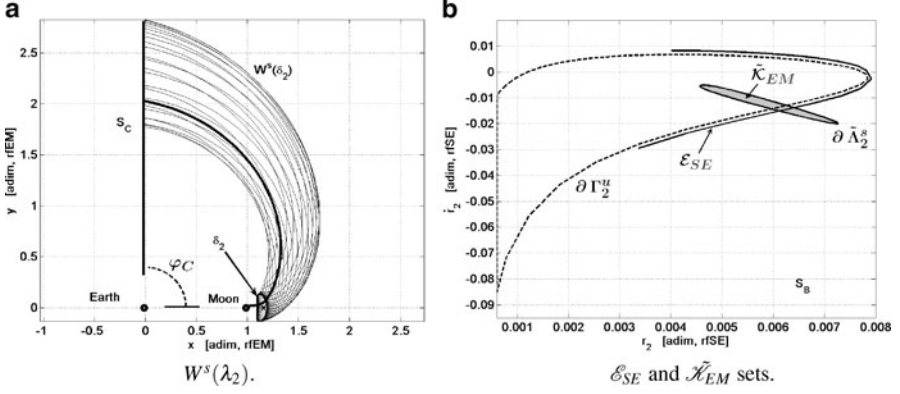


Fig. 4 Stable manifold $W^s(\lambda_2)$ and its section curve $\partial \tilde{\Lambda}_2^s$. The latter is used to define the set of orbits that lead to Moon capture \mathcal{K}_{EM} . In (a), the bold line stands for a sample Moon capture trajectory

manifolds have in trajectory design with the use of attainable sets. This approach can be adapted to design either two-impulse transfers as well as low-thrust transfer to the Moon.

4.1 Moon Ballistic Capture Stage

In analogy with what was described for the Earth escape stage in Sect. 3, by fixing a suitable value of the Jacobi constant in the EM model, C_{EM} , a unique Lyapunov orbit about both $L1$ and $L2$, named λ_1 and λ_2 , respectively, can be defined. Restricting the energy to $C_{EM} \ll C_2$, both λ_1 and λ_2 exist, and the Hill's regions are opened at both $L1$ and $L2$. In order to reach the final orbit about the Moon from the exterior, a capture via $L2$ is considered (see Fig. 4a). This means that the Moon ballistic capture stage is constructed by exploiting the dynamics around $L2$. The stable manifold associated with λ_2 , $W^s(\lambda_2)$, is computed starting from λ_2 and integrating backward until a certain surface of section is reached. Section S_C , making an angle φ_C (counterclockwise) with the x -axis and passing through the Earth, is considered to cut $W^s(\lambda_2)$ ($\varphi_C = \pi/2$ in Fig. 4a). The corresponding section curve, $\partial \Lambda_2^s$ (computed on the (r_1, \dot{r}_1) -plane in the EM model), is diffeomorphic to a circle. The set $\mathcal{K}_{EM} = \tilde{\Lambda}_2^s$ is defined, where $\tilde{\Lambda}_2^s \in S_C$ is the set of points inside $\partial \Lambda_2^s$, set that leads to the Moon capture.

The set \mathcal{K}_{EM} is defined on section S_C in the EM model. However, it is possible to represent \mathcal{K}_{EM} on S_B defined in the SE model through the transformation $\tilde{\mathcal{K}}_{EM} = \mathcal{M}(\mathcal{K}_{EM})$. The operator \mathcal{M} maps states on S_C (EM model) to states on S_B (SE model), provided the two angles φ_C , φ_B . The same conversion is also applied to $\partial \Lambda_2^s$, in order to obtain $\partial \tilde{\Lambda}_2^s = \mathcal{M}(\partial \Lambda_2^s)$ on section S_B from section S_C .

This transformation is basically made up of a rotation and a scaling of the variables in proper units (see Fig. 4b where \mathcal{K}_{EM} and $\partial \tilde{\Lambda}_2^s$ are reported).

Considering only section S_B , it is possible to define the ballistic low-energy Earth–Moon transfers as the orbits belonging to the set $\mathcal{E}_{\text{SE}} \cap \mathcal{K}_{\text{EM}}$. The sets \mathcal{E}_{SE} and \mathcal{K}_{EM} are characterized by different values of the Jacobi constant, C_{SE} and C_{EM} , respectively. In addition, any point in \mathcal{K}_{EM} has a different value of Jacobi constant as \mathcal{M} is not energy preserving. Thus, in order to join together orbits on \mathcal{E}_{SE} and orbits on \mathcal{K}_{EM} , an impulsive maneuver at the patching point may be required.

This approach is the one known in literature to design Earth–Moon low-energy transfers with impulsive maneuvers [9]. Indeed, another impulsive maneuver, beside the one performed at the patching point, is necessary upon Moon arrival to place the spacecraft into a stable orbit around the Moon. This approach is based on either (1) or (11) with $T = 0$, and allows to define low energy transfers like those described in [2, 3, 8].

Two-impulse transfers to the Moon are defined as follows. The spacecraft is assumed to be initially on a parking orbit about the Earth with given eccentricity, e , and perigee/apogee altitude, h_p/h_a . The argument of perigee, ω_E , of this orbit is not fixed. The transfer begins when the spacecraft is at the perigee of this orbit. An impulsive maneuver, carried out by the launch vehicle, places the spacecraft onto a translunar trajectory; from this point on, the spacecraft flies ballistically under the RTBP dynamics until it reaches the Moon neighborhood. Then, a second impulsive maneuver is required to insert the spacecraft into a stable orbit around the Moon. This orbit has moderate eccentricity, e , and periapsis/apoapsis altitude h_p/h_a , prescribed by the mission requirements. The transfer terminates when the spacecraft is at the periapsis of this orbit. While both e and r_p/r_a are given, the orientation, i.e. the argument of periapsis, ω_M , of the final orbit around the Moon is not fixed.

The global transfer is designed recalling the coupled RTBPs approximation. The initial orbit is a trajectory belonging to the set \mathcal{E}_{SE} , whereas the second part is defined using a suitable attainable set in the EM model, according to the formalism introduced in Sect. 4 through (14). As this set constitutes the second part of the trajectory, it is made up of ballistic orbits that are integrated backward, i.e. considering the thrust magnitude $T = 0$. More specifically, as both eccentricity and apsidal altitude are prescribed, the final state of the transfers (i.e., the periapsis point of the orbit about the Moon) is a function of the argument of periapsis and of the final tangential Δv_M impulsive maneuver required to place the spacecraft into a stable lunar orbit, i.e. $\mathbf{y}_f = \mathbf{y}_f(\omega_M, \Delta v_M)$, see Fig. 5a.

According to this approach, the domain of admissible final states becomes

$$\mathcal{Y}^M = \{\mathbf{y}_f(\omega_M, \Delta v_M) | \omega_M \in [0, 2\pi], \Delta v_M \in [0, +\infty]\}, \quad (15)$$

and the attainable set, for some $t \geq 0$ (i.e. $-t$ is a backward integration), containing ballistic capture trajectories with impulsive stabilization is

$$\mathcal{A}_{\mathbf{B}}^M(-t) = \bigcup_{\mathbf{y}_f \in \mathcal{Y}^M} \gamma_{\mathbf{B}}(\mathbf{y}_f(\omega_M, \Delta v_M), -t). \quad (16)$$

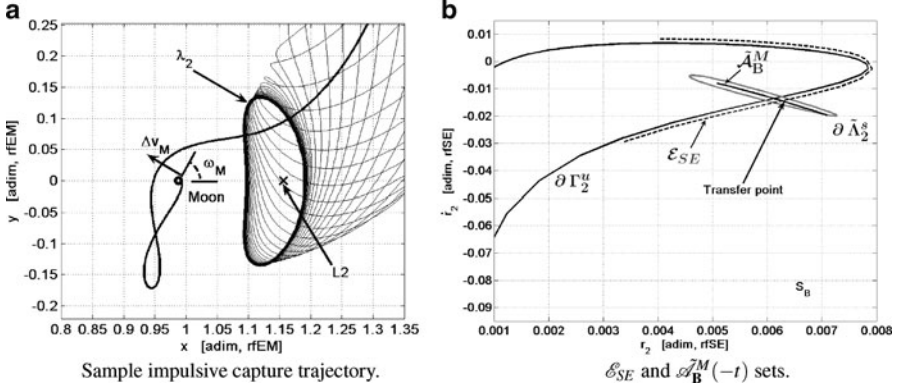


Fig. 5 The first guess impulsive capture solution as the transfer point $\mathcal{B}_{-t}^M = \mathcal{E}_{SE} \cap \mathcal{A}_{\mathbf{B}}^M(-t)$, the latter reported on section S_B in (a)

Each generic Moon capture orbit written in (16), at time $-t$, can be expressed as

$$\gamma_{\mathbf{B}}(\mathbf{y}_i, -t) = \{\phi_{\mathbf{B}}(\mathbf{y}_i, t_i; -\tau) \mid -\tau > -t\}, \quad (17)$$

where $\mathbf{B} = T(\mathbf{v}/v)$, $v = \sqrt{\dot{x}^2 + \dot{y}^2}$, $\mathbf{v} = \{\dot{x}, \dot{y}\}^\top$, and assuming $T = 0$. Since the first part of the transfer is defined on \mathcal{E}_{SE} , the transfer points, if any, that generate two-impulse transfers are contained in the set

$$\mathcal{B}_{-t}^M = \mathcal{E}_{SE} \cap \mathcal{A}_{\mathbf{B}}^M(-t). \quad (18)$$

Once again, the transformation \mathcal{M} is required to map $\mathcal{A}_{\mathbf{B}}^M(-t)$, computed in the EM model, into $\mathcal{A}_{\mathbf{B}}^M(-t)$, defined in the SE model, see Fig. 5b. At this stage just first guesses are defined, states with small, tolerable mismatches can be admitted in \mathcal{B}_{-t}^M as the discontinuities are spread in the subsequent optimization step.

4.2 Moon Low-Thrust Capture Stage

Dealing with the design of low-energy, low-thrust transfers to the Moon, the initial and final conditions at the Earth and the Moon, respectively, are assumed in the same way as described in Sect. 4.1 for the two-impulse trajectories. Actually, this case differs from the previous one in the fact that low-thrust contribution is taken into account. As part of the coupled RTBPs approximation, in connection with the Earth escape stage, the continuously propelled capture stage is used with the low-thrust term preliminary acting only in the EM model. Thus, the initial orbit is a trajectory belonging to the set \mathcal{E}_{SE} , whereas the second part is defined using

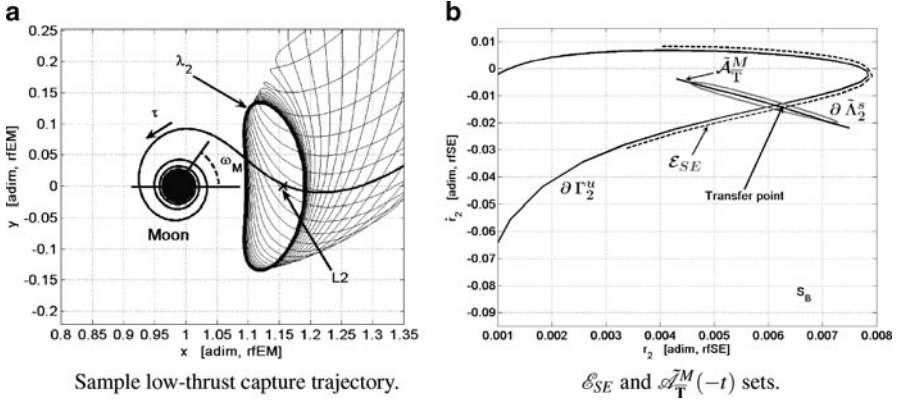


Fig. 6 The first guess low-thrust capture solution as the transfer point $\mathcal{T}_{-t}^M = \mathcal{E}_{SE} \cap \mathcal{A}_T^M(-t)$, the latter reported on section S_B in (b)

a suitable attainable set. As this set constitutes the second part of the trajectory, it is made up of tangential low-thrust orbits that are integrated backward. More specifically, as both eccentricity and apsidal altitude are prescribed, the final state of the transfers (i.e. the periapsis point of the orbit about the Moon) is a function of the argument of periapsis, i.e. $\mathbf{y}_f = \mathbf{y}_f(\omega_M)$, as shown in Fig. 6a. The domain of admissible final states therefore is

$$\mathcal{Y}^M = \{\mathbf{y}_f(\omega_M) | \omega_M \in [0, 2\pi]\}, \quad (19)$$

and the attainable set, for some $t \geq 0$ (i.e. $-t$ is a backward integration), containing low-thrust, ballistic capture trajectories is

$$\mathcal{A}_T^M(-t) = \bigcup_{\mathbf{y}_f \in \mathcal{Y}^M} \gamma_T(\mathbf{y}_f(\omega_M), -t). \quad (20)$$

Since the first part of the transfer is defined on \mathcal{E}_{SE} , the transfer points, if any, that generate low-energy, low-thrust transfers are contained in the set

$$\mathcal{T}_{-t}^M = \mathcal{E}_{SE} \cap \mathcal{A}_T^M(-t). \quad (21)$$

The transformation \mathcal{M} is required to map $\mathcal{A}_T^M(-t)$, computed in the EM model, into $\tilde{\mathcal{A}}_T^M(-t)$, defined in the SE model, as shown in Fig. 6b. It is worth mentioning that first guess solutions are being generated in this step. These preliminary solutions have to be later optimized in a four-body context. Thus, small discontinuities can be tolerated when looking for the transfer point. This means that it is possible to intersect two states such that $\|\mathbf{y}_A - \mathbf{y}_E\| \leq \varepsilon$, where $\mathbf{y}_E \in \mathcal{E}_{SE}$, $\mathbf{y}_A \in \tilde{\mathcal{A}}_T^M(-t)$, and ε is a prescribed tolerance. The greater ε is, the higher number of first guess solutions is found; however, ε should be kept sufficiently small to permit the convergence of the subsequent optimization step.

5 Trajectory Optimization

Once feasible and efficient first guess solutions are achieved, combining attainable sets with Earth-escape sets, an optimal control problem is stated in the BRFBP framework. The model used to take into account low-thrust propulsion and the gravitational attractions of all the celestial bodies involved in the design process (i.e. the Sun, the Earth, and the Moon) is

$$\ddot{x} - 2\dot{y} = \frac{\partial \Omega_S}{\partial x} + \frac{T_x}{m}, \quad \ddot{y} + 2\dot{x} = \frac{\partial \Omega_S}{\partial y} + \frac{T_y}{m}, \quad \dot{\theta} = \omega_S, \quad \dot{m} = -\frac{T}{I_{sp} g_0}. \quad (22)$$

This is a modified version of the classic bicircular four-body problem [17] (see Fig. 7a) and, in principle, incorporates the perturbation of the Sun into the Earth–Moon PCRTBP described by (1). The four-body potential Ω_S reads

$$\Omega_S(x, y, \theta) = \Omega(x, y, \mu_{EM}) + \frac{m_S}{r_S} - \frac{m_S}{\rho_S^2} (x \cos \theta + y \sin \theta). \quad (23)$$

The dimensionless physical constants introduced to describe the Sun perturbation are in agreement with those of the EM model. Thus, the distance between the Sun and the Earth–Moon barycenter is $\rho_S = 3.8878 \times 10^2$, the mass of the Sun is $m_S = 3.2890 \times 10^5$, and its angular velocity with respect to the EM rotating frame is $\omega_S = -9.2518 \times 10^{-1}$. The Sun is located at $(\rho_S \cos \theta, \rho_S \sin \theta)$, and therefore the Sun-spacecraft distance is calculated as

$$r_S^2 = (x - \rho_S \cos \theta)^2 + (y - \rho_S \sin \theta)^2. \quad (24)$$

This low-thrust version of the BRFBP is represented by the sixth-order system of differential equation (22).

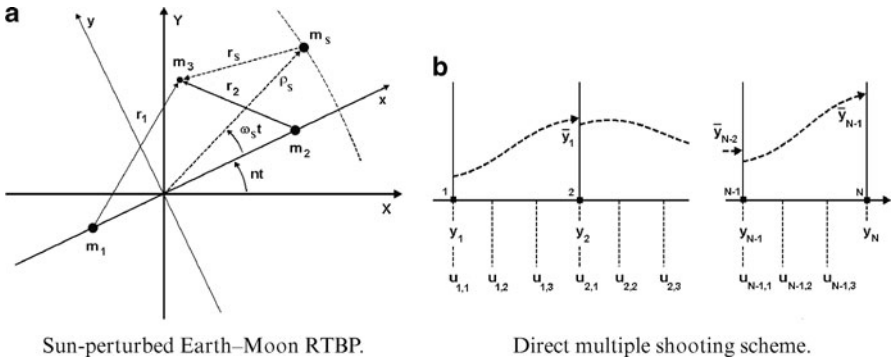


Fig. 7 Optimization process, dynamical model and integration scheme

According to the formalism proposed by Betts [4], the controlled BRFBP described by (22) is written in the first-order form

$$\begin{aligned}
 \dot{x} &= v_x \\
 \dot{y} &= v_y \\
 \dot{v}_x &= 2v_y + \Omega_{Sx} + T_x/m \\
 \dot{v}_y &= -2v_x + \Omega_{Sy} + T_y/m \\
 \dot{\theta} &= \omega_S \\
 \dot{m} &= -T/(I_{sp} g_0),
 \end{aligned} \tag{25}$$

with $v_x = \dot{x}$ and $v_y = \dot{y}$. In a compact explicit form, system (25) reads

$$\dot{\mathbf{y}} = \mathbf{f}[\mathbf{y}(t), \mathbf{T}(t), \mathbf{p}, t], \tag{26}$$

where \mathbf{f} stands for the vector field, $\mathbf{T} = \{T_x, T_y\}^\top$ is the thrust vector, and $\mathbf{y} = \{x, y, \dot{x}, \dot{y}, \theta, m\}^\top$ is the state vector. In addition, the dynamics is also allowed to incorporate some constant parameters \mathbf{p} useful for the definition of the optimal control problem. The aim is finding, according to the standard optimal control theory, the guidance law $\mathbf{T} = \mathbf{T}(t)$, $t \in [t_i, t_f]$, that minimizes a prescribed scalar performance index or objective function

$$J = J(\mathbf{y}, \mathbf{T}, \mathbf{p}, t), \tag{27}$$

while satisfying certain mission constraints. These constraints are represented by the two boundary conditions, defined at the end points of the optimal control problem, and by the inequality conditions, defined along the whole arc. These last quantities are derived specifically for each of the three types of mission to the Moon that are investigated.

The optimal control problem (OCP) is then transcribed into a nonlinear programming, NLP, problem using a direct approach. This method, although suboptimal, generally shows robustness and versatility, and does not require explicit derivation of the necessary conditions of optimality. Moreover, direct approaches offer higher computational efficiency and are less sensitive to variation of the first guess solutions [4]. Furthermore, a multiple shooting scheme is implemented. With this strategy the BRFBP dynamics presented by (25) is forward integrated within $N - 1$ intervals (in which $[t_i, t_f]$ is uniformly split), i.e. the time domain is divided in the form $t_i = t_1 < \dots < t_N = t_f$, and the solution is discretized over the N grid nodes (see Fig. 7b). The continuity of position, velocity, and mass is imposed at their ends [7], in the form of defects $\eta_j = \bar{\mathbf{v}}_j - \mathbf{v}_{j+1} = 0$, for $j = 1, \dots, N - 1$. The quantity $\bar{\mathbf{v}}_j$ stands for the result of the integration, i.e. $\bar{\mathbf{v}}_j = \phi(\mathbf{v}_j, \mathbf{p}, t)$, $t_j \leq t_{j+1}$, and is made up of state variables and control variables (i.e. $\mathbf{v}_j = \{\mathbf{y}_j, \mathbf{T}_{j,k}\}^\top$, for $k = 1, \dots, M - 1$). The control law $\mathbf{T}(t)$ is described within each interval by means of cubic spline functions. The algorithm computes the value of the control at mesh points, satisfying both boundary and path constraints, and minimizing the performance index.

Dynamics described by (22) are highly nonlinear and, in general, lead to chaotic orbits. In order to find accurate optimal solutions without excessively increasing the computational burden, an adaptive nonuniform time grid has been implemented. Thus, when the trajectory is close to either the Earth or the Moon the grid is automatically refined, whereas in the intermediate phase, where a weak vector field governs the motion of the spacecraft, a coarse grid is used. The optimal solution found is assessed a posteriori by forward integrating the optimal initial condition (with RK scheme of the eighth order) and by cubic interpolation of the discrete optimal control solution. Not only the low-thrust portion, but rather the whole transfer trajectory is discretized and optimized, so allowing the low-thrust to act also in regions preliminary made up by coast arcs.

5.1 Two-Impulse Problem Statement

As far as it concerns low energy transfers to the Moon, designed by means of two impulsive maneuvers at both ends of the trajectory, the dynamical system described by (25) is characterized by the thrust magnitude always forced to be equal to zero, in order to describe the coast arcs which compose the whole Earth–Moon transfer.

According to the NLP problem recalled in Sect. 5, the variable vector is

$$\mathbf{x} = \{(\mathbf{y}, \mathbf{B})_1, \dots, (\mathbf{y}, \mathbf{B})_N, t_1, t_N\}^\top, \quad (28)$$

where $\mathbf{B} = \{0, 0\}^\top$ is forced, i.e. the thrust magnitude is $T = 0$ along the whole time domain of the transfer.

The initial conditions read:

$$\psi_i(\mathbf{y}_1, t_1) := \begin{cases} (x_1 + \mu)^2 + y_1^2 - r_i^2 = 0 \\ (x_1 + \mu)(\dot{x}_1 - y_1) + y_1(\dot{y}_1 + x_1 + \mu) = 0, \end{cases} \quad (29)$$

which force the first \mathbf{y}_1 state of the transfer to belong to a circular orbit of radius $r_i = R_E + h_E$, where R_E and h_E stand for the Earth radius and the orbit altitude with respect to the Earth, respectively. In analogy, the final conditions at the Moon are:

$$\psi_f(\mathbf{y}_N, t_N) := \begin{cases} (x_N - 1 + \mu)^2 + y_N^2 - r_f^2 = 0 \\ (x_N - 1 + \mu)(\dot{x}_N - y_N) + y_N(\dot{y}_N + x_N - 1 + \mu) = 0, \end{cases} \quad (30)$$

according to which, the final \mathbf{y}_N state of the transfer is on a circular orbit of radius $r_f = R_M + h_M$, where R_M and h_M stand for the Earth radius and the orbit altitude with respect to the Moon, respectively. Conditions written in (29) and (30) mean that in the reference frame centered in the Earth and the Moon, respectively, the position and velocity vectors are orthogonal [21]. The nonlinear equality constraint

vector, made up of the boundary conditions and the ones representing the dynamics, is therefore written as follows:

$$\mathbf{c}(\mathbf{x}) = \{\psi_i, \eta_1, \dots, \eta_{N-1}, \psi_f\}^\top. \quad (31)$$

Moreover, aiming at avoiding the collision with the two primaries, the following inequality constraints are imposed:

$$\Psi_j^c(\mathbf{y}_j) := \begin{cases} R_E^2 - (x_j + \mu)^2 - y_j^2 \leq 0 \\ R_M^2 - (x_j - 1 + \mu)^2 - y_j^2 \leq 0, \end{cases} \quad j = 2, \dots, N-1. \quad (32)$$

Finally, the flight time is searched to be positive, i.e.

$$\Psi^t = t_1 - t_N \leq 0. \quad (33)$$

The complete inequality constraint vector therefore reads:

$$\mathbf{g}(\mathbf{x}) = \{\Psi_2^c, \dots, \Psi_{N-1}^c, \Psi^t\}^\top. \quad (34)$$

As for the performance index to minimize, this is a scalar that represents the two velocity variations at the beginning and at the arrival of the transfer, i.e. $J(\mathbf{x}) = \Delta v_1 + \Delta v_N$, where

$$\Delta v_1 = \sqrt{(\dot{x}_1 - y_1)^2 + (\dot{y}_1 + x_1 + \mu)^2} - v_i, \quad (35)$$

and where

$$\Delta v_N = \sqrt{(\dot{x}_N - y_N)^2 + (\dot{y}_N + x_N - 1 + \mu)^2} - v_f. \quad (36)$$

In summary, the NLP problem for the two-impulse transfers is formulated as follows:

$$\begin{aligned} \min_{\mathbf{x}} J(\mathbf{x}) \quad & \text{subject to } \mathbf{c}(\mathbf{x}) = 0, \\ & \mathbf{g}(\mathbf{x}) \leq 0. \end{aligned} \quad (37)$$

5.2 Low-Thrust Problem Statement

Dealing with low-energy low-thrust transfers to the Moon, their design is performed by means of an initial impulsive TLI maneuver, followed by a ballistic arc closed by a low-thrust capture around the Moon.

According to the NLP problem recalled in Sect. 5, the variable vector is

$$\mathbf{x} = \{(\mathbf{y}, \mathbf{T})_1, \dots, (\mathbf{y}, \mathbf{T})_N, t_1, t_N\}^\top, \quad (38)$$

where, in this case, the thrust vector is properly $\mathbf{T} = (T_x, T_y)^\top$. Actually, even if the first guess control law is assigned aligned with the synodic velocity of the spacecraft (tangential thrust of magnitude \bar{T}), the optimization process acts on all the variable of (38), and it is free to modify the control history, within the technological limits of the engine.

Then, the equality constraint vector, made up of initial and final conditions, together with the dynamics defect η satisfaction, is proposed in the same way as (31) of Sect. 5.1. Dealing with the inequality constraints, beside the ones preventing the collision of the spacecraft with the two primaries and the one that represents the positive time evolution of the trajectory, a path constraint is added. This is introduced to model the saturation of the low-thrust engine. Thus, the inequality $T(t) \leq T_{\max} = 0.5 \text{ N}$, $t \in [t_i, t_f]$, is imposed along the whole transfer.

The optimal control step aims at finding the guidance law, $\mathbf{T}(t)$, $t \in [t_i, t_f]$, that minimizes the hybrid performance index

$$J = \rho \Delta v_1 + \int_{t_i}^{t_f} \frac{T(t)}{I_{\text{sp}} g_0} dt; \quad (39)$$

the first part stands for the initial impulsive translunar insertion maneuver (provided by the launcher and recalling (35)), while the second contribution corresponds to the propellant mass, $m_p = m_i - m(t_f)$, needed to perform the transfer. The tuning parameter ρ is a weight quantity introduced to wisely balance the two contributions in the objective index computing. It is worth noticing that the optimization process is performed on the NLP variables described by (38), which are already dimensionless and easy to deal with as they show the same order of magnitude. In order to eliminate possible discontinuities, arising when the intersection of the orbits is not exact, each first guess is first preprocessed to minimize

$$J = \frac{1}{2} \int_{t_i}^{t_f} \mathbf{u}^\top \mathbf{u} dt, \quad (40)$$

with $\mathbf{u}(t) = \mathbf{T}(t)/m(t)$. The resulting solution is later used to minimize the objective function described by (39). Numerical experiments have shown that performing an intermediate step with objective function of (40) better enforces the satisfaction of boundary conditions and path constraints [11].

Finally, the NLP problem for the low-energy low-thrust transfers is formulated as proposed by (37) at the end of Sect. 5.1.

6 Optimized Transfer Solutions

In this section the transfer solutions arising from the optimization process are presented. In Sect. 1, two families of trajectories are discussed, according to different types of propulsion system. In the following, the optimized transfers to LLOs are proposed in terms of some relevant performance parameters.

Table 1 Two-impulse transfers and low-energy low-thrust transfers to LLOs. A set of impulsive reference solutions found in literature is also reported

Type	Δv_i [m/s]	Δv_f [m/s]	f_f [adim.]	f_t [adim.]	Δt [days]
Sol.1	3211	–	0.061	0.683	271
Sol.2	3203	–	0.050	0.681	145
Sol.3	3169	–	0.046	0.675	103
Sol.4	3143	650	0.198	0.724	88
Yag	3137	718	0.216	0.730	44
WSB	3161	677	0.206	0.729	90–120
BP	3232	721	0.217	0.739	∞
BE	3161	987	0.284	0.756	55–90
L1	3265	629	0.192	0.734	255
Hoh	3143	848	0.250	0.742	5
Min	3099	622	0.191	0.718	–

Yag two-impulse transfers [21], *WSB* weak stability boundary, *BP* bi-parabolic, *Hoh* Hohmann, *BE* bi-elliptic [3], *L1* via *L1* transit orbits [20], *Min* theoretical minimum [18]

6.1 Trajectories to Low Orbits Around the Moon

Optimal two-impulse and low-energy low-thrust solutions are presented. These transfers start from a circular parking orbit at an altitude of $h_E = 167$ km around the Earth, and end at circular orbit around the Moon, at an altitude of $h_M = 100$ km. The results are shown in Table 1 as follows: the first three solutions (i.e. sol.1, sol.2, sol.3) correspond to the low-energy low-thrust transfers, while solution sol.4 represents a two-impulse low-energy transfer. Then, solutions below the horizontal line are some reference impulsive transfers found in literature [3, 18, 20, 21]; all of them begin from the same LEO and arrive at the same LLO.

Table 1 is organized as follows: the second column Δv_i stands for the initial impulsive maneuver that inserts the spacecraft onto the translunar trajectory. For the solutions computed in this thesis, they are a direct output of the optimization process, described in Sect. 5. The third column Δv_f represents the final impulsive maneuver that permits a stable permanent capture into a circular parking orbit around the Moon. This comes out from the optimization step for sol.4; it is not present for the low-thrust transfers, whereas for the reference solutions this term takes into account all the impulsive maneuvers necessary to carry out the transfers except for Δv_i (i.e. for an Hohmann transfer, only the second burn necessary to place the spacecraft into the final orbit about the Moon is considered; for a WSB transfer, this term has to take into account the possible mid-course maneuver as well as the final maneuver needed to place the spacecraft into the final orbit about the Moon; similar arguments apply also to the bi-elliptic and bi-parabolic transfers).

Then, the fourth column f_f is a direct output (i.e. it is part of the performance index to minimize, divided by $m_{TLI} = 1,000$ kg) only for the first three solutions, and stands for the propellant mass fraction required to perform the transfers, after the

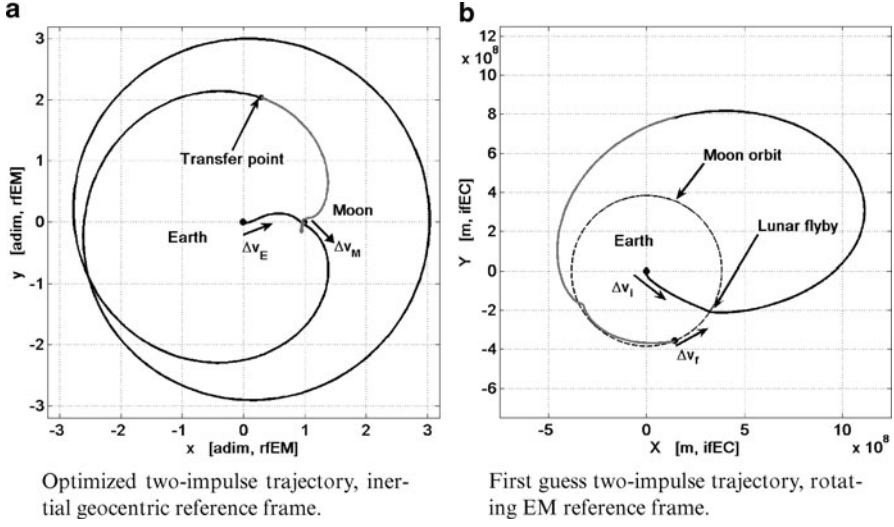


Fig. 8 Optimized two-impulse transfer to a low orbit around the Moon, corresponding to solution 4 in Table 1

translunar insertion. As for the others, the value written in Table 1 is the propellant mass fraction associated with the impulsive Δv_f maneuver. This, through the rocket equation reads

$$f_f = \frac{m_p}{m_{TLI}} = 1 - \exp\left(-\frac{\Delta v_f}{I_{sp}^{ht} g_0}\right), \quad (41)$$

where $I_{sp}^{ht} = 300$ s is assumed as the typical specific impulse related to high-thrust chemical engines. The fifth column f_i represents the overall mass fraction necessary to complete the Earth–Moon transfers. Even if for low-energy low-thrust solutions, according to the design of the Earth escape stage described in Sect. 3, the initial Δv_i is given by the launch vehicle. For a sake of a fair comparison, the cost of this maneuver is considered, as written below:

$$f_t = \frac{m_p}{m_i} = \left[1 - \exp\left(-\frac{\Delta v_i}{I_{sp}^{lt} g_0}\right)\right] + \frac{1}{m_i} \int_{t_i}^{t_f} \frac{T(t)}{I_{sp}^{lt} g_0} dt, \quad (42)$$

where $I_{sp}^{lt} = 3,000$ s is assumed as the specific impulse related to low-thrust electrical engines. Finally, the last column on the right stands for the transfer time. Dealing with the last line, the minimum theoretical cost is computed via energetic considerations, and no solutions corresponding to such solution exist.

First of all, few considerations about the first three low-energy low-thrust transfers: moving from sol.1 to sol.3, the performances of the solutions become better, both in terms of Δv_i and Δt . This is due because the first guess trajectory related to

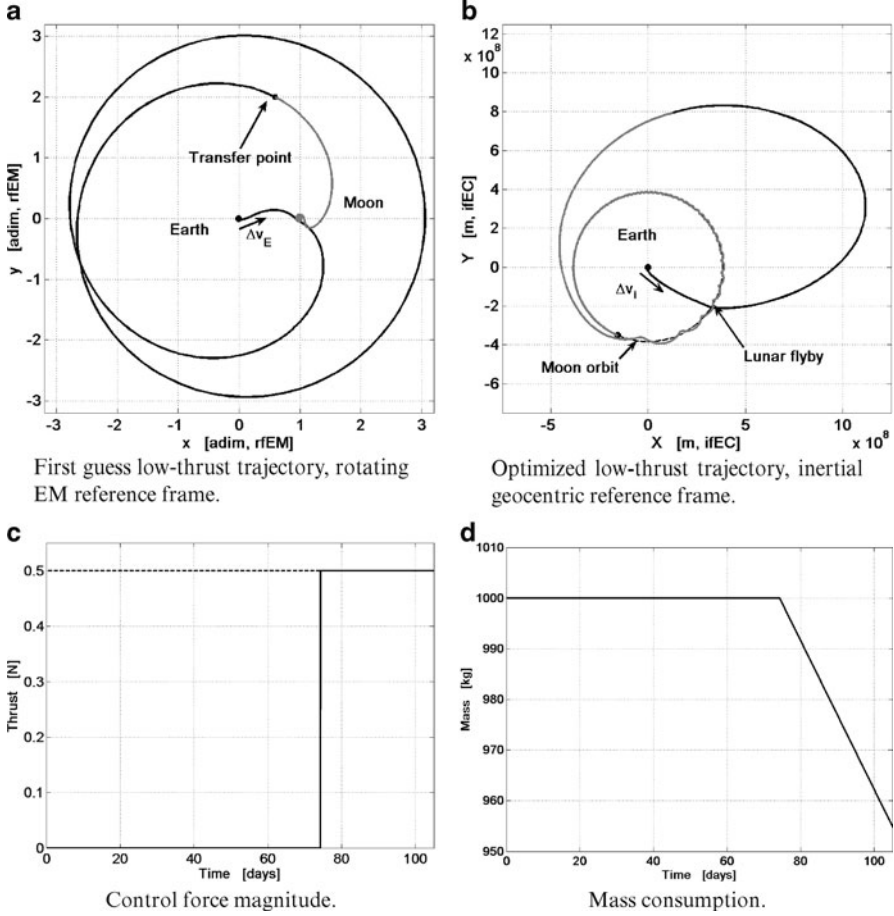


Fig. 9 Optimized low-energy low-thrust transfer to a low orbit around the Moon corresponding to solution 3 in Table 1

sol.3 (in particular the Earth escape stage) is designed taking advantage explicitly of a lunar flyby at the beginning of the transfer, thanks to the Moon-perturbed Sun–Earth restricted three-body model. Moreover, the plane where suitable Poincaré sections are considered and where the set $\mathcal{T}_{-t}^M = \mathcal{E}_{SE} \cap \mathcal{A}_T^M(-t)$ is defined, is chosen tuning shrewdly the φ_B and φ_C angles, in order to reduce the flight time.

In any case, sol.3, shown in Fig. 9, offers the lowest value of the overall mass consumption (see f_t). This happens for two reasons: first the fact that the low-thrust, I_{sp}^{lt} , is one order of magnitude greater than I_{sp}^{ht} . Second, the first guess solutions exploit deeply the dynamics of the RTBPs where they are designed, and later of the Earth–Moon BRFBP where they are optimized. Moreover, these trajectories take explicitly advantage of the initial lunar flyby. The latter can be seen as a kind of aid in the translunar orbit insertion, as it reduces the Δv_i required for that maneuver.

The two-impulse trajectory corresponding to sol.4, shown in Fig. 8, acknowledges these remarks, as it shows the lowest global $\Delta v = 3,793 \text{ m/s}$ (with travel time $\Delta t = 88$ days). These considerations are in total accordance with what can be inferred from the rocket equation (see (42)). In order to reduce the total mass consumption, Δv has to decrease and I_{sp} has to increase. Finally, as far as it concerns the transfer times, the computed solutions presented turn out to be comparable with standard low energy transfers in terms of time of flight, although the latter are outperformed in terms of propellant mass.

Moreover, as far as it concerns sol.3, in Fig. 9c and in Fig. 9d are reported the control history profile and the mass consumption, respectively. It is evident that the thrust magnitude is always below the saturation limit (dashed line in Fig. 9c). Furthermore, when the engine is on duty (at the end of the trajectory in order to perform the low-thrust capture around the Moon), it works at the maximum level: the control profile is optimal as it recalls an *on-off* structure.

7 Conclusions

In this paper, two different techniques to design Earth-to-Moon transfers have been investigated. The optimized solutions are revealed to be efficient, both in terms of propellant mass consumption and flight time.

References

1. Belbruno, E.: Lunar Capture Orbits, a Method of Constructing Earth Moon Trajectories and the Lunar Gas Mission. In: AIAA, DGLR, and JSASS, International Electric Propulsion Conference, 19th, Colorado Springs, CO, May 11-13, 1987. 10 p. (1987)
2. Belbruno, E.: The Dynamical Mechanism of Ballistic Lunar Capture Transfers in the Four-Body Problem from the Perspective of Invariant Manifolds and Hill's Regions. Barcelona, Spain (1994)
3. Belbruno, E., Miller, J.: Sun-Perturbed Earth-to-Moon Transfers with Ballistic Capture. Journal of Guidance, Control and Dynamics **16**, 770–775 (1993)
4. Betts, J.: Survey of Numerical Methods for Trajectory Optimization. Journal of Guidance, Control and Dynamics **21**(2), 193–207 (1998)
5. Conley, C.: Low Energy Transit Orbits in the Restricted Three-Body Problem. SIAM Journal of Applied Mathematics **16**, 732–746 (1968)
6. Dellnitz, M., Junge, O., Post, M., Thiere, B.: On Target for Venus: Set Oriented Computation of Energy Efficient Low Thrust Trajectories. Celestial Mechanics and Dynamical Astronomy **95**, 357–370 (2006)
7. Enright, P., Conway, B.: Discrete Approximations to Optimal Trajectories Using Direct Transcription and Nonlinear Programming. Journal of Guidance, Control and Dynamics **15**, 994–1002 (1992)
8. Gómez, G., Koon, W., Lo, M., Marsden, J., Masdemont, J., Ross, S.: Invariant Manifolds, the Spatial Three-Body Problem, and Space Mission Design. Advances of the Astronautical Sciences **109**, 3–22 (2001)

9. Koon, W., Lo, M., Marsden, J., Ross, S.: Low Energy Transfer to the Moon. *Celestial Mechanics and Dynamical Astronomy* **81**, 63–73 (2001)
10. Llibre, J., Martinez, R., Simó, C.: Transversality of the Invariant Manifolds associated to the Lyapunov Family of Periodic Orbits near L_2 in the Restricted Three-Body Problem. *Journal of Differential Equations* **58**, 104–156 (1985)
11. Mingotti, G., Topputo, F., Bernelli-Zazzera, F.: Numerical Methods to Design Low-Energy, Low-Thrust Sun-Perturbed Transfers to the Moon. In: *Proceedings of 49th Israel Annual Conference on Aerospace Sciences*, Tel Aviv–Haifa, Israel, pp. 1–14 (2009)
12. Mingotti, G., Topputo, F., Bernelli-Zazzera, F.: Combined Optimal Low-Thrust and Stable-Manifold Trajectories to the Earth–Moon Halo Orbits. In: *American Institute of Physics Conference Proceedings*, vol. 886, pp. 100–110 (2007)
13. Ozimek, M., Grebow, D., Howell, K.: Design of Solar Sail Trajectories with Applications to Lunar South Pole Coverage. *Journal of Guidance, Control and Dynamics* **32**(6) (2009)
14. Parker, J., Lo, M.: Shoot The Moon 3D. *Advances in the Astronautical Sciences* **123** (2006)
15. Perozzi, E., Di Salvo, A.: Novel Spaceways for Reaching the Moon: an Assessment for Exploration. *Celestial Mechanics and Dynamical Astronomy* **102**, 201–218 (2008)
16. Schoenmaekers, J., Horas, D., Pulido, J.: SMART-1: With Solar Electric Propulsion to the Moon. In: *Proceedings of the 16th International Symposium on Space Flight Dynamics* (2007)
17. Simó, C., Gómez, G., Jorba, A., Masdemont, J.: The Bicircular Model near the Triangular Libration Points of the RTBP. In: *From Newton to Chaos*, pp. 343–370 (1995)
18. Sweetser, T.: An Estimate of the Global Minimum DV Needed for Earth–Moon Transfer. In: *Proceedings of the 1st AAS/AIAA Spaceflight Mechanics Meeting*, pp. 111–120 (1991)
19. Szebehely, V.: *Theory of Orbits: the Restricted Problem of Three Bodies*. Academic Press New York (1967)
20. Topputo, F., Vasile, M., Bernelli-Zazzera, F.: Earth-to-Moon Low Energy Transfers Targeting L1 Hyperbolic Transit Orbits. *Annals-New York Academy of Sciences* **1065**, 55 (2005)
21. Yagasaki, K.: Sun-Perturbed Earth-to-Moon Transfers with Low Energy and Moderate Flight Time. *Celestial Mechanics and Dynamical Astronomy* **90**(3), 197–212 (2004)

Low-Energy Earth-to-Halo Transfers in the Earth–Moon Scenario with Sun-Perturbation

Anna Zanzottera, Giorgio Mingotti, Roberto Castelli, and Michael Dellnitz

1 Introduction

This paper deals with the design of trajectories connecting LEOs (low Earth orbits) with halo orbits around libration points of the Earth–Moon CRTBP [1] using impulsive maneuvers. The interest for such transfers comes mainly from the importance of halo orbits as possible locations for lunar far-side data relay satellites; in particular, a satellite evolving on a halo orbit will always maintain line of sight contact with the Earth and Moon’s far side [6].

Indeed, it is taken into account a $h_E = 167$ km altitude parking orbit and a family of halo orbits associated to L_2 Lagrangian point. As widely used in literature, suitable first guess trajectories are derived exploiting the *coupled circular restricted three-body problem approximation* [2, 7, 9]. This method consists in the superposition of two different CRTBPs, namely the SE (Sun–Earth) and EM (Earth–Moon) restricted problems. The transfer trajectories are obtained patching together on convenient Poincaré sections the invariant manifolds related to periodic orbits of the two systems. Such transfers are achieved if the Poincaré maps of the two problems intersect almost exactly in the configuration space.

In the current case, intersections between the stable manifolds – related to EM halo orbits – and Earth escape trajectories – integrated in the planar SE CRTBP – have to be detected. This is done using the software package GAIO [4]: n -dimensional Poincaré maps are replaced by a collection of n -dimensional *boxes*, each one identified by a vector containing its center and the radii in each dimension. The implemented box covering structure allows to deal with flows of different dimensions in an efficient way. Moreover, the accuracy of the intersections in the

A. Zanzottera (✉)

Dipartimento di Matematica F. Brioschi, Piazza Leonardo da Vinci, 33,
20133 Milano, Italy

e-mail: anna.zanzottera@mail.polimi.it

configuration space, as well as the discontinuities in terms of Δv , are controlled through the parameters of the box covering. This exploration is systematically performed combining various halo orbits and Earth-escaping trajectories, with a choice of different Poincaré sections.

Then, first guess solutions are optimized through a direct-method approach and multiple-shooting technique [3], in the framework of the Sun-perturbed Earth–Moon bicircular four-body problem [11]. Finally, trajectories with single-impulse and two-impulse maneuvers are presented and compared with results already known in literature [10].

2 Dynamical Models

In this section, we present the dynamical system that we will use to study the motion of a spacecraft in a field of three massive bodies.

2.1 Circular Restricted Three-Body Problem

The circular restricted three-body problem (CRTBP) [1] studies the motion of a massless particle P moving in the gravitational field of two main primaries, with masses $m_1 > m_2$. The primaries are supposed to move under their mutual gravity in circular orbits around the center of mass and their motion is not affected by the third particle. In this paper, P stands for the spacecraft while the role of primaries is played by the Sun and the Earth (SE CRTBP) or by the Earth and the Moon (EM CRTBP).

In a rotating reference frame where the units of measure are normalized so that the distance between the primaries, the modulus of their angular velocity and the total mass are equal to 1, the motion for the third body is governed by the equation

$$\begin{cases} \ddot{x} - 2\dot{y} = \Omega_x \\ \ddot{y} + 2\dot{x} = \Omega_y \\ \ddot{z} = \Omega_z \end{cases} \quad (1)$$

where $\Omega(x, y, z) = \frac{1}{2}(x^2 + y^2) + \frac{1-\mu}{r_1} + \frac{\mu}{r_2} + \frac{1}{2}\mu(1-\mu)$ is the effective potential of the system and the subscripts denote partial derivatives. Here $\mu = m_2/(m_1 + m_2)$ is the mass ratio, while $r_1^2 = (x + \mu)^2 + y^2 + z^2$ and $r_2^2 = (x - 1 + \mu)^2 + y^2$ are the distances from the spacecraft respectively to the larger and the smaller primary. The system (1) admits a first integral of motion, the Jacobi integral, defined as:

$$C(x, y, z, \dot{x}, \dot{y}, \dot{z}) = 2\Omega(x, y, z) - (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) \quad (2)$$

As C^* varies, the family of five-dimensional energy manifolds

$$\mathcal{M}(C^*) = \{(x, y, z, \dot{x}, \dot{y}, \dot{z}) : C(x, y, z, \dot{x}, \dot{y}, \dot{z}) = C^*\}$$

is a foliation of the six-dimensional phase space. For every C^* the solution in the configuration space of the equation $C^* = 2\Omega(x, y, z)$ detects the zero-velocity surface, which bounds the Hill's region where the motion is possible.

The system (1) admits five equilibrium points, referred to as Lagrange points and denoted with Li , $i : 1 \dots 5$: three of them $L1, L2, L3$ lie on the x-axis and represent collinear configurations, while $L4, L5$ correspond to equilateral configurations.

The topology of the Hill's region changes in correspondence to the values C_i of the Jacobi constant relative to the libration points, allowing to open necks between different regions on the configuration space [8].

The collinear libration points, as well as the continuous families of planar and spatial periodic orbits surrounding them, have a saddle-center type stability character: the invariant manifold related to these orbits act as separatrices in the energy manifold and provide dynamical channels in the phase space useful for design low cost spacecraft trajectories.

2.2 Bicircular Restricted Four-Body Problem

The bicircular four-body model (BCRFBP) is a restricted four-body problem where two of the primaries (the Earth and the Moon) are moving in circular orbit around their center of mass B that is at the same time orbiting, together with the last mass (the Sun), around the barycenter of all the system. The motion of the primaries is supposed to be coplanar and with constant angular velocity. The equations of motion of the BCRFBP are written in the EM synodical reference frame and the physical quantities are normalized as in the EM CRTBP. Let m_s be the mass of the Sun, R_s the distance between the Sun and the origin of the frame B and ω_s the angular velocity of the Sun. Moreover let $\theta(t)$ be the angle between the Earth–Moon line and the Sun, and r_s the Sun–spacecraft distance (see Fig. 1):

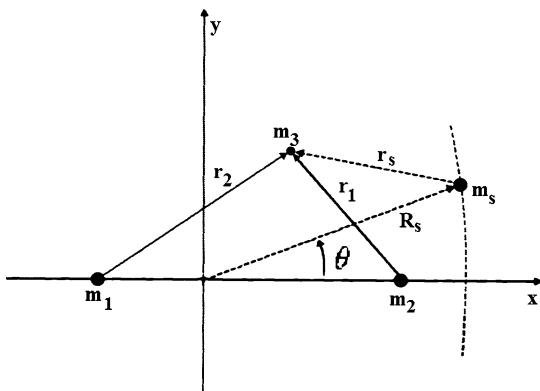
$$r_s^2 = (x - R_s \cos \theta)^2 + (y - R_s \sin \theta)^2 + z^2. \quad (3)$$

The motion of the massless particle solves the system of differential equations

$$\begin{cases} \ddot{x} - 2\dot{y} = \Omega_{Bx} \\ \ddot{y} + 2\dot{x} = \Omega_{By} \\ \ddot{z} = \Omega_{Bz}, \end{cases} \quad (4)$$

where $\Omega_B = \Omega + \frac{m_s}{r_s} - \frac{m_s}{R_s^2}(x \cos \theta + y \sin \theta)$. In literature it is common to refer to the dynamical system (4) as the Sun–Perturbed Earth–Moon Restricted Three-Body problem [11].

Fig. 1 Bicircular four body model



3 Trajectory Design

The aim of this paper is to find trajectories in the bicircular model, starting from a LEO and targeting a halo orbit around $L2$ in the EM-restricted problem. Since the phase space of the four-body system is poor of useful dynamical properties, the design is first performed in the CRTBP and then the initial guess trajectories are optimized to be solution of the bicircular model.

The model adopted as an approximation of the Sun–Earth–Moon-spacecraft restricted four-body problem is the so-called Patched Restricted Three-Body Problem. It consists the superposition of two CRTBPs with a common primary, namely the Sun–Earth CRTBP and the Earth–Moon CRTBP. The structure of the invariant manifolds associated with periodic orbits around the collinear libration points provides natural transfers from and to the smaller primaries. Then, by means of a Poincaré section the two legs of the trajectory are joined together, eventually with an impulsive maneuver, yielding a low-energy ballistic transfer.

According to this procedure, the design of LEO-to-halo trajectories is conceived in two different stages: the *Earth escape stage* and the *halo capture stage*. In the first part the planar SE CRTBP is exploited to leave a LEO orbit of $h_E = 167$ km of altitude, then the invariant manifold structure in the three-dimensional EM CRTBP is followed to obtain a natural transfer to halo orbits. The Poincaré section is chosen as a hyperplane in the phase space passing through the Earth and perpendicular to the (x, y) plane. The inclinations with respect to the positive x -semiaxis in the SE and EM synodical reference frame are denoted respectively with φ_{SE} and φ_{EM} and represent two of the design parameters.

3.1 Earth Escape Stage

As above mentioned, the departure leg of the transfer consists in a trajectory leaving a LEO orbit and integrated in planar SE CRTBP until the Poincaré section is

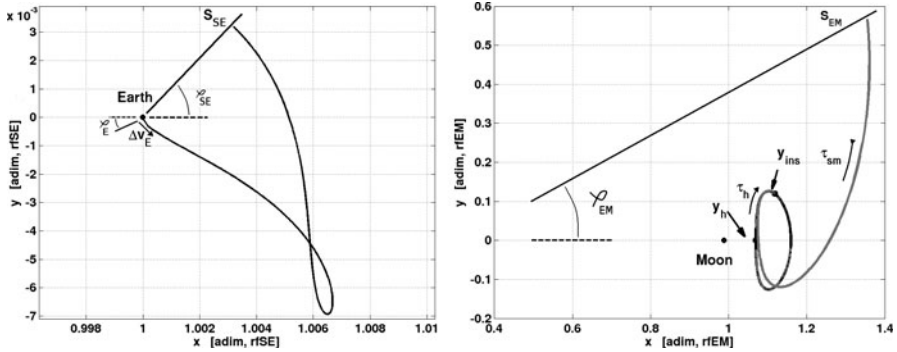


Fig. 2 The two stages of the Earth-to-halo trajectory design. (*left*) Earth escape trajectory. (*right*) Halo arrival trajectory

reached. The launch point $\mathbf{y}_E(\varphi_E, \Delta\varphi_E)$ is identified by the angle φ_E that represents the position on LEO and by the $\Delta\varphi_E$ maneuver applied to insert the spacecraft into the translunar trajectory (see Fig. 2(left)). The maneuver is chosen tangential to the parking orbit with magnitude in the range $I = [3200, 3285]$ m/s. The choice of the interval I follows from the fact that the Jacobi constant, associated to the spacecraft, in the SE restricted model, needs to be decreased from the value related to a LEO orbit, $C_{SE} \approx 3.07053$, to a value just below C_2 , that has been shown to reveal good opportunities for the design.

For a large set of value of $\varphi_E \in [0, 2\pi]$ and every integer value of $\Delta\varphi_E \in I$, the initial state $\mathbf{y}_E(\varphi_E, \Delta\varphi_E)$ is forward integrated in the SE CRTBP at most for one unit of time. For a value of the angle φ_{SE} , let P_{SE} be the set of all the intersections that such paths have with the selected Poincaré section.

3.2 Halo Capture Stage

The second phase of the design exploits the (exterior) stable branch of the EM manifold related to a target halo orbit λ_2 around L_2 . The stable manifold $W^s(\lambda_2)$ is integrated backwards starting from λ_2 until the surface of the section is reached: let P_{EM} the associated Poincaré map.

The halo orbit is identified by the nominal point \mathbf{y}_h where the plane $\{y = 0\}$ is intersected with positive \dot{y} , while the state $\mathbf{y}(t)$ of a generic point on the stable manifold is determined by two time parameters (τ_h, τ_{sm}) . Indeed, denoting with ϕ the flow, let $\mathbf{y}_{ins} = \phi(\mathbf{y}_h, 0; \tau_h)$ be a point along the Halo orbit ($\tau_h \geq 0$) and $\tilde{\mathbf{y}}_{ins}$ a slight shift of \mathbf{y}_{ins} along the stable direction of the linearized system. The time $\tau_{sm} < 0$ is a parameter of the trajectory $\mathbf{y}_{sm} = \phi(\tilde{\mathbf{y}}_{ins}, 0; \tau_{sm})$ on the stable manifold (see Fig. 2(right)).

3.3 Box Covering Technique

As the parameters φ_{SE} and φ_{EM} vary, the design of the trajectory is restricted to the detection on the Poincaré section of two points, $\mathbf{y}_{SE} \in P_{SE}$ and $\mathbf{y}_{EM} \in P_{EM}$. First the map P_{SE} is transformed in EM synodical coordinate system, being $\beta = \varphi_{SE} - \varphi_{EM}$ the angle between the Moon and the Sun-Earth line. Then, in order to obtain a feasible transfer, the points \mathbf{y}_{SE} and \mathbf{y}_{EM} has to match almost exactly in configuration space. Moreover, they should minimize the distance in velocity space. Note that while it is possible to achieve intersections in configuration space between P_{SE} and the subset $P_{EM} \cap \{|z| = 0\}$, a Δv maneuver it's always necessary. Indeed on the $\{z = 0\}$ plane the out of plane component of the velocity of the point \mathbf{y}_{EM} it's always different from zero.

The search of candidate transfer points is performed using a box covering structure implemented in the software package GAIO (Global Analysis of Invariant Objects), see [4] for a detailed description. An n -dimensional box $\mathcal{B}(C, R)$, identified by a center $C = (C_1, \dots, C_N) \in \mathbb{R}^N$ and a vector of radii $R = (r_1, \dots, r_n) \in \mathbb{R}^N$, is defined as

$$\mathcal{B}(C, R) = \cap_{i=1}^N \{(x_1, x_2, \dots, x_N) \in \mathbb{R}^N : |x_i - C_i| < r_i\}.$$

Starting from an initial box \mathcal{B}_0 containing the projection of P_{EM} on the configuration space, a multiple subdivision process is carried out to create families \mathcal{F}_k of smaller boxes $\{\mathcal{B}_k\}$ with the property to cover \mathcal{B}_0 , i.e. $\bigcup \mathcal{B}_k = \mathcal{B}_0$. In the k th subdivision step each rectangle $\mathcal{B}(C, R)$ of the existing collection is subdivided along the j -th coordinate, where j can vary cyclically or can be fixed by the user. Once the radii of the boxes in \mathcal{F}_k reach a prescribed size $\bar{\sigma}$, the Poincaré map is inserted: only those boxes \mathcal{B}_k with nonempty intersection with P_{EM} are stored, otherwise removed. Denoting with \mathcal{F} the collection of the remaining boxes, the feasibility condition previously discussed is fulfilled choosing the possible transfer points in the intersection $\mathcal{F} \cap P_{SE}$.

In the numerical simulation here presented, the Poincaré maps are at first transformed in EM cylindrical coordinates $(r_1, \theta, z, \dot{r}_1, r_1 \dot{\theta}, \dot{z})$ centered on the Earth, then inserted into a collection of boxes with radii $\bar{\sigma}$, at most equal to 10^{-4} in r_1 and z coordinates. In Fig. 3(left), the intersection – in the space of configurations – of the Earth escape trajectories (P_{SE} black line) and the stable manifold related to the final halo target (grey curve $W^s(\lambda_2)$) is shown. The black dots ($W^s_{z=0}(\lambda_2)$) stand for the intersection of P_{SE} with the $\{z = 0\}$ plane. According to Fig. 3(right), it is possible to detect the candidate transfer points as the ones corresponding to $P_{SE} \cap W^s(\lambda_2)$.

4 Trajectory Optimization

This section gives firstly a brief introduction of the trajectory optimization approach used in this work, then formulates in details the minimization problems later solved.

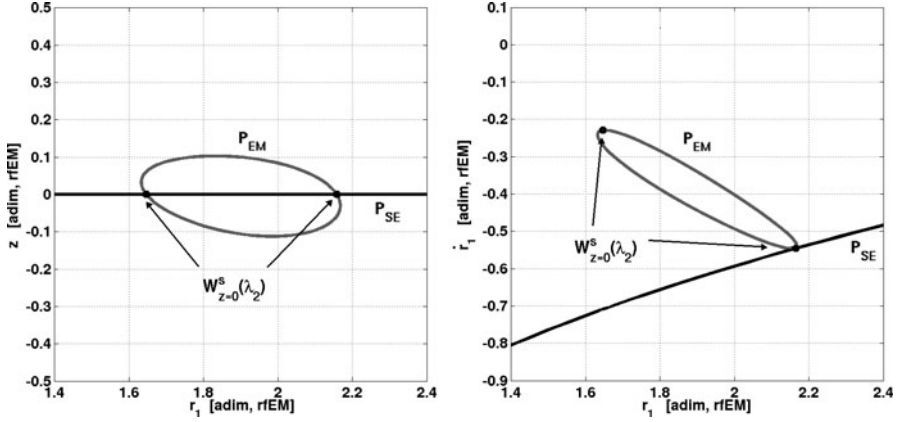


Fig. 3 Transfer point definition on suitable Poincaré maps. (*left*) $P_{SE} \cap W^s(\lambda_2)$ in the r_1, z coordinates. (*right*) $P_{SE} \cap W^s(\lambda_2)$ in the r_1, \dot{r}_1 coordinates

Once feasible and efficient first guess solutions are achieved, combining the two legs of the transfer, an optimization problem is stated. A given objective function is minimized taking into account the dynamic of the process. The dynamical model used to consider the gravitational attractions of all the celestial bodies involved in the design process (i.e. the Sun, the Earth, and the Moon) is the spatial BCRFBP described by (4) and here written in an autonomous fashion:

$$\begin{cases} \ddot{x} - 2\dot{y} = \Omega_{Bx} \\ \ddot{y} + 2\dot{x} = \Omega_{By} \\ \ddot{z} = \Omega_{Bz} \\ \dot{\theta} = \omega_s. \end{cases} \quad (5)$$

According to the formalism proposed by [3], the BCRFBP described by (5) is written in the first-order form

$$\begin{aligned} \dot{x} &= v_x \\ \dot{y} &= v_y \\ \dot{z} &= v_z \\ \dot{v}_x &= 2v_y + \Omega_{Bx} \\ \dot{v}_y &= -2v_x + \Omega_{By} \\ \dot{v}_z &= \Omega_{Bz} \\ \dot{\theta} &= \omega_s, \end{aligned} \quad (6)$$

with $v_x = \dot{x}$, $v_y = \dot{y}$ and $v_z = \dot{z}$. In a compact explicit form, system (6) read as

$$\dot{\mathbf{y}} = \mathbf{f}[\mathbf{y}(t), \mathbf{p}, t], \quad (7)$$

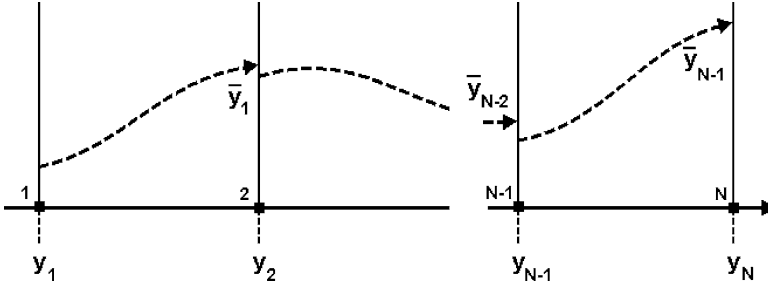


Fig. 4 Direct multiple shooting scheme

where \mathbf{f} stands for the vector field and the state vector is $\mathbf{y} = \{x, y, z, \dot{x}, \dot{y}, \dot{z}, \theta\}^\top$. The aim is finding $\mathbf{y} = \mathbf{y}(t)$, $t \in [t_i, t_f]$, that minimizes a prescribed scalar performance index or objective function

$$J = J(\mathbf{y}, \mathbf{p}, t), \quad (8)$$

while satisfying certain mission constraints. These constraints are represented by the two boundary conditions, defined at the end points of the optimization problem, and by the inequality conditions, defined along the whole arc. These last quantities are derived specifically for the mission investigated. Moreover, \mathbf{p} stands for a vector which brings together some free parameters useful for the optimization process.

The optimization problem, OP, is then transcribed into a nonlinear programming, NLP, problem using a direct approach. This method, although suboptimal, generally shows robustness and versatility, and does not require explicit derivation of the necessary conditions of optimality. Moreover, direct approaches offer higher computational efficiency and are less sensitive to variation of the first guess solutions [3]. Furthermore, a multiple shooting scheme is implemented. With this strategy the BCRFBP dynamics presented by (5) is forward integrated within $N - 1$ intervals (in which $[t_i, t_f]$ is uniformly split), i.e. the time domain is divided in the form $t_i = t_1 < \dots < t_N = t_f$, and the solution is discretized over the N grid nodes (see Fig. 4). The continuity of position and velocity is imposed at their ends [5], in the form of defects $\eta_j = \bar{\mathbf{y}}_j - \mathbf{y}_{j+1} = 0$, for $j = 1, \dots, N - 1$. The quantity $\bar{\mathbf{y}}_j$ stands for the result of the integration, i.e. $\bar{\mathbf{y}}_j = \phi(\mathbf{y}_j, \mathbf{p}, t)$, $t_j \leq t_{j+1}$. The algorithm computes the value of the states at mesh points, satisfying both boundary and path constraints, and minimizing the performance index.

Dynamics described by (5) are highly nonlinear and, in general, lead to chaotic orbits. In order to find accurate optimal solutions without excessively increasing the computational burden, an adaptive nonuniform time grid has been implemented. Thus, when the trajectory is close to either the Earth or the Moon the grid is automatically refined, whereas in the intermediate phase, where a weak vector field governs the motion of the spacecraft, a coarse grid is used. The optimal solution found is assessed a posteriori by forward integrating the optimal initial condition (with a Runge–Kutta 8th order scheme).

4.1 Two-Impulse Problem Statement

In this section, the approach previously described is exploited to obtain optimal transfers with two-impulsive maneuvers.

According to the NLP formalism recalled, the variable vector is

$$\mathbf{x} = \{\mathbf{y}_1, \dots, \mathbf{y}_N, \mathbf{p}, t_1, t_N\}^\top, \quad (9)$$

where $\mathbf{p} = \{\tau_h, \tau_{sm}\}$, is made up of two free optimization parameters useful to describe the final condition of the transfer (see Fig. 2).

The initial conditions read:

$$\psi_i(\mathbf{y}_1, t_1) := \begin{cases} (x_1 + \mu)^2 + y_1^2 + z_1^2 - r_i^2 = 0 \\ (x_1 + \mu)(\dot{x}_1 - y_1) + y_1(\dot{y}_1 + x_1 + \mu) + z_1\dot{z}_1 = 0, \end{cases} \quad (10)$$

which force the first \mathbf{y}_1 state of the transfer to belong to a circular orbit of radius $r_i = R_E + h_E$, where R_E and h_E stand for the Earth radius and the orbit altitude with respect to the Earth, respectively. The transfer ends when the spacecraft flies on the stable manifold related to the final halo. In details, only the continuity in terms of position is imposed, so that the final condition reads

$$\psi_f = \bar{\mathbf{y}}_N - \bar{\mathbf{y}}_{sm} = 0, \quad (11)$$

where it is worth noting that $\bar{\mathbf{y}}_N = \{x_N, y_N, z_N\}^\top$ and $\bar{\mathbf{y}}_{sm} = \{x_{sm}, y_{sm}, z_{sm}\}^\top$. This means that, after the initial impulsive maneuver, a second one is required to inject the spacecraft onto the stable manifold that takes it ballistically to the final halo orbit associated.

The nonlinear equality constraint vector, made up of the boundary conditions and the ones representing the dynamics, is therefore written as follows:

$$\mathbf{c}(\mathbf{x}) = \{\psi_i, \eta_1, \dots, \eta_{N-1}, \psi_f\}^\top. \quad (12)$$

Moreover, aiming at avoiding the collision with the two primaries, the following inequality constraints are imposed:

$$\Psi_j^c(\mathbf{y}_j) := \begin{cases} R_E^2 - (x_j + \mu)^2 - y_j^2 - z_j^2 \leq 0 \\ R_M^2 - (x_j - 1 + \mu)^2 - y_j^2 - z_j^2 \leq 0, \end{cases} \quad j = 2, \dots, N-1. \quad (13)$$

Finally, the flight time is searched to be positive, i.e.

$$\Psi^t = t_1 - t_N \leq 0. \quad (14)$$

The complete inequality constraint vector therefore reads:

$$\mathbf{g}(\mathbf{x}) = \{\Psi_2^c, \dots, \Psi_{N-1}^c, \Psi^t\}^\top. \quad (15)$$

The performance index to minimize is a scalar that represents the two velocity variations at the beginning and at the final node of the transfer, i.e. $J(\mathbf{x}) = \Delta v_1 + \Delta v_N$. In details,

$$\Delta v_1 = \sqrt{(\dot{x}_1 - y_1)^2 + (\dot{y}_1 + x_1 + \mu)^2 + (z_1)^2} - v_i, \quad (16)$$

assuming $v_i = \sqrt{(1-\mu)/r_i}$ as the velocity along the initial circular parking orbit, and

$$\Delta v_N = \sqrt{(\dot{x}_N - \dot{x}_{sm})^2 + (\dot{y}_N - \dot{y}_{sm})^2 + (\dot{z}_N - \dot{z}_{sm})^2}, \quad (17)$$

which represents the discontinuity in terms of velocity between the translunar trajectory and the stable manifold related to the final halo.

In summary, the NLP problem for the two-impulse transfers is formulated as follows:

$$\begin{aligned} \min_{\mathbf{x}} J(\mathbf{x}) \quad & \text{subject to } \mathbf{c}(\mathbf{x}) = 0, \\ & \mathbf{g}(\mathbf{x}) \leq 0. \end{aligned} \quad (18)$$

4.2 Single-Impulse Problem Statement

Aiming to obtain single-impulse trajectories, the equality constraint vector is defined, as in the previous paragraph, by (12) with the exception of the final condition. In this case, at the final point of the transfer, the whole dynamical state is forced to be equal to the one associated to the stable manifold:

$$\boldsymbol{\psi}_f = \mathbf{y}_N - \mathbf{y}_{sm} = 0, \quad (19)$$

where $\mathbf{y}_N = \{x_N, y_N, z_N, \dot{x}_N, \dot{y}_N, \dot{z}_N\}^\top$ and $\mathbf{y}_{sm} = \{x_{sm}, y_{sm}, z_{sm}, \dot{x}_{sm}, \dot{y}_{sm}, \dot{z}_{sm}\}^\top$.

Instead, the inequality constraint vector is defined in the same way as in the two-impulse scenario.

Dealing with the objective index to minimize, this is made up of only the initial velocity maneuver, i.e. $J(\mathbf{x}) = \Delta v_1$, where

$$\Delta v_1 = \sqrt{(\dot{x}_1 - y_1)^2 + (\dot{y}_1 + x_1 + \mu)^2 + (z_1)^2} - v_i. \quad (20)$$

Finally, the NLP problem for the low-energy low-thrust transfers is formulated as proposed by (18) at the end of the previous section.

5 Optimized Transfer Solutions

In this section, the transfers to halos obtained solving the optimization process are presented. In the previous two sections, two families of trajectories are discussed according to the number of impulsive maneuvers that are allowed. In the following, the optimized solutions are proposed in terms of some relevant performance parameters.

5.1 Trajectories to Halos

Optimal two-impulse and single-impulse solutions are presented. These transfers start from a circular parking orbit at an altitude of $h_E = 167$ km around the Earth, and reach a halo orbit around L_2 , with an out-of-plane amplitude of $A_z = 8,000$ km. Referring to Table 1: the first sol.1 corresponds to the two-impulse low energy transfer, while solution sol.2 represents a single-impulse low energy transfer. The last two solutions presented are some reference impulsive transfers found in literature.

More in details, Table 1 is so structured: the second column Δv_i stands for the initial impulsive maneuver that inserts the spacecraft onto the translunar trajectory. The third column Δv_f represents the final impulsive maneuver that permits the insertion of the spacecraft onto the stable manifold related to the target halo. The fourth column Δv_t represents the overall amount of impulsive maneuvers necessary to complete the Earth-to-halo transfers. Finally, the last column on the right stands for the transfer time.

An analysis of the table shows that the single-impulse sol.2 offers the lowest value of the overall impulsive maneuvers (see Δv_t). This happens because the first guess solution takes explicitly advantage of the initial lunar flyby (see Fig. 5(left)). The latter can be seen as a kind of aid in the translunar orbit insertion, as it reduces the Δv_i required for that maneuver. The lunar flyby performs a change of plane of the translunar trajectory that allows the insertion of the spacecraft onto the three-dimensional halo stable manifold without any other maneuver. Summarizing, the single-impulse trajectory corresponding to sol.2 acknowledges these remarks, as it shows the lowest global $\Delta v_t = 3161$ m/s (with travel time $\Delta t = 98$ days).

Table 1 Two-impulse and single-impulse low energy transfers to halos around L_2 . A set of impulsive reference solutions found in literature is also reported [10]

Type	Δv_i (m/s)	Δv_f (m/s)	Δv_t (m/s)	Δt (days)
Sol.1	3110	214	3,324	97
Sol.2	3161	0	3,161	98
Parker.1	3132	618	3,750	–
Parker.2	3235	0	3,235	–

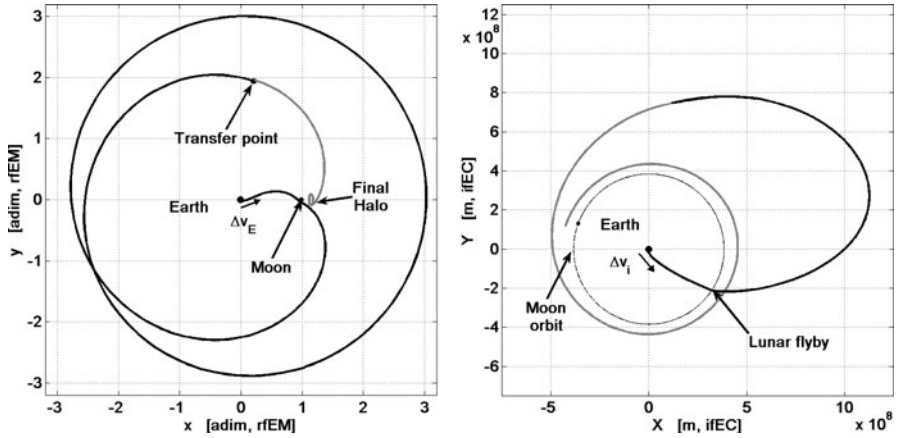


Fig. 5 First guess solution and optimized trajectory corresponding to sol.2 of Table 1. (*left*) First guess solution in Earth–Moon rotating frame. (*right*) Optimized trajectory in Earth centered inertial frame

6 Conclusion and Future Works

In this paper, a technique to design low-energy transfers to halo orbits has been investigated. Through the box-covering technique, an immediate definition of the transfer points in the phase space was possible, formalizing in a systematic way the intersection of three-dimensional manifolds. The optimization approach resulted robust and versatile, allowing to obtain single and two impulse transfer efficiently both in terms of flight time and Δv .

A more detailed analysis of the design process and further results will be presented and discussed in a future paper.

Acknowledgments This work was supported by the Marie Curie Actions Research and Training Network AstroNet, Contract Grant No. MCRTN-CT-2006-035151.

References

1. Szebehely, V.: Theory of Orbits: the Restricted Problem of Three Bodies, Academic Press New York (1967)
2. Belbruno, E.A. and Miller, J.K.: Sun-Perturbed Earth-to-Moon Transfers with Ballistic Capture. *Journal of Guidance Control and Dynamics* **16**, 770–775 (1993)
3. Betts, J.T.: Survey of Numerical Methods for Trajectory Optimization. *Journal of Guidance control and dynamics* **21**, 193–207 (1998)
4. Dellnitz, M. and Junge, O.: Set Oriented Numerical Methods for Dynamical Systems. *Handbook of dynamical systems* **2**, North-Holland (2002)

5. Enright, P.J. and Conway, B.A.: Discrete Approximations to Optimal Trajectories Using Direct Transcription and Nonlinear Programming. *Journal of Guidance Control and Dynamics* **15**,994–1002 (1992)
6. Farquhar, R.: Future Missions for Libration-Point Satellites. *Astronautics and Aeronautics* **7**, 52–56 (1969)
7. Gómez, G. and Koon, WS and Lo, MW and Marsden, JE and Masdemont, J. and Ross, SD: Connecting orbits and invariant manifolds in the spatial restricted three-body problem. *Nonlinearity* **17**, IOP Publishing (2004)
8. Koon, W.S. and Lo, M.W. and Marsden, J.E. and Ross, S.D.: Heteroclinic Connections Between Periodic Orbits and Resonance Transitions in Celestial Mechanics. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **10**, (2000)
9. Koon, W. and Lo, M. and Marsden, J. and Ross, S.: Low Energy Transfer to the Moon. *Celestial Mechanics and Dynamical Astronomy* **81**, 63–73 (2001)
10. Parker, J.S.: Families of Low-Energy Lunar Halo Transfers. *AAS/AIAA Spaceflight Dynamics Conference* **90**, 1–20 (2006)
11. Simó, C. and Gómez, G. and Jorba, A. and Masdemont, J.: The Bicircular Model near the Triangular Libration Points of the RTBP. *From Newton to Chaos* pp. 343–370 (1995)

On the Relation Between the Bicircular Model and the Coupled Circular Restricted Three-Body Problem Approximation

Roberto Castelli

1 Introduction

Starting from the work of Belbruno and Miller [1], where the perturbation of the Sun has been shown to decrease the amount of fuel necessary for a Earth-to-Moon transfer, the Sun–Earth–Moon-spacecraft restricted 4-body model has been commonly adopted to describe the spacecraft motion.

One of the techniques used to approximate the 4-body dynamics, or in general the n -body problem, is the Coupled restricted three-body problem approximation: partial orbits from different restricted problems are connected into a single trajectory, yielding energy efficient transfers to the Moon [12], interplanetary transfers [6] or very complicated itineraries [10].

The procedure requires the choice of a Poincaré section where the phase spaces of the two different models have to intersect: the analysis of the Poincaré maps of the invariant manifolds reduces the design of the trajectory to the selection of a point on the section. The Poincaré section plays also a role in the accuracy of the approximation of the undertaken dynamical system: indeed the encounter with the Poincaré section is the criteria for switching from the first to the second restricted three-body problem. In the mentioned works, the Poincaré section is chosen *a priori* in order to accomplish certain design constraints or to simplify the selection of the connection point and it usually consists in a hyperplane passing through one of the primaries or lying along one of the coordinated axis.

Although it has been shown that for design purpose the solutions in a simplified model like the Circular restricted three-body problem (CR3BP) are very good approximations to real trajectories in the complicated and full system [15], this

R. Castelli (✉)

University of Paderborn, WarburgerStr. 100, 33098 Paderborn, Germany

e-mail: robertoc@math.upb.de

work deepens from a more theoretical point of view on the role played by the two restricted three-body problems in the approximation of the 4-body system.

The undertaken model considered here for the 4-body dynamics is the Bicircular model (BCP), [16], while the two restricted problems are the Earth–Moon CR3BP and the Sun–(Earth+Moon) CR3BP, where, in the last case, the Sun and the Earth–Moon barycenter act as primaries. The comparison of the mentioned systems leads to the definition of *Regions of Prevalence* in the space where one of the restricted problems performs, at least locally, the best approximation of the Bicircular model and therefore it should be preferred in designing the trajectory.

Then, setting the Poincaré section according to this prevalence, the coupled CR3BP approximation is implemented to design low energy transfers leaving Lyapunov orbits in the Sun–Earth system and leading to the Moon’s region.

The plan of the paper is the following. In the next section, the CR3BP is briefly recalled and the equations of motion for the BCP in the inertial reference frame are written. Then, in Sect. 3, the comparison between the BCP and each one of the restricted problems is performed: this analysis enables to define, in Sect. 4, the regions of prevalence of the two restricted systems in the approximation of the 4-body model. Section 5 concerns the design of the transfer trajectory, while Sect. 6 deepens on the numeral scheme used to analyze the Poincaré maps for the selection of the connection points. Finally in the last section, some of the results are discussed.

2 Dynamical Models

2.1 Circular Restricted Three-Body Problem

The CR3BP is a simplified case of the general Three-Body Problem and models the motion of a massless particle under the gravitational influence of two bodies, with masses $M_1 < M_2$, that are revolving with constant angular velocity in circular orbit around their center of mass, see [17]. In the following, only the planar motion is considered.

In a rotating reference frame centered in the center of mass and with the units of measure normalized so that the total mass, the distance between the primaries and their angular velocity are equal to 1, the motion $z(t) = x(t) + iy(t)$ of the massless particle evolves following the differential equation

$$\frac{d^2z}{dt^2} + 2i\frac{dz}{dt} - z = - \left[\frac{(1-\mu)(z+\mu)}{\|z+\mu\|^3} + \frac{\mu(z-(1-\mu))}{\|z-(1-\mu)\|^3} \right] \quad (1)$$

where $\mu = M_2/(M_1 + M_2)$ is the mass ratio. Note that the primaries are fixed on the x -axis and the bigger and the smaller mass are placed in positions $(-\mu, 0)$ and $(1-\mu, 0)$ respectively.

In (x, y) components, the equation of motion assumes the form

$$\ddot{x} - 2\dot{y} = \Omega_x, \quad \ddot{y} + 2\dot{x} = \Omega_y$$

where $\Omega(x, y) = (x^2 + y^2)/2 + (1 - \mu)/r_1 + \mu/r_2 + \mu(1 - \mu)/2$ is the potential function. The subscripts of Ω denote the partial derivatives, while $r_{1,2}$ are the distances between the moving particle and the primaries. The advantage to study the dynamics in a rotating frame is that the system (1) is Hamiltonian and autonomous and admits a first integral called Jacobi constant

$$J(x, y, \dot{x}, \dot{y}) = -(\dot{x}^2 + \dot{y}^2) + 2\Omega(x, y).$$

Therefore, the phase space is foliated in 3-dimensional energy manifolds

$$E(h) = \{(x, y, \dot{x}, \dot{y}) \in \mathbb{R}^4 : J(x, y, \dot{x}, \dot{y}) = h\}$$

whose projections onto the configuration space are known as Hill's regions. For any fixed value of h , the Hill's region prescribes the region where the particle is allowed to move.

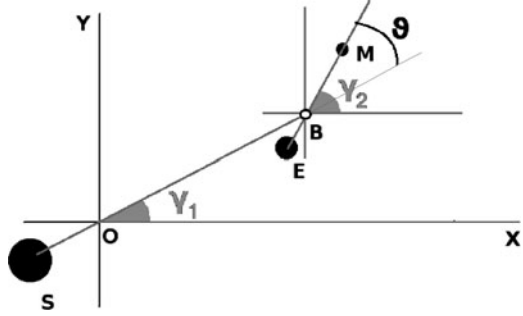
The potential Ω admits five critical points, the Lagrangian points $L_i, i = 1 \dots 5$, that represent equilibrium points for the vector field. The points $L_{4,5}$ correspond to equilateral triangle configurations, while the remaining are placed on the x -axis and correspond to collinear configurations of the masses. Of particular interest for mission design are L_1 and L_2 and the periodic orbits surrounding them that play the role of gates in the Hill's region, see for instance [11].

2.2 Bicircular Model

The Bicircular model (BCP), see [4], consists in a restricted four-body problem where two of the primaries are rotating around their center of mass, which is meanwhile revolving together with the third mass around the barycenter of the system. The massless particle is moving under the gravitational influence of the primaries and does not affect their motion. It is assumed that the motion of the primaries, like the motion of the test particle, is coplanar. The low eccentricity of the Earth's and Moon's orbit and the small inclination of the Moon's orbital plane, make the Bicircular a quite accurate model to describe the dynamics of a spacecraft in the Sun–Earth–Moon system; see for instance [18] and [14].

Referring to Fig. 1, let S, E, M be the positions of the three primaries, namely the Sun, the Earth, and the Moon while B and O indicate respectively the Earth–Moon barycenter and the total center of mass of the system. For a given time–space system of measure, let us define w_1 and w_2 the angular velocities of the couple S and B around O and of the couple E and M around B ; L_S and L_M the distances

Fig. 1 Positions of the primaries in the inertial reference frame



\overline{SB} and \overline{EM} ; M_m, M_e, M_s the masses of the Moon, the Earth and the Sun and G the gravitational constant. Moreover let μ_s and μ_m be the mass ratios

$$\mu_m = \frac{M_m}{M_e + M_m}, \quad \mu_s = \frac{M_e + M_m}{M_e + M_m + M_s}. \quad (2)$$

With respect to an inertial reference frame (X, Y) whose origin is fixed in the barycenter O and where τ denotes the time coordinate, the positions of the primaries are given by

$$\begin{aligned} S &= -\mu_s L_S e^{i(\phi_0 + w_1 \tau)} \\ E &= (1 - \mu_s) L_S e^{i(\phi_0 + w_1 \tau)} - \mu_m L_M e^{i(\phi_0 + w_2 \tau)} \\ M &= (1 - \mu_s) L_S e^{i(\phi_0 + w_1 \tau)} + (1 - \mu_m) L_M e^{i(\phi_0 + w_2 \tau)}. \end{aligned}$$

In order to lighten the notation, we define

$$\gamma_1(\tau) = \phi_0 + w_1 \tau, \quad \gamma_2(\tau) = \phi_0 + w_2 \tau.$$

The motion $Z(\tau) = X(\tau) + iY(\tau)$ of the spacecraft is governed by the second order differential equation

$$\frac{d^2 Z}{d\tau^2} = -G \left[\frac{M_s (Z - S)}{\|Z - S\|^3} + \frac{M_e (Z - E)}{\|Z - E\|^3} + \frac{M_m (Z - M)}{\|Z - M\|^3} \right]. \quad (3)$$

Later on the BCP will be compared with two different restricted three-body problems: the CR3BP_{EM} and the CR3BP_{SE}. The role of primaries is played by the Earth and the Moon in the first system and by the Sun and the barycenter B with mass $M_b = M_e + M_m$ in the latter. Three different reference frames and different units of measure are involved in the analysis: the inertial reference frame and the SE-synodical reference frame, whose origin is set in the center of mass O , and the EM-synodical reference frame centered on the point B .

2.3 Change of Coordinates

Following the notation previously adopted, let (X, Y, τ) be the space–time coordinates in the inertial reference frame and the small letters (x, y, t) the coordinates in the rotating systems. When necessary, in order to avoid any ambiguity, the subscript (x_s, y_s, t_s) and (x_m, y_m, t_m) are used to distinguish the set of coordinates in the CR3BP_{SE} and in the CR3BP_{EM} respectively. In complex notation

$$Z := X + iY, \quad z_m := x_m + iy_m, \quad z_s := x_s + iy_s$$

the relations between the inertial and the synodical coordinates are given by

$$\begin{aligned} Z &= L_S z_s e^{i\gamma_1}, & \tau &= \frac{t_s}{w_1} \\ Z &= (1 - \mu_s) L_S e^{i\gamma_1} + L_M z_m e^{i\gamma_2}, & \tau &= \frac{t_m}{w_2}. \end{aligned}$$

Concerning the two synodical systems, the relation between the time coordinates t_s and t_m is easily derived

$$t_s = \frac{w_1}{w_2} t_m$$

while the transformation between the space coordinates (x_s, y_s) and (x_m, y_m) depends on the mutual position of the primaries. Let θ be the angle between the positive x_s -semiaxis and the positive x_m -semiaxis, see Fig. 1:

$$\theta(\tau) := \gamma_2 - \gamma_1 = \theta_0 + (w_2 - w_1)\tau.$$

For any value of θ , the position and the velocity of a particle in the two different synodical systems satisfy the relations

$$\begin{aligned} z_m &= \frac{L_S}{L_M} e^{-i\theta} (z_s - (1 - \mu_s)) \\ \frac{dz_m}{dt_m} &= \frac{L_S}{L_M} \frac{w_1}{w_2} e^{-i\theta} \left[i \left(1 - \frac{w_2}{w_1} \right) (z_s - (1 - \mu_s)) + \frac{dz_s}{dt_s} \right] \end{aligned} \quad (4)$$

and

$$\begin{aligned} z_s &= \frac{L_M}{L_S} e^{i\theta} z_m + (1 - \mu_s) \\ \frac{dz_s}{dt_s} &= \frac{L_M}{L_S} \frac{w_2}{w_1} e^{i\theta} \left[i \left(1 - \frac{w_1}{w_2} \right) (z_m) + \frac{dz_m}{dt_m} \right]. \end{aligned}$$

A second integration provides the relations between the accelerations in the two systems:

$$\frac{d^2 z_s}{dt_s^2} = \frac{L_M}{L_S} \left(\frac{w_2}{w_1} \right)^2 e^{i\theta} \times \left[- \left(1 - \frac{w_1}{w_2} \right)^2 z_m + 2i \left(1 - \frac{w_1}{w_2} \right) \frac{dz_m}{dt_m} + \frac{d^2 z_m}{dt_m^2} \right]. \quad (5)$$

Moreover, as a consequence of the third Kepler's law,

$$w_1^2 L_S^3 = G(M_s + M_e + M_m), \quad w_2^2 L_M^3 = G(M_e + M_m). \quad (6)$$

In this work, the Sun–Earth–Moon scenario is considered and the physical parameters adopted in the numerical simulations are set according with the Jet Propulsion Laboratory ephemeris (available on-line at <http://ssd.jpl.nasa.gov/?constants>). In particular the mass ratios are

$$\mu_s = 3.040423402066 \times 10^{-6}, \quad \mu_m = 0.012150581$$

being the masses of the bodies

$$M_s = 1.988924 \times 10^{30} \text{ kg} \quad M_e = 5.973712 \times 10^{24} \text{ kg}$$

$$M_m = 7.347686 \times 10^{22} \text{ kg}.$$

In the inertial reference frame (X, Y, τ) the time is measured in seconds, the distances L_S and L_M are equal to

$$L_S = 149,597,870 \text{ km}, \quad L_M = 384,400 \text{ km},$$

while the values of the angular velocities w_1 and w_2 are

$$w_1 = 1.99098898 \times 10^{-7} \frac{\text{rad}}{\text{s}}, \quad w_2 = 2.6653174179 \times 10^{-6} \frac{\text{rad}}{\text{s}}.$$

3 The Comparison of the BCP with the CR3BPs

For applications in celestial mechanics and mission design, the dynamical model should be chosen to reproduce as better as possible the real force system. Moreover, it could be an advantage if the model presents useful mathematical properties so that dynamical system theory could be applied to heighten understanding and perspectives. These features are usually in conflict and, for mission design purposes,

the Circular restricted three-body problem reveals to be a good compromise. Indeed, this system admits a rich phase portrait that could be used to construct energy-efficient spacecraft trajectories. In particular, the structure of certain invariant sets and associated invariant manifolds have been exploited to accomplish a wide class of mission requirements, [6, 7, 9, 10, 13, 18] and enables to heuristically explain other theories like the Weak Stability Boundary theory, formalized in the Coupled CR3BP model, [1–3, 8]. On the other side, in different situations like the Sun–Earth–Moon scenario, solutions in the restricted model have been shown to be good approximations to trajectory in the real system.

Following the Coupled CR3BP approach, a complete trajectory in a multibody system is built patching together selected arcs computed in different restricted three-body problems. The sequence of the restricted models and the criteria where to connect the different solution arcs are usually chosen a priori to accomplish the mission requirements and to facilitate the design. Indeed the Poincaré section where the intersection of the manifolds is analyzed is usually set on straight lines in the configuration space or on the boundary of the three-body sphere of influence as in [15].

The incoming analysis concerns the relation between the Bicircular 4-body system and a couple of circular restricted three-body problem and aims to define the coupled CR3BP approximation so that each restricted model is considered where it provides the best approximation of the more complicated system. The distance between the Bicircular model and each one of the CR3BP is estimated as the norm of the difference of the differential equations governing their dynamics, once they are written in the same reference frame and in the same units of measure. In particular, the comparison is performed in the synodical frame proper of the considered restricted problem, while the units of measure in both the cases will be the dimensional one.

3.1 Comparison with CR3BP_{SE}

Denote with $\bar{z} = z_s L_S$ the spatial coordinate in the Sun–Earth rotating reference frame, measured in km. Since $Z = \bar{z} e^{i\gamma_1}$, in this setting the positions of the primaries are given by

$$\begin{aligned}\bar{S} &= -\mu_s L_S \\ \bar{E} &= (1 - \mu_s) L_S - \mu_m L_M e^{i(\gamma_2 - \gamma_1)} \\ \bar{M} &= (1 - \mu_s) L_S + (1 - \mu_m) L_M e^{i(\gamma_2 - \gamma_1)}.\end{aligned}$$

The double derivative of $Z(\tau)$ turns into

$$\frac{d^2 Z}{d\tau^2} = \left(\frac{d^2 \bar{z}}{d\tau^2} + 2i w_1 \frac{d\bar{z}}{d\tau} - w_1^2 \bar{z} \right) e^{i\gamma_1}$$

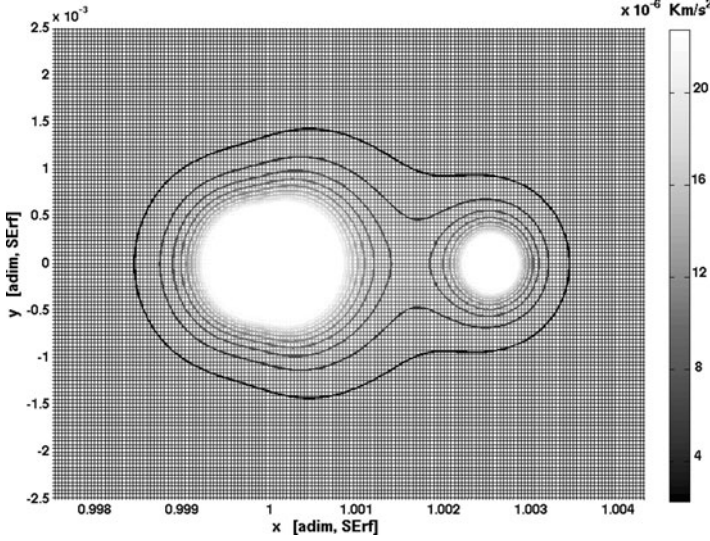


Fig. 2 Level curves of Δ_{SE} for $\theta = 0$

then, substituting the new variables into (3), the equation of motion associated to the BCP assumes the form

$$\frac{d^2 \bar{z}}{d\tau^2} + 2iw_1 \frac{d\bar{z}}{d\tau} - w_1^2 \bar{z} = -G \left[\frac{M_s(\bar{z} - \bar{S})}{\|\bar{z} - \bar{S}\|^3} + \frac{M_e(\bar{z} - \bar{E})}{\|\bar{z} - \bar{E}\|^3} + \frac{M_m(\bar{z} - \bar{M})}{\|\bar{z} - \bar{M}\|^3} \right]. \quad (7)$$

In the same time–space reference frame, the CR3BP_{SE} is described by

$$\frac{d^2 \bar{z}}{d\tau^2} + 2iw_1 \frac{d\bar{z}}{d\tau} - w_1^2 \bar{z} = -G \left[\frac{M_s(\bar{z} - \bar{S})}{\|\bar{z} - \bar{S}\|^3} + \frac{M_b(\bar{z} - \bar{B})}{\|\bar{z} - \bar{B}\|^3} \right].$$

It follows the difference between the two models

$$\begin{aligned} \Delta_{SE}(\bar{z}) &= \|BCP - CR3BP_{SE}\| \\ &= G \left\| -\frac{M_e(\bar{z} - \bar{E})}{\|\bar{z} - \bar{E}\|^3} - \frac{M_m(\bar{z} - \bar{M})}{\|\bar{z} - \bar{M}\|^3} + \frac{M_b(\bar{z} - \bar{B})}{\|\bar{z} - \bar{B}\|^3} \right\|. \end{aligned}$$

The gap Δ_{SE} arises because in the restricted three-body problem the Earth–Moon system is considered as a unique body concentrated in its center of mass rather than a binary system.

The distance between the two systems rapidly decreases to zero as the evaluation point is out of two disks around the primaries. For any different mutual position of the three primaries, the picture of Δ_{SE} is different but self-similar up to rotation around the point B ; in Fig. 2, the value of Δ_{SE} is plotted for $\theta = 0$.

3.2 Comparison with CR3BP_{EM}

Following the same procedure as before, the distance between the CR3BP_{EM} and the BCP is achieved. Again, let $\bar{z} = z_m L_M$ be used to denote the complex coordinates in a rotating reference frame and dimensional units of measure. Reminding that the origin of the Earth–Moon rotating frame is placed in the barycenter B that is revolving around the center of mass O , the inertial coordinate Z and \bar{z} are linked by the formula

$$Z = B + \bar{z}e^{i\gamma_2}, \quad B = (1 - \mu_s)L_S e^{i\gamma_1}.$$

The positions of the primaries

$$\bar{S} = (S - B)e^{-i\gamma_2} = -L_S e^{i(\gamma_1 - \gamma_2)}$$

$$\bar{E} = (E - B)e^{-i\gamma_2} = -\mu_m L_M$$

$$\bar{M} = (M - B)e^{-i\gamma_2} = (1 - \mu_m)L_M$$

and the acceleration of the particle

$$\frac{d^2 Z}{d\tau^2} = \left(\frac{d^2 \bar{z}}{d\tau^2} + 2iw_2 \frac{d\bar{z}}{d\tau} - w_2^2 \bar{z} - w_1^2 (1 - \mu_s)L_S e^{i(\gamma_1 - \gamma_2)} \right) e^{i\gamma_2}$$

yield the differential equation for the BCP in dimensional EM-synodical coordinates

$$\begin{aligned} & \frac{d^2 \bar{z}}{d\tau^2} + 2iw_2 \frac{d\bar{z}}{d\tau} - w_2^2 \bar{z} - w_1^2 (1 - \mu_s)L_S e^{i(\gamma_1 - \gamma_2)} \\ &= -G \left[\frac{M_s(\bar{z} - \bar{S})}{\|\bar{z} - \bar{S}\|^3} + \frac{M_e(\bar{z} - \bar{E})}{\|\bar{z} - \bar{E}\|^3} + \frac{M_m(\bar{z} - \bar{M})}{\|\bar{z} - \bar{M}\|^3} \right]. \end{aligned} \quad (8)$$

The term $-w_1^2(1 - \mu_s)L_S e^{i(\gamma_1 - \gamma_2)}$ represents the centrifugal acceleration of B or, equivalently, the gravitational influence of the Sun on the Earth–Moon barycenter, indeed (2) and (6) imply $(1 - \mu_s)w_1^2 = \frac{GM_s}{L_S^3}$. The difference between (8) and

$$\frac{d^2 \bar{z}}{d\tau^2} + 2iw_2 \frac{d\bar{z}}{d\tau} - w_2^2 \bar{z} = -G \left[\frac{M_e(\bar{z} - \bar{E})}{\|\bar{z} - \bar{E}\|^3} + \frac{M_m(\bar{z} - \bar{M})}{\|\bar{z} - \bar{M}\|^3} \right]$$

that describes the motion in the Earth–Moon restricted problem, gives the distance between the two models

$$\Delta_{EM}(\bar{z}) = \|BCP - CR3BP_{EM}\| = GM_s \left\| \frac{(\bar{S} - \bar{z})}{\|\bar{z} - \bar{S}\|^3} - \frac{(\bar{S} - \bar{B})}{\|\bar{S} - \bar{B}\|^3} \right\|.$$

The error originates because in the CR3BP_{EM} the influence of the Sun on the spacecraft is considered as the same influence that the Sun produces on the center B

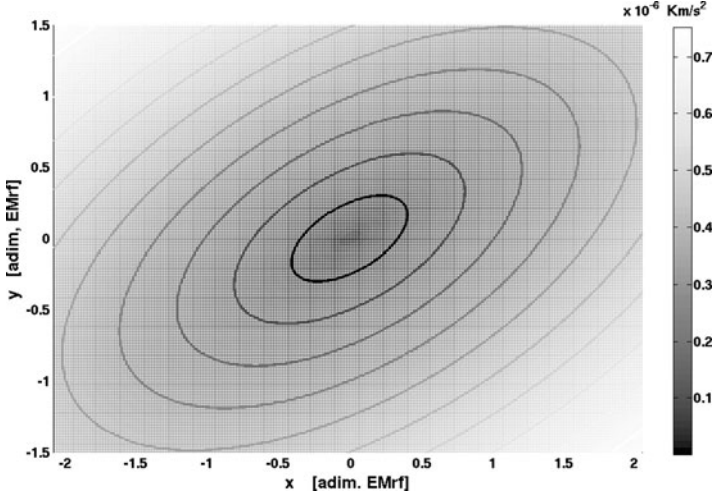


Fig. 3 Level curves of Δ_{EM} for $\theta = \pi/3$

of the rotating frame. Indeed the error vanishes whenever the spacecraft is placed in the origin of the reference frame and grows when it moves away, see Fig. 3.

4 Regions of Prevalence

The prevalence of each CR3BP is investigated according to which one produces the lowest error if it is considered in place of the BCP. To this aim, once a system of coordinates is chosen, let be defined the function

$$\begin{aligned} \Delta E(z) &= (\Delta_{SE} - \Delta_{EM})(z) \\ &= G \left\| -\frac{M_e(z-E)}{\|z-E\|^3} - \frac{M_m(z-M)}{\|z-M\|^3} + \frac{M_b(z-B)}{\|z-B\|^3} \right\| \\ &\quad - GM_s \left\| \frac{(S-z)}{\|z-S\|^3} - \frac{(S-B)}{\|S-B\|^3} \right\|. \end{aligned}$$

In any point z , one of the restricted models has to be preferred according with the sign of ΔE : where $\Delta E < 0$ the CR3BP_{SE} provides a better approximation of the BCP, otherwise the CR3BP_{EM}.

For a given relative phase θ of the primaries, denote with $\Gamma(\theta)$ the zero level set of the function ΔE

$$\Gamma(\theta) := \{z : \Delta E(z) = 0\}.$$

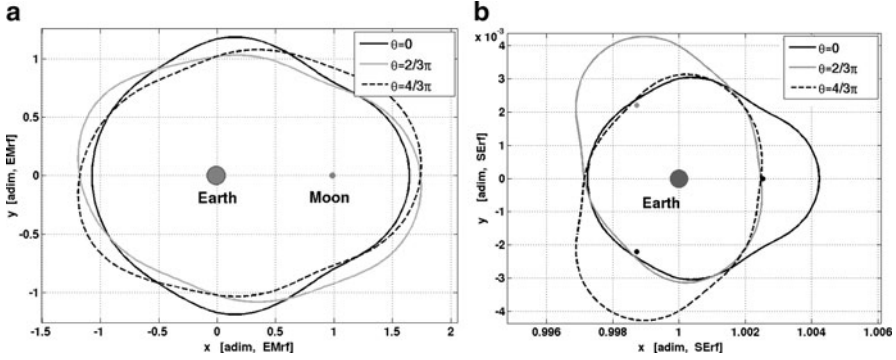


Fig. 4 $\Gamma(\theta)$ with $\theta = 0, 2/3\pi, 4/3\pi$ in EM and SE reference frame

Numerical simulations show that $\Gamma(\theta)$ is a closed simple curve: we refer to the two regions bounded by $\Gamma(\theta)$ as the *Regions of Prevalence* of the two restricted problems. In the bounded region $\Delta_{EM} < \Delta_{SE}$, while in the exterior region the opposite holds.

In Fig. 4, the zero level set of ΔE is drawn for different choices of the angle θ and in different system of coordinates. For any angle θ the Earth, the Moon as like as the L_1 and L_2 Lagrangian points related to the $CR3BP_{EM}$ belong to the EM region of prevalence, while the $CR3BP_{SE}$ Lagrangian points are placed in the exterior region.

5 The Coupled CR3BP Approximation

The Coupled CR3BP concerns the approximation of the four-body problem with the superposition of two Circular restricted three-body problems, [12]. Here the design of trajectories leaving a Lyapunov orbit around L_1 and L_2 in the $CR3BP_{SE}$ and directed to the vicinity of the Moon is considered: therefore, denoting with $W_{(SE),EM,i}^{(u),s}(\gamma)$ any (un)-stable manifold related to Lyapunov orbits γ around L_i in the (SE) or EM restricted problem, the intersections of $W_{EM,2}^s(\gamma_1)$ with $W_{SE,1}^u(\gamma_2)$ and $W_{SE,2}^u(\gamma_2)$ have been exploited. According with the prevalence regions previously defined, the Poincaré section is set on the curve $\Gamma(\theta)$. The procedure to design the transfer trajectories is the following: first the angle θ is chosen and the curve $\Gamma(\theta)$ in both the synodical systems is set. After, for a couple of Lyapunov orbits γ_1, γ_2 , the manifolds $W_{EM,2}^s(\gamma_1)$ and $W_{SE,1,2}^u(\gamma_2)$ are computed, each in their own coordinate frame, until the corresponding curve $\Gamma(\theta)$ is encountered. The resulting Poincaré map are then transformed into the same coordinates system and the analysis of their intersection leads to the selection on the point *Int* where the two solution arcs need to be connected.

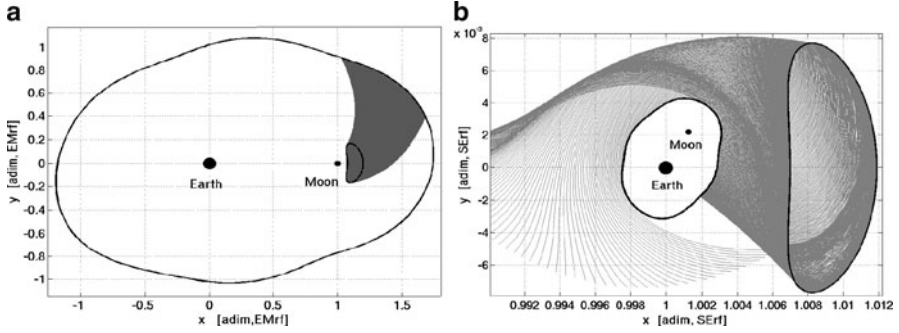


Fig. 5 Intersection of $W_{SE,2}^u(\gamma_2)$ and $W_{EM,2}^s(\gamma_1)$ with $\Gamma(\pi/3)$

As shown in Fig. 5a, for almost every Lyapunov orbits around L_2 in the EM system and every θ , the external stable manifold $W_{EM,2}^s(\gamma_1)$ invests completely the curve $\Gamma(\theta)$: the resulting Poincaré map, topologically equivalent to a circle, bounds the region \mathcal{B} of those initial data leading to the Moon's region. On the other side, Fig. 5b, depending on the angle θ and on the selected Jacoby constant, the unstable manifolds $W_{SE,1,2}^u(\gamma_2)$ may intersect only partially the curve $\Gamma(\theta)$.

For our purpose, the point *Int* has to be selected in the set $\mathcal{B} \cap W_{SE,1,2}^u$. Indeed, patching together the trajectories obtained integrating the point *Int* backward in time in the CR3BP_{SE} and forward in the CR3BP_{EM}, it follows an orbit that, starting from the SE-Lyapunov, after have passed through the EM Lyapunov gateway, will travel close to the Moon.

6 The Box-Covering Approach

In this section, the numerical technique used to detect the connecting point *Int* is discussed.

Once the intersection of the stable manifold relative to a Lyapunov orbit in the EM system with the curve $\Gamma(\theta)$ is computed, the four-dimensional Poincaré map is covered with box structures implemented in the software package GAIO (Global Analysis of Invariant Objects), see [5].

An N -dimensional box $B(C, R)$ is defined as a multidimensional rectangle and it is identified by a center $C = (C_1, \dots, C_N)$ and a vector of radii $R = (r_1, \dots, r_N)$:

$$B(C, R) = \bigcap_{i=1}^N \{(x_1, x_2, \dots, x_N) \in \mathcal{R}^N : |x_i - C_i| \leq r_i\}.$$

Following a multiple subdivision process, starting from a box B_0 , a larger family of smaller boxes is created with the property to cover B_0 : the depth d of a family

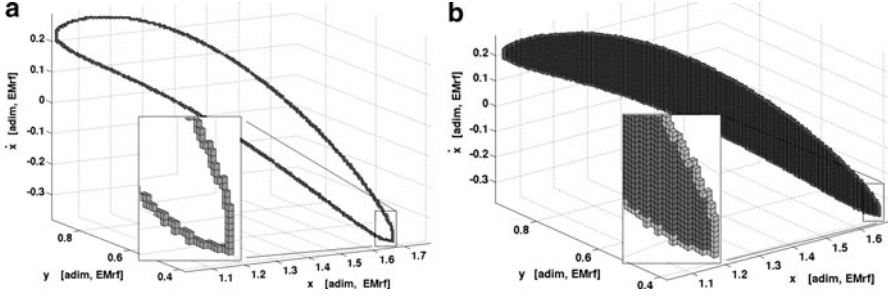


Fig. 6 Box covering of the Poincaré map of $W_{EM,2}^s$ and of the transfer region \mathcal{B}

denotes the number of such subdivision iterations. Once the family $\mathcal{F}(d)$ of boxes is created, the Poincaré map is therein inserted: only those boxes of $\mathcal{F}(d)$ containing at least one point of the Poincaré map are considered, the others are neglected, Fig. 6a. Denote with \mathcal{P} the family of boxes used for the covering of the Poincaré map. In order to detect the points Int , the interior region \mathcal{B} needs to be covered as well, see Fig. 6b. The definition of the centres of the boxes used to cover \mathcal{B} is made “by columns”: from the set of boxes in \mathcal{P} whose centres have the same (x, y) -coordinates, let be selected the two boxes with the maximal \dot{x}_{\max} and minimal \dot{x}_{\min} value of the \dot{x} -coordinate. Then a new set of centres $\{C_k = (x, y, \dot{x}_k, \dot{y}_k)\}_{k=1}^K$ are defined, where $\dot{x}_k = \dot{x}_{\min} + k\Delta v$ and \dot{y}_k is obtained from the Jacobi constant. Here Δv is twice the radius in the \dot{x} -direction of the boxes in $\mathcal{F}(d)$ and $K = (v_{\max} - v_{\min})/\Delta v$.

In the presented simulations, the covering is performed at $d = 32$: depending on the size of the Poincaré map the radii of the covering boxes result in the range $[4 \times 10^{-4}, 2 \times 10^{-3}]$ EM units.

Then, for a value of the Jacobi integral in the SE system, the Poincaré map of $W_{SE,1}^u(\gamma_2)$ or $W_{SE,2}^u(\gamma_2)$ is computed and, using (4), it is transformed in EM synodical coordinates, being θ the angle between the primaries. Finally, all those points of the SE Poincaré map lying in one of the boxes covering \mathcal{B} are considered as transfer points.

7 Some Results

The existence of connection points is tested starting from a database of 60 Lyapunov orbits in the CR3BP_{SE} both around L_1 and L_2 and 60 Lyapunov orbits around L_2 in the CR3BP_{EM}. The Jacobi constant varies in the range $[3.0004, 3.00084]$ for the SE system and in the interval $[3.053, 3.177]$ for the EM system and 32 equispaced values of θ have been considered.

Figure 7 concerns transfers leaving a Lyapunov orbit around L_1 in the Sun–Earth system: every dark sign marks a point in the intersection $\mathcal{B} \cap W_{SE,1}^u$, i.e. $\Delta v = 0$ connections. The coordinates represent the Jacobi constant of the connection point

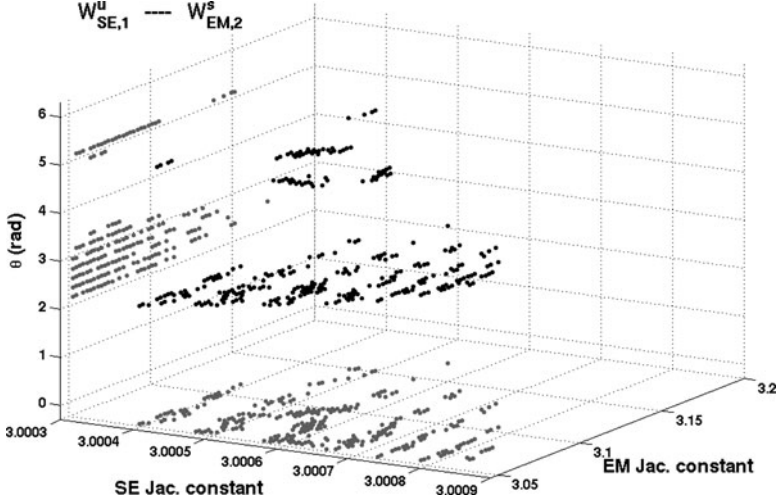


Fig. 7 Zero ΔV connections between $W_{SE,1}^u$ and $W_{EM,2}^s$

respectively in the SE and EM system and the angle θ of the Poincaré section $\Gamma(\theta)$ where the connection is detected. The lighter points are the projections of the previous ones onto the coordinates planes. Starting from one intersection, backward and forward integration in the two CR3BPs produce the complete transfer. If no differently specified, all the evaluations are done in the SE-synodical reference frame and in SE-units of measure. In the following figures, the darker and the lighter lines concern the pieces of trajectory integrated in the $CR3BP_{SE}$ and in the $CR3BP_{EM}$ respectively. In Fig. 8, the bigger picture depicts the orbit from a Lyapunov orbit around L_2 to the Moon region, while the smaller ones show the values of $\Delta_{SE}(t)$ and $\Delta_{EM}(t)$ evaluated along the trajectory. The integral $Total \Delta V = \int_{t_0}^{t_c} \Delta_{SE}(t) dt + \int_{t_c}^{t_{fin}} \Delta_{EM}(t) dt$ is used as a measure of the overall distance between the Coupled CR3BP approximation and the Bicircular model along a selected trajectory. Here t_0 is the last time when the spacecraft is far from the Earth more than 2.5 times the Earth–Moon distance and t_{fin} is the first moment the spacecraft is 10,000 km close to the Moon, while t_c denotes the instant when the Poincaré section is crossed. Referring to Fig. 9, in the bigger figure the dotted line remarks the circle inside which the above integration starts, the small black circles show the position of the Moon when the spacecraft is on the section and at the end of the travel, the black line denotes the Poincaré section at the crossing time. In the upper of the smaller boxes the values of Δ_{EM} and Δ_{SE} are plotted together, while the last graph shows the value of the integral $Total \Delta V$. Starting from t_0 the error Δ_{SE} is integrated up to the crossing moment t_c , then two different integrations are done: in the first case the error Δ_{EM} is considered till the final time t_{fin} (lighter line), in the second case again Δ_{SE} is integrated for a short interval of time (darker line). The last figure depicts

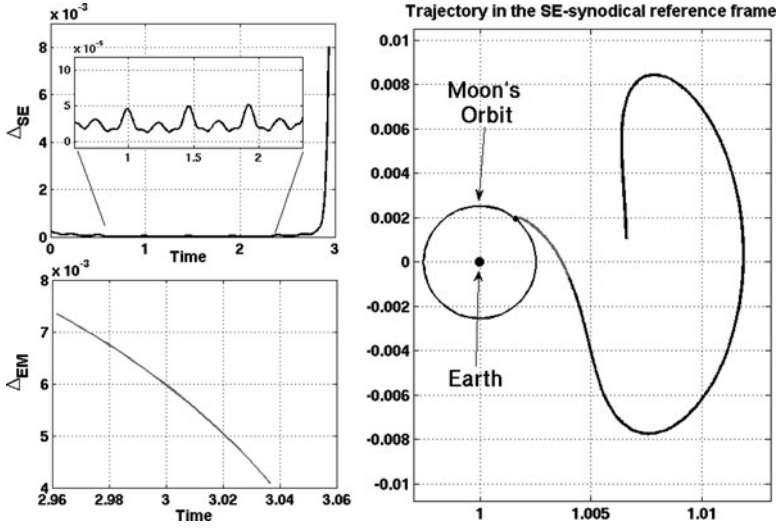


Fig. 8 Example of transfer trajectory and related errors Δ_{SE} , Δ_{EM}

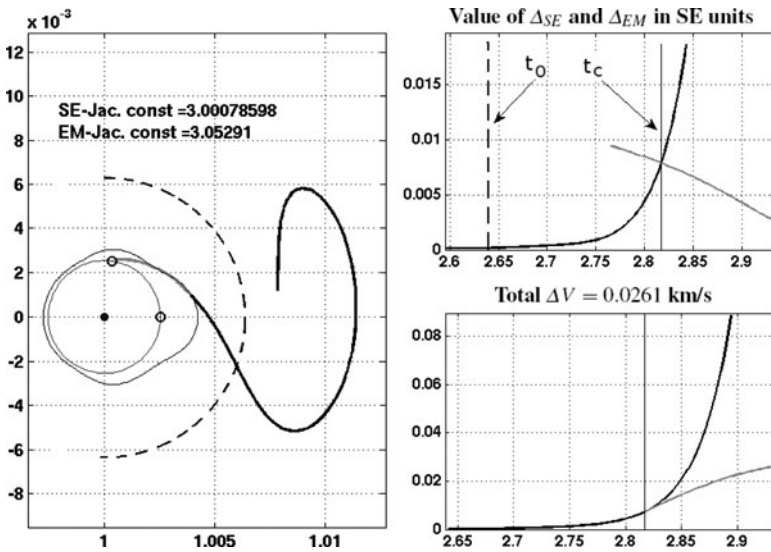


Fig. 9 Example of trajectory. Analysis of the error

the optimality of the choice of the Poincaré section $\Gamma(\theta)$ in minimizing the local approximation error.

A more detailed analysis and a comparison with trajectories designed adopting different Poincaré sections will be discussed in a future work.

References

1. Belbruno, E., Miller J.: Sun-perturbed Earth-to-Moon transfer with ballistic capture. *Journal of Guidance Control and Dynamics*, **16**, 1993.
2. Belbruno, E.A.: The dynamical mechanism of ballistic lunar capture transfers in the four body problem from the perspective of invariant manifolds and Hills regions. Institut d'Estudis Catalans, CRM, Preprint No. 270, 1994.
3. Belbruno, E.: *Capture Dynamics and Chaotic Motions in Celestial Mechanics: With Applications to the Construction of Low Energy Transfers*. Princeton Univ. Press, 2004.
4. Cronin, J., Richards, P.B., Russell, L.H.: Some periodic solutions of a four-body problem. *Icarus*, **3**(5-6) 423–428 1964.
5. Dellnitz, M., Froyland, G., Junge, O.: The algorithms behind GAIO - Set oriented numerical methods for dynamical systems. In: *Ergodic theory, analysis, and efficient simulation of dynamical systems*, pages 145–174. Springer, 2000.
6. Dellnitz, M., Junge, O., Post, M., Thiere, B.: On target for Venus – Set oriented computation of energy efficient low thrust trajectories. *Celestial Mech. Dynam. Astronom.*, **95**(1-4), 357–370, 2006.
7. Demeyer, J.R.C., and Gurfil, P.: Transfer from Earth to Distant Retrograde Orbits using Manifold Theory, *AIAA Journal of Guidance, Control and Dynamics*, **30**(5), 1261–1267, 2007.
8. F. Garca, F., Gómez, G.: A note on weak stability boundaries. *Celestial Mech. Dynam. Astronom.*, **97**(2), 87–100, 2007.
9. Gómez, G., Jorba, A., Masdemont, J., Simo, C.: Study of the transfer from the Earth to a halo orbit around the equilibrium point L1. *Celestial Mech. Dynam. Astronom.*, **56**(4), 541–562, 1993.
10. Gómez, G., Koon, W.S., Lo, M.W., Marsden, J.E., Masdemont, J., Ross, S.D.: Connecting orbits and invariant manifolds in the spatial restricted three-body problem. *Nonlinearity*, **17** 1571–1606, 2004.
11. Koon, W. S., Lo, M. W., Marsden, J. E., Jerrold, E., Ross, S. D.: Heteroclinic connections between periodic orbits and resonance transitions in celestial mechanics. *Chaos*, **10**(2), 427–469, 2000.
12. Koon, W. S., Lo, M. W., Marsden, J. E., Ross, S. D.: Low energy transfer to the Moon. *Celestial Mech. Dynam. Astronom.*, **81**(1-2), 63–73, 2001.
13. Lo, M. W., Williams, B.G., Bollman, W.E., Han, D., Hahn, Y., Bell, J.L., Hirst, E.A., RCorwin, R.A., Hong, P. E., Howell, K. C., Barden, B., Wilson R.: Genesis Mission Design, *The Journal of the Astronautical Sciences* 49, 169–184, 2001.
14. Mingotti G., Topputo F., Bernelli-Zazzera F.: Low-energy, low-thrust transfers to the Moon. *Celestial Mech. Dynam. Astronom.*, **105**(1-3) 61–74, 2009.
15. Parker, J.S.: Families of low-energy Lunar halo transfer. *Proceedings of the AAS/AIAA Space Flight Mechanics Meeting*, 483–502, 2006.
16. Simó, C., Gómez, G., Jorba, À., Masdemont, J.: The Bicircular model near the triangular libration points of the RTBP. In: *From Newton to chaos*, volume 336 of NATO Adv. Sci. Inst. Ser. B Phys., pages 343–370. 1995.
17. Szebehely, V.: *Theory of orbits, the restricted problem of three bodies*. Academic Press, New York and London, 1967.
18. Yagasaki, K.: Sun-perturbed Earth-to-Moon transfers with low energy and moderate flight time. *Celestial Mech. Dynam. Astronom.*, **90**(3-4): 197–212, 2004.

Adaptive Remeshing Applied to Reconfiguration of Spacecraft Formations

Laura Garcia-Taberner and Josep J. Masdemont

1 Introduction

Formation flying of spacecraft is a concept that has an important role in technology applications and continues growing in importance, mainly for science and astrophysical missions. The reason is that formation flying enables a set of some small (and cheaper) spacecraft to act as a virtual larger satellite, obtaining better information than a bigger one, with flexibility about the space observations that a formation can perform in the future.

Some formation flying missions are planed to be in orbits about the Earth, but also relevant science missions have big interest on the vicinity of libration points. The L_1 libration Sun–Earth+Moon point is nowadays the best place for observations of the Sun. The L_2 libration point is also an interesting place for deep space observations, where large telescopes or interferometry baselines could be located. This is the reason because projects like the Terrestrial Planet Finder [11] or Darwin [10] were planned to be in a libration point orbit about L_2 .

The key technology that must be implemented in spacecraft formation flying is the control of a formation when doing observations. Mutual distances between spacecraft must be kept with high precision. A study of some of these control methodologies can be found in Farrar, Thein and Folta [1] and references therein.

But also there are many other technologies that have to be taken into account for formation flying. The one we consider in this paper is the reconfiguration of the formation. This technique is important in the lapses between observations. It can be necessary to change the orientation, the target point of the formation, or maybe also the shape or diameter of the cluster to give flexibility to the mission.

L. Garcia-Taberner (✉)

Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona,
Escola Politècnica Superior, 17071 Girona, Spain
e-mail: laura.garcia@ima.udg.edu

Some representative techniques of reconfigurations are the proximity maneuvering using artificial potential functions studied by McInnes [8], or the technique used by Hadaegh, Beard, Wang and McLain [6].

In this paper, we consider the reconfiguration of a formation by means of an optimization problem which uses the finite element methodology to obtain the controls that must be applied to each spacecraft. Additionally, our methodology can solve the problem of the transfer or deployment of the formation, which is another key problem in formation flying.

In this paper, we mainly focus on the obtention of a suitable optimal mesh using an adaptive remeshing strategy, that assures us that the error produced by the finite element methodology is small enough. Once the methodology gives us the trajectory and an optimal mesh, we check it considering a vector field of full JPL-ephemeris.

2 The FEFF Methodology

The main purpose of the FEFF methodology is to compute reconfigurations of a formation of spacecraft. In this paper, we consider that the spacecraft are in the vicinity of a halo orbit of 120,000 km of z -amplitude about L_2 in the Sun–Earth system, but the methodology can be applied to any other libration point orbit.

Here we present just a brief description of the basic FEFF methodology. More details about this point can be found in [2, 4].

We consider a formation of N spacecraft which must perform a reconfiguration in a fixed time T . The spacecraft are in a small formation (i.e., the distance between them is only of a few hundreds of meters, both in the initial and the final configurations). Our objective is to find a trajectory for each of the spacecraft which guides it to the goal position, with the minimum fuel consumption and avoiding collisions with other spacecraft.

As the formations are small with respect to the amplitude of the halo orbit, we consider the linearized equations of motion about the nonlinear orbit. We compute the trajectories using these equations and then we will deal with the nonlinear part. For each of the spacecraft, we have the equation

$$\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t), \quad (1)$$

where $\mathbf{A}(t)$ is a 6×6 matrix and \mathbf{X} refers to the state of the spacecraft. The origin of the reference frame for the \mathbf{X} coordinates is the nominal point on the base halo orbit at time t being the orientation of the coordinate axis parallel to the one of the RTBP.

The goal of the methodology is to find a set of optimal controls for each of the spacecraft. Including the initial and final states of the spacecraft in the reconfiguration problem and the controls, the equations that we deal with are

$$\begin{cases} \dot{\mathbf{X}}_i(t) = \mathbf{A}(t)\mathbf{X}_i(t) + \bar{\mathbf{U}}_i(t), \\ \mathbf{X}_i(0) = \mathbf{X}_i^0, \\ \mathbf{X}_i(T) = \mathbf{X}_i^T, \end{cases} \quad (2)$$

where \mathbf{X}_i^0 and \mathbf{X}_i^T stand for the initial and final state of the i -th spacecraft of the formation, and $\bar{\mathbf{U}}_1, \dots, \bar{\mathbf{U}}_N$, are the controls we are searching.

The key of procedure FEF is that it uses the finite element methodology to obtain these controls (see [9] for references about the finite element method and [4] for a more detailed exposition about the FEF methodology). Essentially, the time interval $[0, T]$ which we consider for the reconfiguration is split in M subintervals of the domain that we call elements. This mesh can have elements of different length and can be different for each of the spacecraft, depending on the nature of the trajectories of reconfiguration. We impose that controls are some maneuvers (in form of delta-v) that we apply in the points where two elements join (the nodes). The finite element methodology gives us a relation between the positions of the spacecraft in the nodes and the maneuvers, $\Delta \mathbf{v}$, that we must apply.

Procedure FEF reduces the reconfiguration problem to an optimization problem with constraints. The functional that must be minimized has to be related to fuel consumption, and since it is related to the sum of the norm of the delta-v, the functional we want to minimize is

$$J_1 = \sum_{i=1}^N \sum_{k=0}^{M_i} \rho_{i,k} \|\Delta \mathbf{v}_{i,k}\|, \quad (3)$$

where $\|\cdot\|$ denote the Euclidean norm and $\rho_{i,k}$ are weight parameters that can be used, for instance, to penalize the fuel consumption of selected spacecraft with the purpose of balancing fuel resources (here for clarity we consider that $\rho_{i,k}$ multiplies the modulus of the delta-v, but in a similar way we can impose a weight on each component).

The most important constraint in our problem is the avoidance of collision between spacecraft. This enters in the optimization problem as a constraint, imposing that each spacecraft is surrounded by a security sphere and the spheres of all the spacecraft cannot collide. We note that the partition of the time interval made by the finite element methodology gives us an efficient implementation to check this constraint.

3 Adaptive Remeshing Applied to Reconfigurations

When searching the optimal controls for the spacecraft, we must take into account two facts. One of them is related to ill conditioning problems and the other is related to the error due to the finite element approximation. Both of them lead us to think about a remeshing strategy.

When we consider the functional of equation (3), we see that it is ill conditioned to compute derivatives when delta-v values are small. But our objective is to find

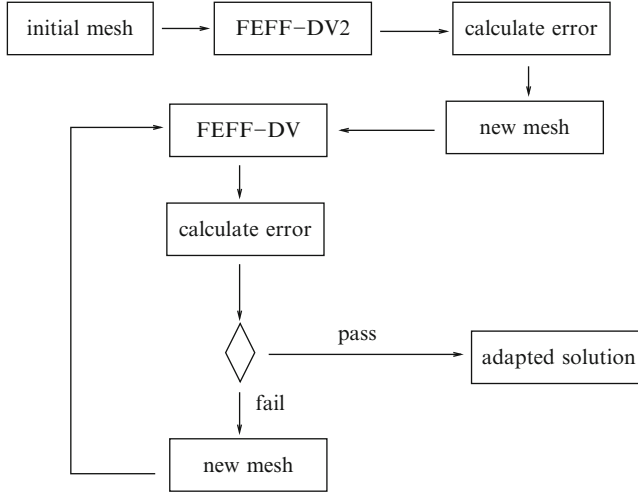


Fig. 1 Schema of the procedure of adaptive remeshing

small delta-v! In order to avoid this problem, we have two strategies: the first one consists on considering an alternative functional to minimize,

$$J_2 = \sum_{i=1}^N \sum_{k=0}^{M_i} \rho_{i,k} \|\Delta \mathbf{v}_{i,k}\|^2, \quad (4)$$

which is also related to fuel consumption and it is not ill conditioned. The second one is based on considering a remeshing strategy to suppress the nodes with a small delta-v.

On the other hand, the approximation of the solution via the finite element method gives us some errors associated with the approximation we make on each one of the elements of the mesh.

To solve these two facts, we will consider an adaptive remeshing strategy applied to our reconfiguration problem. The general idea of adaptive remeshing is that, given a threshold value e , to find a mesh that provides an approximate solution with error (understood as the difference between the solution of the problem and its approximation inside of an element) less than e in some norm.

In Fig. 1, we give a schema of the general idea of our procedure. It has two different phases. In the first one, it starts computing the trajectories for the spacecraft using the finite element method with a given mesh. In this first phase the objective is to find a rough approximation of the solution, so we start with a small number of elements (a maximum of 10). All of these elements have the same length and all the spacecraft start with the same mesh. Once we have obtained the trajectories, it computes an estimation of the error. This error is essentially obtained by comparison between the gradient obtained using the finite element model and the one obtained

by integration of the equations of motion. The second step of the procedure is an iterative process which performs adaptive remeshing, recomputes the approximate solution with the new mesh, and ends when the error is below a given tolerance. When remeshing is necessary, the new mesh is adapted using the estimation of errors of the previous mesh.

Adaptive remeshing methods penalize the elements where the error is considered big, dividing them in smaller elements. On the other hand, if the estimation of the error is small in an element, then this element is made bigger in the next iteration. Since, essentially our estimation of the error is related to the value of the delta-v maneuvers to be implemented, this method tends to increase the length of the elements which have associated small delta-v and tends to decrease the length of the elements which have associated big delta-v's.

Essentially to decide whether the current mesh is good enough or not we use a criterion which compares the modulus of the estimated error ($||e||$) with the total gradient of the solution,

$$||\bar{u}|| = \int_0^T \mathbf{v}_2 dt.$$

We accept the mesh when

$$||e|| \leq v ||\bar{u}||,$$

where v is the acceptability criteria. We discuss the value of v taken on Sect. 4.

In order to compute the new mesh, we use the Li and Bettess remeshing strategy (see [7]). This strategy is based on the idea that the error distribution on an optimal mesh is uniform,

$$||\hat{e}_k|| = v ||\bar{u}|| / \sqrt{\hat{M}},$$

where v is again the acceptability criteria, e_k is the computed error on element k , M is the number of elements of the mesh and the hat distinguishes the parameters of the new mesh. The strategy consists of finding the new length of the elements using the number of elements of the new mesh, \hat{M} . Let us denote d , the dimension of the problem and m , the maximum degree of the polynomials used in the interpolation. Then, according to Li and Bettess, the number of elements needed by the new mesh is,

$$\hat{M} = (v ||\bar{u}||)^{-d/m} \left(\sum_{k=1}^M ||e_k||^{d/(m+d/2)} \right)^{(m+d/2)/m}.$$

Since we work with linear elements in dimension one, we have $m = 1$ and $d = 1$. The recommended number of elements of the new mesh is

$$\hat{M} = (v ||\bar{u}||)^{-1} \left(\sum_{k=1}^M ||e_k||^{2/3} \right)^{3/2}.$$

Once we have the estimation of the number of elements, we can find the length of the new elements:

$$\hat{h}_k = \left(\frac{v||\bar{u}||}{\sqrt{\hat{M}}||e_k||} \right)^{1/m+d/2} h_k,$$

that in our case, turns out to be

$$\hat{h}_k = \left(\frac{v||\bar{u}||}{\sqrt{\hat{M}}||e_k||} \right)^{3/2} h_k.$$

4 Simulations with Adaptive Remeshing

When computing reconfigurations of spacecraft, we have two limiting cases: if there is no collision risk when the spacecraft follow a linear trajectory, we know that the optimal trajectory is a bang–bang control for each one of the spacecraft. This is the most critical case for our procedure, since the optimal maneuver consists in two delta-v: one at departure and another one at the arrival position. The remaining nodal delta-v must be zero, and so this is a case where the computation of derivatives for J_1 is very ill conditioned. On the other hand, for cases with collision risk, our methodology must tend to low thrust when the diameter of the mesh tends to zero. In this section, we present an example of each of these two cases.

4.1 A Bang–Bang Example

In this kind of problem, since there are no collision hazards, collision avoidance does not affect the trajectories of the spacecraft and the optimal trajectory for each spacecraft is independent from the others. For this reason, we can reduce the computations to obtain the optimal trajectory for a single spacecraft.

In order to exemplify the procedure, we consider a shift of a single spacecraft. We consider the reference frame for (2) aligned with respect to the RTBP reference frame, but with origin on the nominal point of the base halo orbit (when $t = 0$ this point corresponds to the “upper” position of the Halo orbit, this is when it crosses the RTBP plane $Y = 0$ with $Z > 0$). The initial condition for this example is taken 100 m far from the base nominal halo orbit in the X direction, and the goal is to transfer it to a symmetrical position with respect to the halo orbit in 8 h. This is to 100 m in the opposite X direction doing a shift of 200 m for the spacecraft transfer maneuver.

For this particular case, we know that the optimal solution is a bang–bang control with maneuvers of 0.69 cm/s at departure and arrival.

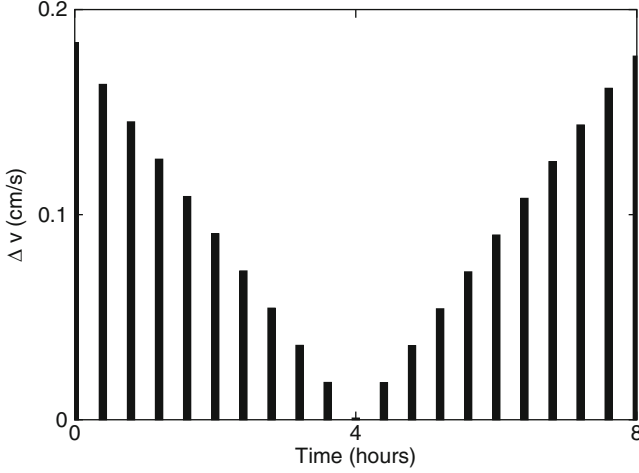


Fig. 2 Delta- v obtained with the minimization of functional (4) in the case of no collision risk. The methodology FEFF converges to a bang–bang solution with this initial seed

Our procedure starts by minimizing the functional (4) and obtains a trajectory with the delta- v profile of Fig. 2. This is a typical profile result for the reconfigurations without collision risk.

We compute now the reconfiguration considering different values for the parameter v . We note that this parameter does not only appear in the acceptability criteria but it is also used to obtain the new mesh. If we take a small value of v , we can end up with a mesh with a big number of nodes, which results in an optimization problem with a very large number of variables, that could be unsolvable in practice. Note that a mesh with 100 elements and a single spacecraft ends up with an optimization problem with 594 variables. In the other way around, if we use a big v , we could end up accepting some meshes with big errors. In Table 1, we have a summary of the results obtained for different values of the parameter v , the number of iterations needed to reach the bang–bang solution and the number of elements after the first iteration of the methodology.

We note that when v is very small, there is no convergence. The case with v equal to 0.0001 makes the optimal procedure awkward. When v is 0.001, the final number of elements is greater than 1 (that we know is our final target number) although there are some very small delta- v . When v is big, moreover, there is no convergence; the final mesh contains more elements than expected, because it passes the acceptability criteria before converging to the bang–bang control. We can conclude that, in this bang–bang case, the best values for v are inside the range $[0.04, 0.06]$. With values larger than 0.06, the algorithm does not converge.

Table 1 Number of iterations necessary to obtain the bang–bang solution depending on ν . We have indicated by “fail” the cases where the procedure does not converge

ν	Elements iteration 1	Iterations
0.0001	3,008	Fail
0.001	301	Fail
0.002	149	25
0.005	61	16
0.01	31	14
0.02	15	10
0.03	11	6
0.04	7	4
0.05	6	4
0.06	4	2
0.07	4	Fail

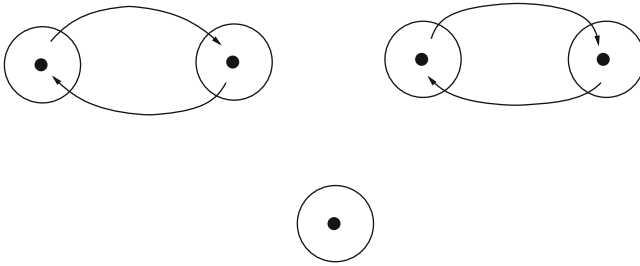


Fig. 3 Example of reconfiguration with collision risk: the switch of two pairs of spacecraft of the TPF formation

4.2 A Low-Thrust Example

In reconfigurations where the bang–bang trajectories end up with collisions with some of the spacecraft, FEFF obtains trajectories which of course are different from bang–bang. Our objective is to study whether these trajectories could tend to low thrust arcs.

For this case we consider a configuration based on the Terrestrial Planet Finder (TPF) model (see [11]). We assume that the satellites are initially contained in the local plane $Z = 0$, with the interferometry baseline aligned on the X axis. We simulate the swap between two pairs of satellites in the baseline: each inner satellite changes its location with the outer satellite which is closest to it in position (this is, inner satellites are maneuvered to attain outer positions and vice-versa as shown in Fig. 3). Again we consider 8 h for the reconfiguration. The process of switching positions has a collision risk and simple bang–bang controls are no longer valid.

As in the bang–bang case, we consider different values for the parameter ν . A discussion like in the previous section is also valid here: using a small ν , we can end up with a mesh with many elements. For example, taking $\nu = 0.0005$, in the first iteration we have around 1,000 elements. We do not only have the problem

Table 2 Number of iterations and elements obtained with the swapping example of TPF depending on ν

ν	N_1	Iteration	N_F
0.0001	3,504	Fail	
0.001	350	10	232
0.002	175	8	202
0.005	70	8	171
0.01	34	7	89
0.02	18	6	45
0.03	12	4	33
0.04	9	3	27
0.05	6	3	15
0.06	6	3	9
0.07	5	2	7

of having very small elements: the optimization problem that we end up with which has 29,970 variables and it is not desirable. Again, if we take a big ν , we can end up with a mesh with big errors or with a mesh with only a few elements.

In Table 2, we display a summary of the results obtained for different values of the parameter ν , the number of iterations until the methodology converges (Iter), the number of elements in the first iteration (N_1) and the number of elements in the last iterate (N_F).

As in the previous case, when ν is small, the number of elements is big, and the computation of the optimum is very expensive. Also, taking ν big, the number of elements may not be enough.

We note that now the best values are inside the range $[0.005, 0.05]$. With values larger than 0.05, the number of elements is very small and with values smaller than 0.005 the number of elements makes the computation much more expensive.

Since the value $\nu = 0.05$ is appropriated for the two cases, we consider it for our computations.

4.3 Considerations About the Value of ν

We have seen in the previous sections that a desirable value of ν must be in the range $[0.005, 0.05]$. This range gives us an idea of the value of ν we must choose.

We have applied the reconfiguration procedure to a test bench of 25 reconfigurations, which include switches between spacecraft located at opposite vertices of polygons (6), switches in the TPF formation (9) and parallel shifts (10) of different size with a number of spacecraft from 3 to 10. Ten of the reconfigurations are converging to a bang–bang solution, the other 15 reconfigurations are converging to low thrust. We have applied the methodology using different values of ν and we have computed the mean of the number of iterations of the adaptive process necessary to converge. The results can be seen in Table 3.

Again we conclude that the best value of ν is 0.05.

Table 3 Mean of the number of iterations as a function of v for the 25 test bench reconfigurations

v	0.005	0.01	0.02	0.03	0.04	0.05	0.055	0.06
It.	10.2	8.4	7.1	4.2	3.7	3.2	4.3	5.2

5 Dealing with Nonlinearities

We have pointed out that the size of the formations is very small (a few hundreds of meters) when comparing it to the size of the halo orbit. This fact gives us the possibility to work with linearized equations about the halo orbit. In this section, we present the results of some simulations to show that the use of linearized equations is really a good model.

Our objective in this section is to give a way to measure how the truncated nonlinear terms, as well as other perturbations, affect the nominal reconfiguration trajectory, and the corrections that should be applied to the nominal maneuvers (corrective maneuvers) in order to reach the mission goal. For this purpose, we consider the trajectories given by the FEFF methodology as the nominal path for the spacecraft and compute the corrective maneuvers that guide the spacecraft through the nominal nodal states.

The corrective maneuvers are computed using a similar strategy to [5]. The main idea is that on each element we have the nodal states given by the FEFF methodology, and the difference between these states and the true states is corrected with some small maneuvers. For more details, see [3] or [4].

In the same line as in previous computations, we consider two different cases: the ones which end up in low thrust trajectories (there are many elements on the mesh, and these elements are small) and the ones that end up in bang–bang (the mesh is formed by a single element which is the same as the reconfiguration time).

5.1 The Bang–Bang Case

Like we have pointed out in the previous section, in examples where there is no collision risk, we can focus in the results of a single spacecraft, because the results are independent.

In order to test the suitability of the linear approximation, we consider the shift of a spacecraft in the x direction (200 or 400 m in 8 or 24 h). In Table 4, we give the value of the delta- v given by FEFF methodology (Δv_L), the number of corrections that are performed on each element (n), the maximum of corrective maneuvers ($\Delta \hat{v}_{L\max}$), the total amount of corrective maneuvers ($\Delta \hat{v}_{LJ}$), and the percentage of the corrective maneuvers with respect to the maneuvers given by FEFF.

We note that all these corrective maneuvers are very small, both in absolute values and in percentage.

Table 4 Corrective maneuvers for some cases of bang–bang. Model equations considered correspond to JPL ephemeris and all delta-v are given in cm/s

Case	Δv_L	n	$\Delta \hat{v}_{LJmax}$	$\Delta \hat{v}_{LJ}$	%
200 m 8 h	0.69	3	3.3×10^{-3}	6.7×10^{-3}	0.96
	0.69	4	2.5×10^{-3}	6.1×10^{-3}	0.88
	0.69	5	2.2×10^{-3}	5.8×10^{-3}	0.84
	0.69	6	2.2×10^{-3}	5.6×10^{-3}	0.80
200 m 24 h	0.23	3	3.6×10^{-3}	7.5×10^{-3}	3.25
	0.23	4	2.7×10^{-3}	6.7×10^{-3}	2.89
	0.23	5	2.2×10^{-3}	6.4×10^{-3}	2.77
	0.23	6	2.2×10^{-3}	6.1×10^{-3}	2.65
400 m 8 h	2.8	3	5.3×10^{-3}	9.7×10^{-3}	0.35
	2.8	4	3.9×10^{-3}	9.2×10^{-3}	0.33
	2.8	5	3.1×10^{-3}	8.1×10^{-3}	0.29
	2.8	6	2.7×10^{-3}	7.5×10^{-3}	0.27

Table 5 Corrective maneuvers for some cases of low thrust. Model equations considered correspond to JPL ephemeris and all delta-v are given in cm/s

Case	Δv_L	n	$(\Delta v/l)_{Jmax}$	Δv_J	%
Swap	0.63	3	9.1×10^{-3}	7.5×10^{-3}	1.19
2 sats	0.63	4	8.5×10^{-3}	6.9×10^{-3}	1.10
100 m	0.63	5	7.7×10^{-3}	6.7×10^{-3}	1.06
24 h	0.63	6	5.9×10^{-3}	6.4×10^{-3}	1.01
TPF	2.34	3	9.9×10^{-3}	1.3×10^{-2}	0.51
Swap	2.34	4	8.2×10^{-3}	1.1×10^{-2}	0.44
	2.34	5	6.4×10^{-3}	1.0×10^{-2}	0.42
	2.34	6	2.6×10^{-3}	9.2×10^{-3}	0.37
TPF	1.26	3	8.3×10^{-3}	1.0×10^{-2}	0.77
Symmetry	1.26	4	8.0×10^{-3}	9.4×10^{-3}	0.73
Baseline	1.26	5	5.9×10^{-3}	8.9×10^{-3}	0.66
	1.26	6	5.5×10^{-3}	8.1×10^{-3}	0.60

5.2 The Low Thrust Case

When we compute the corrective maneuvers with cases of low thrust we must take into account that the elements are very small, and we will do a lot of corrective maneuvers during the reconfiguration time. So, each one of these maneuvers is expected to be smaller than the ones in the case of bang–bang.

We have considered three cases of low-thrust examples, presented in Table 5. The first one is the swap of two spacecraft which are at a distance of 100 m. The reconfiguration time is 24 h. The second example is the TPF swap presented in the previous section. And the third example deals with another case based on the TPF formation: to change the position of the collector toward a point symmetric to the departure position with respect to the interferometry baseline. The parameters of the table are the same ones of the previous example, except for the maximum of

the delta- v , $(\Delta v/l)_{\text{Jmax}}$. In this case, we give the maximum of delta- v divided by the length of the element instead, this is thrust acceleration as is usual for low-thrust trajectories.

We remark that delta- v s are again small, both in absolute value and in percentage. So, we can conclude that for small formations (of a few hundreds of meters) and small reconfiguration times (of a few hours), the corrective maneuvers are small and the linear approximation about the nonlinear orbit is good enough.

6 Conclusions

This paper presents a methodology to find trajectories for reconfigurations of spacecraft, using the finite element method.

We have shown that the strategy of adaptive remeshing is suitable for the problem. It converges toward a bang–bang solution when there is no collision risk (that is known in this case to be the optimal control), and when there exists collision risk in the reconfiguration process, the procedure tends to low-thrust arcs.

The computations have been done for formations of a few hundreds of meters, with reconfiguration times of a few hours. We have shown, through some general examples, that the linearized model about the nonlinear orbit is suitable for the nominal computations.

Acknowledgments This research has been supported by the MICINN-FEDER grant, MTM2009-06973, the CUR-DIUE grant 2009SGR859, and the European Marie Curie grant Astronet, MCRTN-CT-2006-035151.

References

1. Farrar, M., Thein, M. and Folta, D. C. (2008). A Comparative Analysis of Control Techniques for Formation Flying Spacecraft in an Earth/Moon-Sun L2-Centered Lissajous Orbit. *AIAA Paper* 2008–7358, Aug. 2008.
2. Garcia-Taberner, L. and Masdemont, J. J. (2009). Maneuvering Spacecraft Formations using a Dynamically Adapted Finite Element Methodology. *Journal of Guidance, Control, and Dynamics* Vol. 32, No. 5, pp. 1585–1597.
3. Garcia-Taberner, L. and Masdemont, J. J. (2010). FEFF methodology for spacecraft formations reconfiguration in the vicinity of libration points. *Acta Astronautica*, doi:10.1016.
4. Garcia-Taberner, L. (2010). Proximity Maneuvering of Libration Point Orbit Formations Using Adapted Finite Element Methods. *PhD Dissertation* Universitat Politècnica de Catalunya. Barcelona.
5. Gómez, G., Lo, M., Masdemont, J.J. and Museth, K. (2002). Simulation of Formation Flight near Lagrange Points for the TPF Mission. *Advances in the Astronautical Sciences* Vol. 109, 61–75.
6. Wang, P.K.C. and Hadaegh, F. Y. (1999). Minimum-fuel Formation Reconfiguration of Multiple Free-flying Spacecraft. *Journal of the Astronautical Sciences* Vol. 47, No. 1,2, pp. 77–102.

7. Li, L. Y., Bettess, P., Bull, J.W., Bond, T. and Applegarth, I. (1995). Theoretical formulation for adaptive finite element computations. *Communications in Numerical Methods in Engineering* 11, 858-868.
8. Badawy, A. and McInnes, C. R. (2008). On-Orbit Assembly Using Superquadric Potential Fields. *Journal of Guidance, Control, and Dynamics* Vol. 31, No. 1, pp. 30–43.
9. Reddy, J. N. (1993). An Introduction to the Finite Element Method. McGraw-Hill, New York.
10. The Science and Technology Team of Darwin and Alcatel Study Team. (2000) Darwin. The Infrared Space Interferometer. Concept and Feasibility Study Report. *European Space Agency, ESA-SCI* 12, iii+218.
11. The TPF Science Working Group. (1999) The Terrestrial Planet Finder: A NASA Origins Program to Search for Habitable Planets. *JPL Publication* 99-3, 1999.

A Cartographic Study of the Phase Space of the Elliptic Restricted Three Body Problem: Application to the Sun–Jupiter–Asteroid System

Cătălin Galeş

1 Introduction

In the last two decades, various numerical methods have been applied to celestial mechanics to distinguish between regular and chaotic motions. Depending on the mathematical model, the information offered and the amount of computation, each method has its own advantages and disadvantages. We shall briefly recall some widely used tools in investigating the dynamics of nonintegrable systems. Thus, the frequency map analysis [22, 23] has been used for studying small body dynamics by Nesvorný and Ferraz-Mello [31] and Celletti et al. [4]; the method of twist angles introduced by Contopoulos and Voglis [6] has been tested on the standard map and compared with other methods of analysis in [9]; the method of short time average of the stretching numbers [11] has been applied to the circular and elliptic restricted three-body problem in [35]; the fast Lyapunov indicator [12, 13] has been investigated in detail and applied to various dynamical problems; the MEGNO technique [5] has been utilized to perform a stability analysis of extra solar planets in [18, 19]; the time series given by the intervals between successive crossing of a given plane of section have been used to reconstruct the phase space in the restricted three-body problem [16].

In this chapter, we use the fast Lyapunov indicator (FLI) in order to perform a cartographic study of a given portion of the phase space of the restricted three-body problem. Introduced by Froeschlé et al. [12, 13], this tool is easy to implement, cheap in computational time, and very sensitive for the detection of weak chaos and for distinguishing between regular resonant orbits and regular nonresonant ones. These features have been unveiled by testing it on symplectic mappings [10, 24] as well as on continuous flows [8]. Moreover, it was shown that FLI is a very useful numerical

C. Galeş (✉)

Faculty of Mathematics, Al. I. Cuza University of Iaşi, Romania

e-mail: cgales@uaic.ro

tool for revealing the geography of resonances (the so-called Arnold's web), for detecting the transition between the stable Nekhoroshev's regime and the diffusive Chirikov's one [14, 20, 36] and even for detection of diffusion along resonances (Arnold's diffusion) [15, 25, 26, 36]. These topics have been investigated in the framework of quasi-integrable Hamiltonian systems. In this sense, some model systems have been carefully chosen.

The FLI tool has been directly utilized to study the stability of extrasolar planets [33], to solve spacecraft preliminary trajectory design problems [38], to investigate the dynamics associated to nearly integrable dissipative systems [1].

In this chapter, we use the FLI tool to investigate the phase space of the restricted three body problem and we focus on the Sun–Jupiter–Asteroid system. We recall that FLI has been first applied to asteroidal motion. Froeschlé et al. [13] integrated a model consisting of the four giant planets and the Sun and studied the dynamics of the asteroids orbiting between the 3/1 and 5/2 Kirkwood gaps. Here, we implement the FLI tool for the elliptic restricted three-body model (the case Sun–Jupiter–infinitesimal mass) and prove that FLI reveals after a very short computational time, the entire structure of the phase space (the regular and chaotic regions, the geography of resonances, the libration regions, and other details closely connected with the degeneracy property of the restricted three-body problem).

This analysis is relevant in the study of the global dynamics in the asteroid belt. This topic, with particular attention to the mechanisms of chaotic diffusion, is extensively studied in literature (see for example the book of Morbidelli [27] and the review article by Tsiganis [37]). The theories of chaotic diffusion aim to predict the long-term behavior of ensembles of asteroids, rather than individual orbits. In this sense, the Lyapunov time, the size, and the shape of the chaotic regions and the diffusion coefficients are the main parameters needed to understand the long-term effect of chaotic diffusion in the asteroid belt. To compute them, analytical models and numerical studies of two-body mean motion resonances of different order have been accomplished (see [21, 27, 30, 37] and references therein).

Here, a numerical study is given in the framework of the planar elliptic restricted three body problem in order to give an estimate of the size and the shape of the chaotic regions. We integrate the variational equations along with the equations of motion for a set of 500×500 test particles placed on a regular grid in the plane (a, e) , where the semimajor axis a range from 1.5 to 6 AU, while the eccentricity $e \in [0, 0.5]$. The total time span covered by our integration has a maximum length of 8,400 y (700 periods of Jupiter). On the obtained dynamical maps, the structure of the above portion of the phase space is clearly displayed.

We recall that the phase space of the circular and elliptic restricted three-body problem has been investigated previously by Sándor et al. [35] by using the method of short time average of the stretching numbers. In the semimajor axis-eccentricity plane of the test particle, Sándor et al. [35] studied some features of the region $a \in [3 \text{ AU}, 4.5 \text{ AU}]$, $e \in [0, 0.4]$ for three cases of the eccentricity of the primaries: $e_J = 0$, $e_J = 0.048$ and $e_J = 0.1$.

Another study [37] is based on the computation of the Lyapunov time in the framework of 2D and 3D elliptic three-body problem for 1 Gy in order to give an estimate of the removal time from different mean motion resonances.

Here, we consider the same problem, but we use a different method. The FLI is computationally cheap, robust, and easy to implement. In contrast with the above mentioned numerical methods, FLI is also very sensitive to detect weak chaos. The dynamical maps show not only the regular regions, the chaotic ones, and the geography of mean motion resonances of low order, but there are also displayed very thin layers associated with mean motion resonances of moderate order. Moreover, inside the mean motion resonances there are revealed other small chaotic lines which are due to the resonance splitting. We discuss our numerical results in the light of the theory of Hamiltonian systems.

2 Basic Formulation

In this section, we recall the definition of the FLI and describe the techniques utilized to integrate the equations of motion and the variational equations.

2.1 The Fast Lyapunov Indicator (FLI)

Let us consider the dynamical system

$$\frac{dX}{dt} = F(X(t)), \quad X \in \mathbb{R}^n, \quad t \in \mathbb{R}, \quad (1)$$

where F is a continuously differentiable function. Given an initial condition $X(0) \in \mathbb{R}^n$, let us consider the evolution $V(t) \in \mathbb{R}^n$ of an initial vector $V(0) \in \mathbb{R}^n$ of norm 1, obtained by integrating the variational equations

$$\frac{dV}{dt} = \frac{\partial F}{\partial X}(X(t))V, \quad (2)$$

where $X(t)$ is the evolution of $X(0)$. Then, the maximum Lyapunov exponent is defined by

$$MLE(X(0)) = \lim_{t \rightarrow \infty} \frac{\ln \|V(t)\|}{t}. \quad (3)$$

Numerically, one works on finite times. Thus, one estimates MLE by computing the Lyapunov characteristic indicator defined by

$$LCI(t; X(0)) = \frac{\ln \|V(t)\|}{t}. \quad (4)$$

at a large time t . The fast Lyapunov indicator is defined by (see [10, 24])

$$FLI(T; X(0)) = \sup_{0 < t < T} \ln \|V(t)\|. \quad (5)$$

The computation of FLI on a relatively short time is enough to discriminate between chaotic and regular orbits. The FLI of a regular orbit increases linearly, while for a chaotic orbit, the FLI increases exponentially. The FLI is very sensitive to detect weak chaos. Moreover, FLI may be used to discriminate among regular motion between nonresonant and resonant orbits. Because of the differential rotation, the norm of the vector V , asymptotically grows as $\|V(t)\| \cong \alpha t$, the coefficient α depending on the nature of invariant curve (torus or libration island). As against the MLE, there is a disadvantage. Namely, FLI depends on the initial conditions and if the system is Hamiltonian, on the choice of the canonical variables. But, once some reference orbits have been computed for which the chaotic (or regular) nature has been determined, the FLI allows the investigation of a large number of orbits and the computations of dynamical maps like the ones presented in this chapter.

2.2 The Restricted Three Body Problem

The three-body problem simplified by setting one mass to zero is widely known as the restricted three-body problem. In this case, the equations for the two primaries S (Sun) and J (Jupiter) decouple, so that their motion is Keplerian. According to the value of the eccentricity of their orbits we speak of the circular or elliptic problem. Here, we consider the elliptic problem. We define our system of units such that $G(m_S + m_J) = 1$, where G denotes the constant of gravitation, m_S the mass of the Sun, and m_J the mass of Jupiter. Let us introduce the notation

$$\mu = Gm_J = \frac{m_J}{m_S + m_J} = 9.54 \cdot 10^{-4}. \quad (6)$$

Then, the heliocentric equations of motion of Jupiter and the infinitesimal body are [29]

$$\begin{aligned} \frac{d^2 \mathbf{r}_J}{dt^2} &= -\frac{\mathbf{r}_J}{\|\mathbf{r}_J\|^3}, \\ \frac{d^2 \mathbf{r}}{dt^2} &= -(1 - \mu) \frac{\mathbf{r}}{\|\mathbf{r}\|^3} + \mu \left[\frac{\mathbf{r}_J - \mathbf{r}}{\|\mathbf{r}_J - \mathbf{r}\|^3} - \frac{\mathbf{r}_J}{\|\mathbf{r}_J\|^3} \right], \end{aligned} \quad (7)$$

where \mathbf{r}_J and \mathbf{r} are their position vectors relative to the Sun. We have numerically integrated the above equations using as starter a single step method (Runge–Kutta), while a symmetric multistep method of 12th order (proposed by Quinlan and Tremaine [34]) performs most of the propagation. We recall that the numerical integration is subject to several sources of integration error, namely truncation

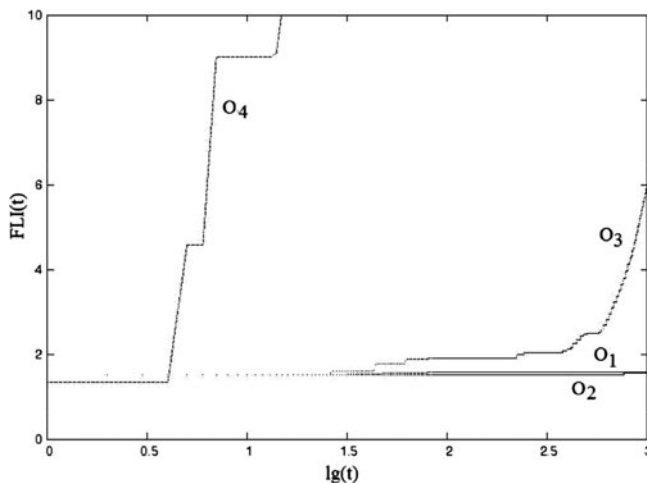


Fig. 1 Evolution of FLI as a function of $lg(t)$ for the orbits: O_1 (regular orbit), O_2 (regular resonant orbit), O_3 (weak chaotic orbit) and O_4 (strong chaotic orbit). The unit of time is the period of Jupiter

error, round-off error, error in the initial conditions and masses of the bodies, error in the physical model etc. Quinlan and Tremaine [34] showed that symmetric multistep methods reduce the truncation error and illustrated by numerical examples the superiority of the symmetric multistep methods over other methods. Based on those discussed in [34] concerning the optimum choice of the order and the coefficients of the integrator, we chose a 12th order method in our computations. For the corresponding variational equations, we have utilized again the Runge–Kutta method. For each initial condition, the total time span covered by the integration has a length of maximum 8,400 y (700 periods of Jupiter). The computations for a longer time (over 1,000 periods of Jupiter) did not change significantly the result. Therefore, we used a shorter computation time for the whole analysis. Moreover, since FLI increase exponentially with time for chaotic orbits, we stopped the integration if $\ln(FLI)$ exceeded the value 1.8.

The initial conditions have been chosen such that the second primary (Jupiter) describes an elliptic motion with eccentricity $e_J = 0.048$ and at time $t = 0$, Jupiter is at perihelion. We have utilized a Cartesian coordinate system centered on the Sun with the x -axis pointing towards the perihelion of Jupiter.

We suppose the massless body moves in a coplanar heliocentric elliptic orbit perturbed by Jupiter. Its initial conditions are given once the semimajor axis a (in AU), the eccentricity e , the argument of perihelion ω (the angle between the perihelion line and the x line), and the mean anomaly M are prescribed.

In order to exemplify the behavior of FLI for different kind of orbits, we plotted the evolution of the values of FLI for four orbits: O_1 (regular orbit), O_2 (regular resonant orbit), O_3 (weak chaotic orbit), and O_4 (strong chaotic orbit) during 1,000 periods of Jupiter (see Fig. 1). The orbits correspond to the following initial conditions: O_1 : $a = 2.43$ AU, $e = 0.15$, $\omega = 0^\circ$, $M = 0^\circ$; O_2 : $a = 2.49$ AU, $e = 0.15$,

$\omega = 0^\circ, M = 0^\circ$; O_3 : $a = 2.53 AU, e = 0.15, \omega = 0^\circ, M = 0^\circ$; and O_4 : $a = 4.7 AU, e = 0.15, \omega = 0^\circ, M = 0^\circ$. The regular orbit O_1 is close to the 3 : 1 resonance, O_2 is in the libration region of the 3 : 1 resonance, O_3 is on the separatrix of the 3 : 1 resonance, while O_4 is located in the unstable region (see Fig. 3). For the regular orbits O_1 and O_2 , FLI increases linearly. The FLI of the chaotic orbits O_3 and O_4 increases exponentially, but with different rates.

3 Results

In this Section, we describe numerically the structure of a selected region of the phase space of the planar elliptic restricted three-body problem (PERTBP) by computing dynamical maps, such as the ones presented in Figs. 3–7. In order to give a theoretically based interpretation to our numerical analysis, we recall first some important definitions and analytical results.

The phase space of the PERTBP is four dimensional. Thus, from a theoretical point of view, to describe a portion of it, a dense network of points covering a subset of \mathbb{R}^4 should be investigated. However, since our model is a quasi-integrable Hamiltonian system, we resort to the space of the actions.

In the case of quasi-integrable Hamiltonian systems having a nondegenerate integrable part, the KAM theorem (see for example Celletti and Chierchia [2] for a recent description of the state of the art of the theory) assures the persistence of invariant tori carrying motion with diophantine frequencies, provided the perturbations are small enough. In other words, the nondegenerate integrable approximation H_0 gives a foliation of the phase space in invariant tori, the actions being constants and the angles circulating linearly with time. When a small perturbation εH_1 is added, the KAM theorem ensures that some invariant tori with diophantine frequencies continue to be invariant for the complete Hamiltonian $H_0 + \varepsilon H_1$. The size ε determines which tori continue to be invariant among all the unperturbed ones with diophantine frequencies.

Clearly, the integrable part of the PRTBP, namely the two-body problem is highly degenerate, i.e. the Hamiltonian does not depend on some action variables and so the hypotheses of the KAM theorem and also of the stability theorem of Nekhoroshev are not satisfied in their standard form. For this reason, the PRTBP exhibit very complicated dynamics. In the phase space, we identify: regular regions, chaotic areas, and resonant regions with its libration and chaotic zones. We discuss their size as function of the resonance order and the parameters entering into the perturbing function, in particular the argument of perihelion and the mean anomaly.

Let us recall now the Hamiltonian and the canonical variables of the planar restricted three-body problem. With our choice of dimensions and the Cartesian coordinate system, the modified Delaunay variables are [7, 27]

$$\begin{aligned} \Lambda &= \sqrt{(1-\mu)a}, & \lambda &= M + \omega, \\ \Omega &= \Lambda(1 - \sqrt{1-e^2}), & -\omega &. \end{aligned} \tag{8}$$

The autonomous Hamiltonian has the form

$$H = -\frac{(1-\mu)^2}{2\Lambda^2} + n_J \Lambda_J - \mu f(\Lambda, \Omega, \lambda, \omega, \lambda_J), \quad (9)$$

where $n_J = \dot{\lambda}_J$ is the mean motion of Jupiter, λ_J is the mean longitude, and Λ_J is the conjugate action corresponding to Jupiter. The disturbing function f , also depends on the constant elements of Jupiter's orbit (for a detailed description of the perturbation theories the reader is referred to the books of Murray and Dermott [29] and Ferraz-Mello [7]). The integrable part of H is strongly degenerate. The fast angle λ is nondegenerate, while the slow angle ω is degenerate.

Now, we recall that the resonances

$$(p+q)\dot{\lambda} - p\dot{\lambda}_J = 0, \quad p, q \in \mathbb{Z} \quad (10)$$

are called mean motion resonances. The integer q is the order of resonance and determines its strength. It is known that asteroids can develop chaotic motion as a result of resonant perturbations, exerted by the major planets. The asteroid–Jupiter mean motion resonances of low order ($1 \leq q \leq 4$) force the resonant asteroids to become planet crossers on a short timescale (of order of a few million years), while in mean motion resonances of moderate order ($5 \leq q \leq 7-9$) the times required to become a planet crosser are much longer (from tens of millions of years to of order 1 Gyr) (see [27]).

In order to investigate numerically the topology of the phase space, we proceeded as follows: in the domain $0 \leq e \leq 0.5$, $1.5 \text{ AU} \leq a \leq 6 \text{ AU}$, of the action plane (Λ, Ω) we considered a grid of 500×500 equidistant initial conditions. We considered various choices of the initial angles ω and M . We report here only the figures obtained for: $\omega = 0^\circ, M = 0^\circ$ (Fig. 3); $\omega = 90^\circ, M = 0^\circ$ (Fig. 4); $\omega = 180^\circ, M = 0^\circ$ (Fig. 5); $\omega = 60^\circ, M = 0^\circ$ (Fig. 6); $\omega = -60^\circ, M = 0^\circ$ (Fig. 7); For each point on the grid, we calculated the final value of the logarithm of FLI for maximum 700 periods of Jupiter. The results are reported in the (Figs. 3–7), where the gray scale is used in such a way that white color corresponds to chaotic orbits, whereas the darker the color is, the more stable the orbit is.

Moreover, we represented the distribution of the main belt and Trojans asteroids (Fig. 2) using the osculating semimajor axis and eccentricity, in order to display the location of the mean motion resonances with Jupiter. We recall the location of some low order resonances: 4 : 1 ($a = 2.06 \text{ AU}$); 3 : 1 ($a = 2.5 \text{ AU}$); 5 : 2 ($a = 2.82 \text{ AU}$); 7 : 3 ($a = 2.95 \text{ AU}$); 2 : 1 ($a = 3.277 \text{ AU}$); 7 : 4 ($a = 3.58 \text{ AU}$); 5 : 3 ($a = 3.7 \text{ AU}$); 3 : 2 ($a = 3.9 \text{ AU}$); 1 : 1 ($a = 5.2 \text{ AU}$).

Figures 3–7 are in a good agreement with the analytical and numerical studies on the global dynamics of the asteroid belt (see [27, 29, 37] and references therein). Analyzing these maps, we may conclude the following: Depending on the initial proper elements a, e, ω, M , the infinitesimal mass can undergo chaotic or regular motions, and the phase space can be partitioned into the following zones:

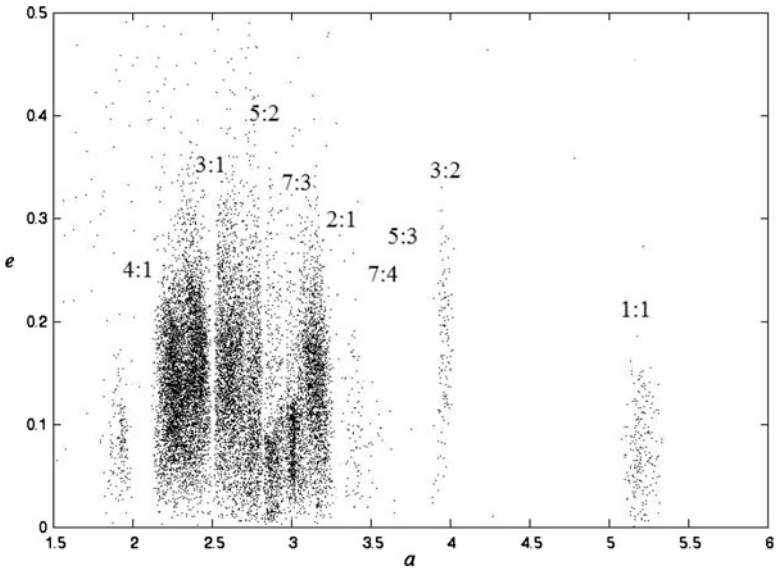


Fig. 2 Distribution of the main-belt and Trojans asteroids on the (a, e) -plane

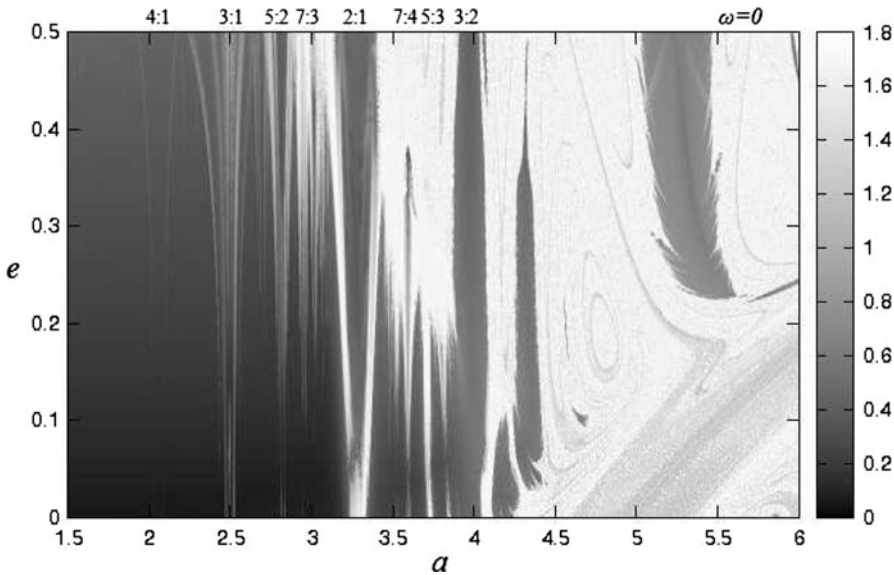


Fig. 3 Maps of the logarithm of the final value of FLI after 700 periods of Jupiter. The test particles were placed on a regular 500×500 grid in (a, e) plane and the initial angles was $\omega = 0^\circ$, $M = 0^\circ$. The values of $\lg(FLI)$ are color-coded, according to the scale shown on the right (*black* = regular, *white* = chaotic)

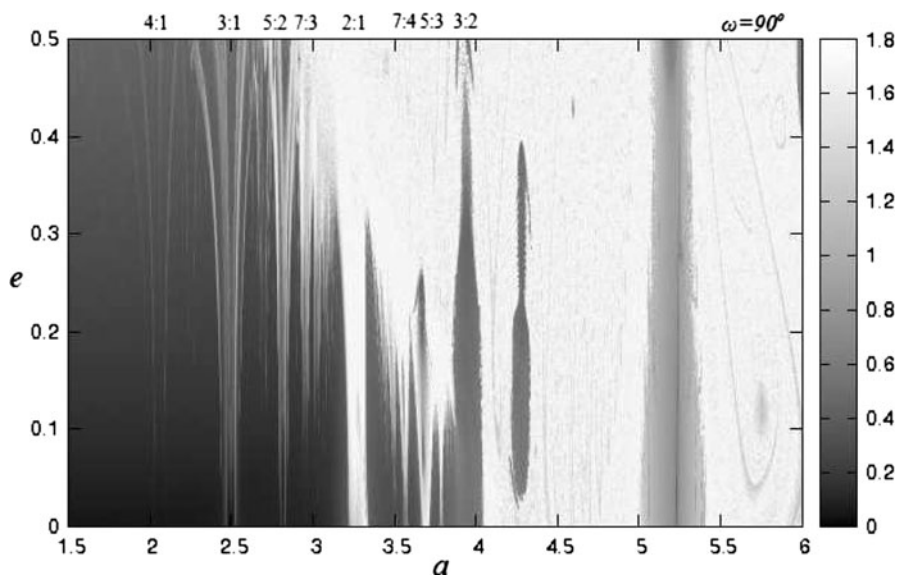


Fig. 4 Same as Fig. 3, but for $\omega = 90^\circ$, $M = 0^\circ$

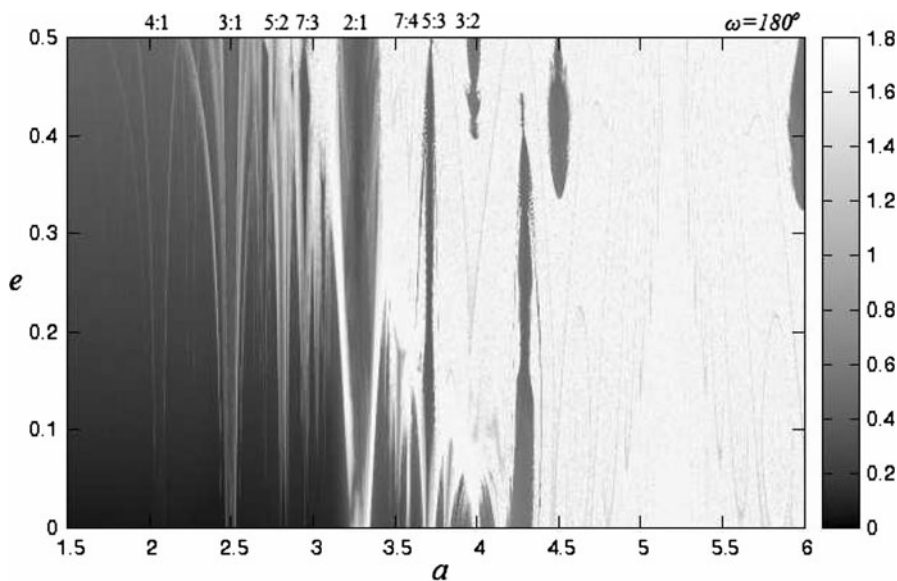


Fig. 5 Same as Fig. 3, but for $\omega = 180^\circ$, $M = 0^\circ$

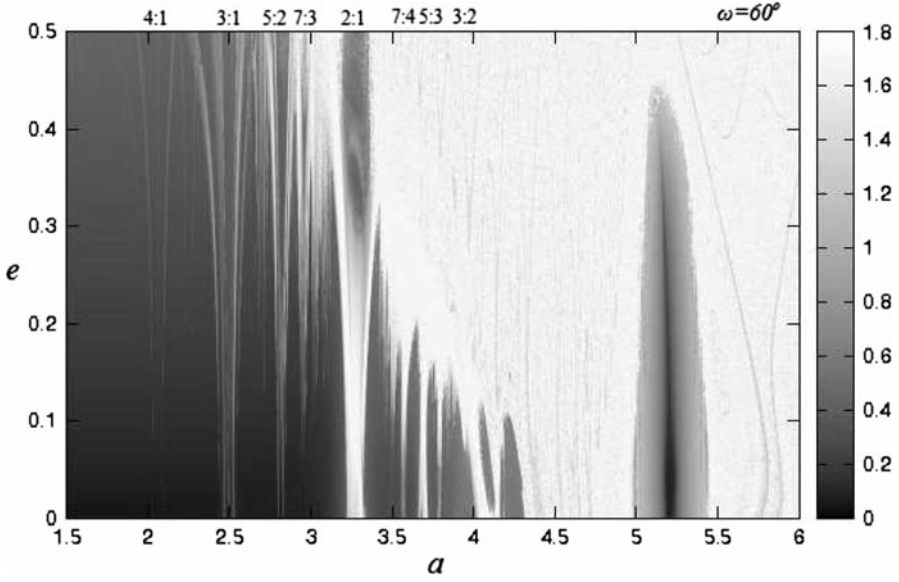


Fig. 6 Same as Fig. 3, but for $\omega = 60^\circ$, $M = 0^\circ$

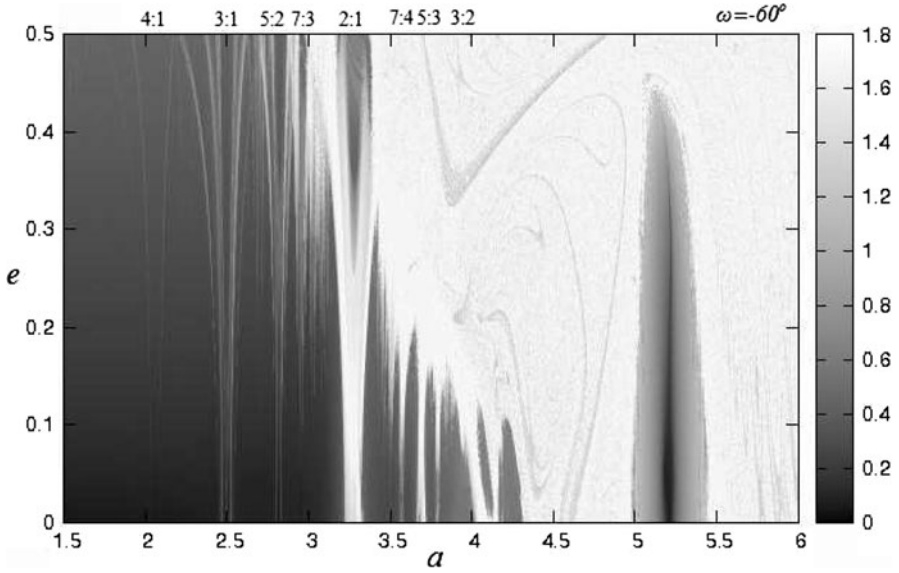


Fig. 7 Same as Fig. 3, but for $\omega = -60^\circ$, $M = 0^\circ$

The *stability region* located between 1.5 and 3.2 AU is distinguished on each map. Moreover, the maps computed for $\omega = 90^\circ$, $\omega = 60^\circ$ and $\omega = -60^\circ$ (Figs. 4, 6, and 7) show the regular zones associated with 1 : 1 jovian resonance, while on

the Figs. 3 and 4 the regular region corresponding to 3 : 2 resonance is exhibited. Clearly, since we did not consider the perturbation exerted by Mars and Saturn, the stability regions displayed on our charts have a sensible bigger area than they should normally have when additional planets are taken into account. The results are in concordance with the analytical studies. Although many quasi-integrable systems, including the gravitational ones, are degenerate, some stability results have been proved by adapting the perturbation techniques to the considered model. For example, Morbidelli and Guzzo [28] studied the stability of the asteroid belt dynamical system and showed that a Nekhoroshev-like result might be constructed on domains that exclude all the mean motion resonances of order smaller than the logarithm of the mass of Jupiter. Moreover, they proved that Nekhoroshev result is strictly connected with the existence of a specific structure of the phase space even for a degenerate system. It is clear that the dark regions on the Figs. 3–7 match with these analytical studies.

Inside this domain, we expect to be many KAM tori. We recall that in the framework of the restricted circular planar three-body problem, Celletti and Chierchia [3] obtained a stability result for the asteroid Victoria ($a = 2.33$ AU, $e = 0.22$) by showing the existence of two nearby trapping KAM tori at the same energy level.

The *strong chaotic motions* arise as a consequence of the resonance overlap criterion. This criterion was discussed by Wisdom [39]. One could consider the phase space as being made of a succession of resonances, each independent of the others and having its own librational and chaotic regions, as in the case of the perturbed pendulum. An obvious example is the sequence of the first order interior resonances of the form $p + 1 : p$. Since each resonance has a well-defined width in semimajor axis, and since the separation of adjacent resonances becomes smaller as the perturber is approached, there will come a point at which these overlap (for example, the resonances: 4 : 3 ($a = 4.29$ AU); 5 : 4 ($a = 4.48$ AU); 6 : 5 ($a = 4.6$ AU) etc.) As a consequence, we would expect a cleared zone in the asteroid belt beyond 4.3 AU. This is in good agreement with the observations (see Fig. 2). The chaotic zone predicted by the above described overlap criterion is obtained in each picture. For $a \geq 4.3$ AU a large white zone, whose shape depends on ω , is obtained by numerical integration.

The stability region is crossed by a series of “V”-shaped layers of various sizes. These layers correspond to the *mean motion resonances of low order* (4 : 1 ($a = 2.06$ AU); 3 : 1 ($a = 2.5$ AU); 5 : 2 ($a = 2.82$ AU); 7 : 3 ($a = 2.95$ AU); 2 : 1 ($a = 3.277$ AU); 7 : 4 ($a = 3.58$ AU); 5 : 3 ($a = 3.7$ AU); see also the Fig. 2). Moreover, between 2.5 and 3.27 AU, we can distinguish very thin white lines which for high eccentricity overlap with the mean motion resonance of low order. These lines correspond to moderate-order mean motion resonances. From a mathematical point of view, the dynamics inside mean motion resonances can be explained by analogy to a perturbed pendulum. In the pendulum case the phase space has two regions, the libration and the circulation zones, separated by separatrix. When the perturbing function is taken into account, the separatrix disappears. Its place is taken by a chaotic region, whose size depends on the perturbation.

For each mean motion resonance of low order (see for example the resonance 3 : 1 ($a = 2.5$ AU)), we recognize its chaotic border (separatrix) and the libration region. Moreover, inside mean motion resonances we have other small chaotic lines, whose size and location depend again on ω . These regions are due by the contribution to the disturbing functions of the possible resonant arguments. Due to degeneracy of the problem, each resonance splits into a multiplet of resonances. For example, the angles associated with the 3 : 1 resonance are $\varphi_1 = 3\lambda_J - \lambda - 2\omega$, $\varphi_2 = 3\lambda_J - \lambda - \omega - \omega_J$ and $\varphi_3 = 3\lambda_J - \lambda - 2\omega_J$. Since ω have a small but nonzero frequency, φ_1 , φ_2 , and φ_3 have zero derivatives at different locations. Therefore, the 3 : 1 mean motion resonance splits in a natural way into a threeplet of resonances. The exact location of each component is given by $\dot{\varphi}_k = 0$, $k = 1, 2, 3$. The image of a given mean motion resonance varies from map to map since $\dot{\omega}$ depends on ω and M via Lagrange's equations.

Finally, we note that the libration region around mean motion resonance decreases with the order q and increases with the eccentricity e . In other words the layer around a mean motion resonance has a “V”-shape and occupies smaller area once the order q increases. These results are in agreement with the analytical perturbation theory (see, for example, the book of Murray and Dermott [29]), which guaranties higher terms in disturbing function once the order q is low and the eccentricity e is high. In fact, the coefficients of the resonant terms are proportional to e^q .

4 Conclusions

The two-body problem is highly degenerate. For this reason, even small perturbations of the two-body problem, like the restricted three-body problem, may exhibit very complicated dynamics. In this work, we considered the planar elliptic restricted three-body problem (the Sun–Jupiter–Asteroid system) and by using the fast Lyapunov indicator we studied numerically the global topology of the phase space. On each dynamical map, regular regions, chaotic zones, and “V”-shaped layers around the mean motion resonances of low order, predicted by analytical theories (see e.g. [27, 29, 37] and references therein) are revealed by the FLI in a very short computational time. The degeneracy of the problem, pointed out by the resonance splitting, is clearly illustrated in the Figs. 3–7. Secular and mean-motion resonances of low order are known to lead to fast chaotic transport of asteroid orbits on million-year time scales [17, 27].

On our dynamical maps, some thin layers associated with mean motion resonances of moderate order are displayed. These resonances together with the three-body mean motion resonances (asteroid–Jupiter–Saturn) (see [32]) form a dense network of thin chaotic layers throughout the asteroid belt, where small-amplitude variations of proper elements of asteroids accumulate slowly over time. This effect is known as chaotic diffusion [37]. The results obtained here for the PERTBP encourage the application of the FLI to study this fine chaotic structure of the asteroid belt.

Acknowledgment This work was supported from the POSDRU/89/1.5/S/49944 Project.

References

1. Celletti, A.: Weakly dissipative systems in Celestial Mechancs. *Lect. Notes Phys.* **729**, 67–90 (2008)
2. Celletti, A., Chierchia, L.: KAM tori for N-body problems (a brief history). *Celest. Mech. Dynamical Astron.* **95**, 117–139 (2006)
3. Celletti, A., Chierchia, L.: KAM stability and celestial mechanics. *Memoirs of the American Mathematical Society* vol. 187, n. 878 (2007)
4. Celletti, A., Froeschlé, C., Lega, E.: Frequency analysis of the stability of asteroids in the framework of the restricted three-body problem. *Celest. Mech. Dynamical Astron.* **90**, 245–266 (2004)
5. Cincotta, P., Simó, C.: Simple tools to study global dynamics in non-axisymmetric galactic potentials-I. *Astron. Astrophys. Suppl. Ser.* **147**, 205–228 (2000)
6. Contopoulos, G., Voglis, N.: A fast method for distinguishing between order and chaotic orbits. *Astron. Astrophys.* **317**, 73–81 (1997)
7. Ferraz-Mello, S.: *Canonical perturbations theories. Degenerate systems and resonance.* Springer, New York (2007)
8. Fouchard, M., Lega, E., Froeschlé, Ch., Froeschlé, C.: On the relationship between fast Lyapunov indicator and periodic orbits for continuous flows. *Celest. Mech. Dynamical Astron.* **83**, 205–222, (2002)
9. Froeschlé, C., Lega, E.: Twist angles: a method for distinguishing islands, tori and weak chaotic orbits. Comparison with other methods of analysis. *Astron. Astrophys.* **334**, 355–362, (1998)
10. Froeschlé, C., Lega, E.: On the structure of symplectic mappings. The fast Lyapunov indicator: a very sensitive tool. *Celest. Mech. Dynamical Astron.* **78**, 167–195 (2000)
11. Froeschlé, C., Froeschlé, Ch., Lohinger, E.: Generalized Lyapunov characteristic indicators and corresponding Kolmogorov like entropy of the standard mapping. *Celest. Mech. Dynamical Astron.* **83**, 205–222 (1993)
12. Froeschlé, C., Lega, E., Gonczi, R.: Fast Lyapunov indicators. Application to asteroidal motion. *Celest. Mech. Dynamical Astron.* **67**, 41–62, (1997)
13. Froeschlé, C., Gonczi, R., Lega, E.: The fast Lyapunov indicator: a simple tool to detect weak chaos. Application to the structure of the main asteroidal belt. *Planet. Space Sci.* **45**, 881–886 (1997)
14. Froeschlé, C., Guzzo, M., Lega, E.: Graphical evolution of the Arnold web: from order to chaos. *Science* **289**, 2108–2110 (2000)
15. Froeschlé, C., Guzzo, M., Lega, E.: Analysis of the chaotic behaviour of orbits diffusing along the Arnold web. *Celest. Mech. Dynamical Astron.* **95**, 141–153 (2006)
16. Gidea, M., Deppe, F., Anderson, G.: Phase space reconstruction in the restricted three-body problem. In: *New Trends in Astrodynamics and Applications III* (Eds. E. Belbruno), AIP Conference Proceedings Volume 886, (2007)
17. Gladman, B. et al.: Dynamical lifetime of objects injected into asteroid belt resonances. *Science* **277**, 197–201 (1997)
18. Goździewski, K.: A dynamical analysis of the HD 37124 planetary system. *Astron. Astrophys.* **398**, 315–325 (2003)
19. Goździewski, K., Bois, E., Maciejewski, A.J., Kiseleva-Eggleton, L.: Global dynamics of planetary systems with the MEGNO criterion. *Astron. Astrophys.* **378**, 569–586 (2001)
20. Guzzo, M., Lega, E., Froeschlé, C.: On the numerical detection of the effective stability of chaotic motions in quasi-integrable systems. *Physica D* **163**, 1–25 (2002)
21. Holman, M., Murray, N.: Chaos in high order mean motion resonances in the outer asteroid belt. *Astron. J.* **112**, 1278–1293 (1996)
22. Laskar, J.: The chaotic motion of the Solar System. A numerical estimate of the size of the chaotic zones. *Icarus* **88**, 266–291, (1990)
23. Laskar, J., Froeschlé, C., Celletti, A.: The measure of chaos by the numerical analysis of the fundamental frequencies. Applications to the standard mapping. *Physica D* **56**, 253–269 (1992)

24. Lega, E., Froeschlé, C.: On the relationship between fast Lyapunov indicator and periodic orbits for symplectic mappings. *Celest. Mech. Dynamical Astron.* **81**, 129–147 (2001)
25. Lega, E., Guzzo, M., Froeschlé, C.: Detection of Arnold diffusion in Hamiltonian systems. *Physica D* **182**, 179–187 (2003)
26. Lega, E., Guzzo, M., Froeschlé, C.: Diffusion in Hamiltonian quasi-integrable systems. *Lect. Notes Phys.* **729**, 29–65 (2008)
27. Morbidelli, A.: *Modern Celestial Mechanics. Aspects of the Solar System Dynamics*. Taylor and Francis, London (2002)
28. Morbidelli, A., Guzzo, M.: The Nekhoroshev theorem and the asteroid belt dynamical system. *Cel. Mech Dynamical Astron.* **65**, 107–136 (1997)
29. Murray, C.D., Dermott, S.F.: *Solar System Dynamics*. Cambridge University Press, UK (1999)
30. Murray, N., Holman, M.: Diffusive chaos in the outer asteroid belt. *Astron. J.* **114**, 1246–1259 (1997)
31. Nesvorný, D., Ferraz-Mello, S.: On the asteroidal population of the first-order Jovian resonances. *Icarus* **130**, 247–258 (1997)
32. Nesvorný, D., Morbidelli, A.: Three-body mean motion resonances and the chaotic structure of the asteroid belt. *Astron. J.* **116**, 3029–3037 (1998)
33. Pilat-Lohinger, E.: Eccentric orbits in double stars. *Proceedings of the 3rd Austrian-Hungarian Workshop on Trojan and related Topics* (Eds. F. Freistetter, R. Dvorak and B. Érdi), 35–45 (2003)
34. Quinlan, G.D., Tremaine, S.: Symmetric multistep method for the numerical integration of planetary orbits. *Astron. J.* **100**, 1694–1700 (1990)
35. Sándor, Z., Balla, R., Téger, F., Érdi, B.: Short time Lyapunov indicators in the restricted three-body problem. *Celest. Mech. Dynamical Astron.* **79**, 29–40 (2001)
36. Todorovic, N., Lega, E., Froeschlé, C.: Local and global diffusion in the Arnold web of a priori unstable systems. *Celest. Mech. Dynamical Astron.* **102**, 13–27 (2008)
37. Tsiganis, K.: Chaotic diffusion of asteroids. *Lect. Notes Phys.* **729**, 111–150 (2008)
38. Villac, B.F.: Using FLI maps for preliminary spacecraft trajectory design in multi-body environments. *Celest. Mech. Dynamical Astron.* **102**, 29–48 (2008)
39. Wisdom, J.: The resonance overlap criterion and the onset of stochastic behavior in the restricted three-body problem. *Astron. J.* **85**, 1122–1133 (1980)

Parameter Identification of the Langmuir Model for Adsorption and Desorption Kinetic Data

Dumitru Baleanu, Yeliz Yolcu Okur, Salih Okur,
and Kasim Ocakoglu

1 Introduction

The humidity adsorption and desorption kinetic data of spin-coated 50 nm Ruthenium polypridyl complex (Ru-PC K314) film has been measured under relative humidity between 11 % and 97% using Quartz Crystal Microbalance (QCM) technique. QCM has been extensively used for the determination and investigation of the kinetics of adsorption/desorption of adsorbate molecules for monolayer films [3, 4, 8]. QCM technique denotes a powerful approach for determining the sensing properties of materials before a sensor device design during development stages.

Langmuir model has been used successfully for monolayer films to analyze adsorption kinetics. For multilayer films, the Langmuir model cannot be used due to the diffusion of adsorbed molecules between layers. Therefore, it should be modified to determine the adsorption and desorption rates for the diffusion effect.

The chapter is organized as follows. In Sect. 2, we will give more information about the experiment. In Sect. 3, we briefly discuss about the celebrated Langmuir model, which was first introduced in [9]. In Sect. 4, we briefly mention about nonlinear regression analysis. In Sect. 5, we introduce the model that we propose for humidity adsorption and desorption kinetic data of spin-coated 50 nm Ruthenium polypridyl complex (Ru-PC K314) film and estimate the parameters of the model for the measured data. The last section constitutes a brief conclusion of the paper.

D. Baleanu (✉)

Department of Mathematics and Computer Science, Çankaya University, Ankara Turkey

First author is on leave from The National Institute for Laser, Plasma and Radiation, Physics, Institute of Space Sciences, Magurele Bucuresti, P.O. Box MG-23, R 76911, Romania
e-mail: dumitru@cankaya.edu.tr

2 Experimental

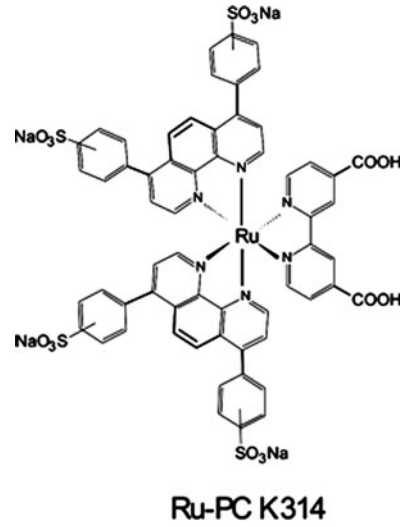
For the preparation of QCM electrodes, gold-coated quartz crystal electrodes were placed into ethanol and ultrasonically cleaned, then rinsed by deionized water. 1 mg/ml Ru-PC was dissolved in deionized water. 5 μ l of solution was spin-coated on to quartz crystal with 2,000 rpm. After drying at room conditions, it was kept in dessicator at room temperature for 3 hours. Then the quartz crystal coated with Ru-PC film was used to record both the reference frequency at 11% and the frequency changes up to 97% relative humidity. The thicknesses of films were measured using a Dektak profilometer from Veeco and found to be 50 nm.

A Time-Resolved Electrochemical Quartz Crystal Microbalance (EQCM) with the model of CHI400A Series from CH Instruments (Austin, USA) has been used to measure the change in the resonance frequency of quartz crystals between gold electrodes via both serial and usb interface connected to a computer. The QCM works with oscillation frequencies between 7.995 and 7.950 MHz. The density (ρ) of the crystal is 2.684 g/cm³, and the shear modulus (μ) of quartz is 2.947×10^{11} g/cm.s². Around oscillation frequency of 7.995-MHz, a net change of 1 Hz corresponds to 1.34 ng of materials adsorbed or desorbed onto the crystal surface of an area of 0.196 cm². For further information about QCM technique and its applications, see [2, 13, 14].

The signals coming from a QCM electrode and a commercial RH humidity sensor were simultaneously measured during the adsorption and desorption process. Both the relative humidity and temperature were also recorded during measurements while maintaining the temperature around 23°C. For this purpose, a EI-1050 selectable digital relative humidity and temperature probe with a response time of 4 s and a resolution of 0.03% RH was used with a USB controlled LabJack U12 ADC system combined with a single chip sensor module (SHT11) manufactured by Sensirion (Staeafa, Switzerland).

[Ru^{II} (bis(4,7-diphenyl-1,10-phenanthroline-disulfonic acid disodium salt) (4,4'-dicarboxy-2,2'-bipyridine)], [Ru-PC K314] (Fig. 1), was synthesized according to procedure given in the literature [12]. [RuCl₂(p-cymene)]₂, 1,10-phenanthroline-disulfonic acid disodium salt and 4, 4'-dicarboxy-2, 2'-bipyridine were obtained from Aldrich. All organic solvents were purchased from Merck and Fluka, and used without further purification. The final metal complex was characterized as following: ¹H NMR (400 MHz, D₂O) δ ppm: 9.56 (t, J = 1.7 Hz, 1H, NCHCHC, L2), 9.43 (t, J = 1.3 Hz, 1H, NCHCHC, L2), 8.73 (s, 1H, NCHCH, L1), 8.62 (s, 1H, NCHCH, L1), 8.10 (t, J = 1.1 Hz, 1H, CCHC, L2), 8.02 (t, J = 1.3 Hz, 1H, CCHC, L2), 7.78 (d, J = 3.6 Hz, 1H, NCHCHC, L2), 7.73 (d, J = 3.8 Hz, 1H, NCHCHC, L2), 7.62 (m, J = 1.4 Hz, 4H, CCHCHC, L1), 7.44 (q, J = 1.1 Hz, 2H, NCHCHC, L1), 7.40 (q, J = 1.2 Hz, 2H, NCHCHC, L1), 7.24 (m, J = 2.4 Hz, 4H, CCHCHCSO₃Na, L1), 7.18 (m, J = 2.6 Hz, 4H, CCHCHCSO₃Na, L1), 6.95 (m, J = 2.4 Hz, 4H, CCHCHCSO₃Na, L1), 6.89 (m, J = 2.4 Hz, 4H, CCHCHCSO₃Na, L1). UV (MeOH); λ_{max} : 483 (1.38), 314 (2.66), 278 (4.21), 221 (4.28). MS (MALDI): m/z = 1419.1 [M+H]⁺. Anal. Calc. For C₆₀H₃₆N₆Na₄O₁₆S₄Ru (1418.25): C, 50.81; H, 2.56; N, 5.93. Found: C, 50.79; H, 2.39; N, 5.88% [12].

Fig. 1 Chemical structure of ruthenium polyridyl complex



3 Langmuir Model

In this work, we deal with Langmuir model, which was first developed by Irving Langmuir in 1916 [5,6,9]. In order to analyze the adsorption and desorption kinetics of gas vapor molecules onto organic or inorganic films, Langmuir adsorption isotherm model is applied [1, 7, 15, 17, 18]. As a result, the model describes the rate of surface reaction for forming a monolayer on the surface by using the below equation,

$$\frac{d\theta}{dt} = k_a(1 - \theta) - k_d\theta. \quad (1)$$

Here, θ is a unitless quantity, which means the fraction of surface coverage, k_a and k_d denote the rate constants for the adsorption and desorption processes.

In this study, QCM has been used to measure the fractional coverage θ a function of time during the adsorption and desorption of water vapor molecules.

Hence, the difference between the oscillation frequency shift Δf of coated and uncoated QCM is directly proportional to the adsorbed mass of moisture molecules. The relationship between the surface adsorption kinetics and frequency shift (Δf) of QCM can be expressed as below

$$\frac{\Delta f}{dt} = (k_a C + k_d) \Delta f + k_a C \Delta f_{\max}. \quad (2)$$

where C is the water vapor concentration in the air. Moreover, note that during the adsorption process, (Δf) is equal to $\max \Delta f_{\max}$ for a very long time period.

4 Nonlinear Regression

The classical regression model assumes that the population regression function is the linear function of n independent variables, say x_i for $i = 1, \dots, n$ and the linear regression model can be written as in the following general matrix form:

$$y = \beta X + u,$$

where β is the vector of parameters, X is the vector of independent variables x_1, \dots, x_n , and u is the stochastic error term. The reason that we add a stochastic term is there are many unexpected and unobservable cases that will affect us to do some errors in modeling. The measurement error in calculations, missing variables in the model are the examples for such kind of unobservable errors. In order to have a realistic model, we must add a stochastic term to the model, the so-called disturbance term.

When we fit a model to our data, we obtain best-fit values that we can interpret in the context of the model. In many cases, the conditional expectation of the dependent variable is not a linear function of independent variables. Because of this reason, it is more practical to model with nonlinear systems to have more realistic models. For further information about nonlinear regression analysis, see [10] and [16].

5 The Main Result

The aim of this study is to fit the measured data of spin-coated 50 nm Ruthenium polypyridyl complex (Ru-PC K314) film to a curve by an appropriate methodology. Polynomial fitting, cubic spline, and linear regression are some fundamental techniques for this procedure. However, most of them ignore the theoretical part of the study and just focus on curve fitting. We use nonlinear regression methodology to fit a curve regarding fundamentals of the theory. Moreover, it is showed that many type of the data are best analyzed by using nonlinear least squares [11].

We assume that the difference between the oscillation frequency shift $y := \Delta f$ of coated and uncoated QCM can be modeled by a nonlinear regression function as follows

$$y_t = f(t, y, k_a, k_d) + \varepsilon_t,$$

for all $t \in [0, T]$. Here, f is a nonlinear function of time and the process itself with parameters k_a and k_d . Indeed,

$$f(t, y, k_a, k_d) = \frac{k_a C y_{\max}}{k_a C + k_d} + \left(y_0 - \frac{k_a C y_{\max}}{k_a C + k_d} e^{-(k_a C + k_d)(t-t_0)} \right), \quad (3)$$

where t_0 and y_0 are the initial values for t and y , respectively. Therefore, for the adsorption data, we use the following model:

$$y_t = \frac{k_a C y_{\max}}{k_a C + k_d} + \left(y_0 - \frac{k_a C y_{\max}}{k_a C + k_d} e^{-(k_a C + k_d)(t-t_0)} \right) + \varepsilon_t, \quad (4)$$

where the expectation and the variance of ε is zero and a constant number, respectively, i.e.,

$$E[\varepsilon_t] = 0$$

and

$$\text{Var}(\varepsilon_t) = \sigma^2$$

for a positive constant σ .

Note that for modeling adsorption data Langmuir is convenient since

$$\begin{aligned} \lim_{t \rightarrow t_0} y_t &= A_0, \\ \lim_{t \rightarrow \infty} y_t &= \frac{k_a C_1 \triangle f_{\max}}{k_a C_1 + k_d}. \end{aligned}$$

However, for desorption data Langmuir model is not a good candidate because we have to guarantee that

$$\begin{aligned} \lim_{t \rightarrow t_0} y_t &= y_{\max}, \\ \lim_{t \rightarrow \infty} y_t &= M, \end{aligned}$$

for a constant M which sufficiently small number and close to the minimum of the data series. Regarding all of these properties, we suggest an homogeneous exponential model for the desorption data instead of Langmuir model. We proposed the following one:

$$y_t = y_0 \exp(-(k_a C_2 + k_d)(t - t_0)) + \text{noise}. \quad (5)$$

5.1 Data Analysis

We have 3 adsorption and desorption cycles of data of Ruthenium polypyridyl complex film, which is measured by QCM technique (Figs. 2–4). The starting nodes, the maximum values, and the length of the time series are summarized by the following tables (Tables 1–3). Here are the details of the data of Ruthenium polypyridyl complex film:

Fig. 2 Plot of observed data for adsorption and desorption for cycle 1

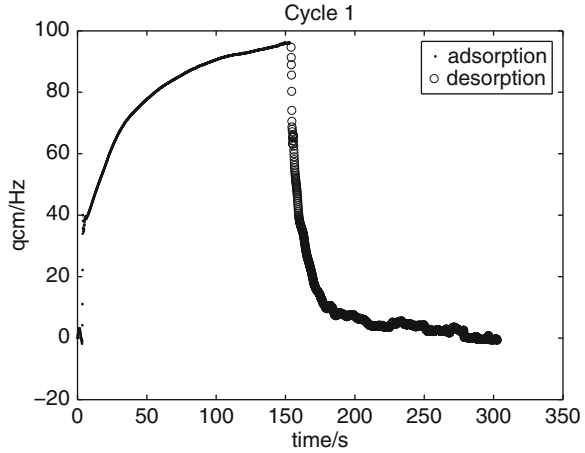
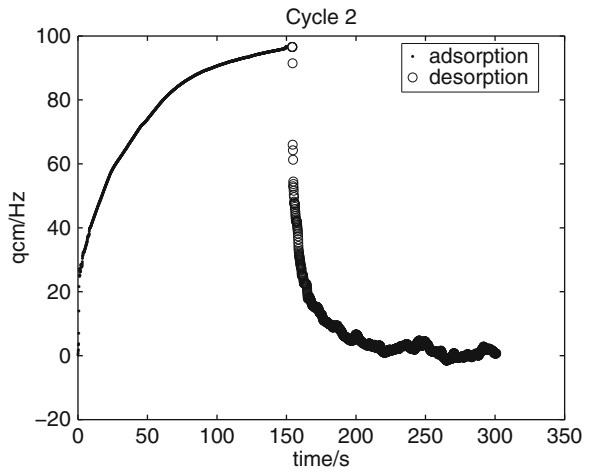


Fig. 3 Plot of observed data for adsorption and desorption for cycle 2



5.2 Parameter Estimation

Parameter estimation is the crucial part of mathematical modeling. An estimator attempts to approximate the unknown parameters using the measurements. In this paper, we use nonlinear least squares methodology to find the parameters of adsorption and desorption concentration rate. The parameters describe the physical setting such that it represents the distribution of the data.

However, we have realized two facts. First, the initial values of the original data does not satisfy the usual conditions of a function. Because of this reason we cut some initial points from the measured data in each cycle before estimating the parameters of the nonlinear least squares. Moreover, we realized that the desorption data series are noisy. So, first we denoise the original desorption series by wavelet

Fig. 4 Plot of observed data for adsorption and desorption for cycle 3

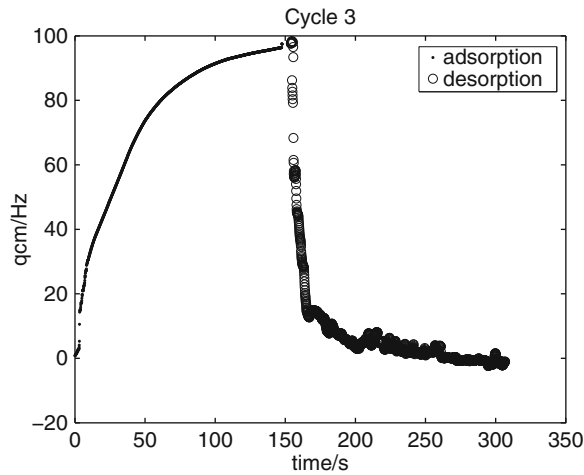


Table 1 Cycle 1

Properties	Adsorption	Desorption
Start	-3.81×10^{-7}	94.65
End	96.07	-0.63
Max	96.14	94.65
Length	1,532	1,482

Table 2 Cycle 2

Properties	Adsorption	Desorption
Start	0.18	96.57
End	96.67	0.75
Max	96.74	96.60
Length	1,499	1,469

Table 3 Cycle 3

Properties	Adsorption	Desorption
Start	0.81	97.56
End	97.43	-1.01
Max	97.53	98.33
Length	1,479	1,528

methodology. For each cycle, we have found the most appropriate wavelet family to denoise. It is explicitly specified at each table below. Every calculation is done by MATLAB[®] program (Tables 4–6).

For each cycle, we calculate the error terms (which are exact values minus estimated values) and show that they have zero expectation with a constant variance. We also check for the heteroscedasticity of the error terms in order to guarantee that the error terms have constant variance. We found that no ARCH effects exist.

Table 4 Cycle 1 estimation

Properties	Adsorption	Desorption
k_a	32.499	32.499
k_d	$1.02e-004$	0.103
# Cut points	50	–
Wavelet family	–	“Haar”

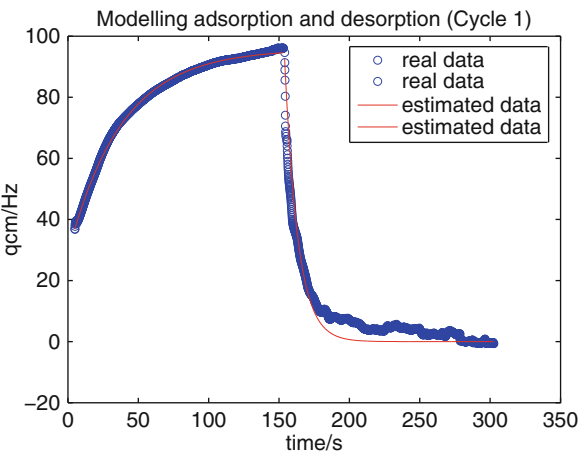
Table 5 Cycle 2 estimation

Properties	Adsorption	Desorption
k_a	32.331	32.331
k_d	$1.074e-004$	0.155
# Cut points	9	–
Wavelet family	–	“db3”

Table 6 Cycle 3 estimation

Properties	Adsorption	Desorption
k_a	34.554	34.554
k_d	$8.441e-005$	0.131
# Cut points	34	–
Wavelet family	–	“rbio1.5”

Fig. 5 Plot of estimated and observed data for cycle 1



Indeed, in each cycle of data, we observe that in the first cycle, the error terms has a mean of -1.1×10^{-15} and variance 0.0051. In the second cycle of data, the error terms has a mean of -0.9×10^{-15} and variance 0.0271. In the last cycle of data, their mean and variance are -0.9×10^{-15} and 0.05, respectively.

Let us show our results by plotting the estimated and observed data over time period. Note that for each cycle at the following graph, the blue circles show the original data of spin coated 50 nm Ruthenium polypridyl complex (Ru-PC K314) film, which was measured by QCM technique and the red line shows the estimated data by using the mathematical model introduced in the previous section (Figs. 5–7).

Fig. 6 Plot of estimated and observed data for cycle 2

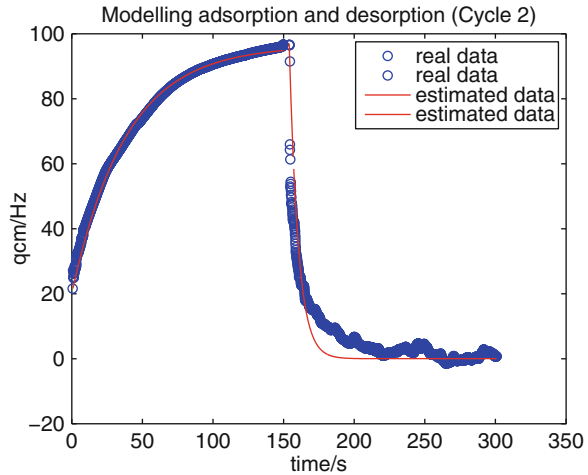
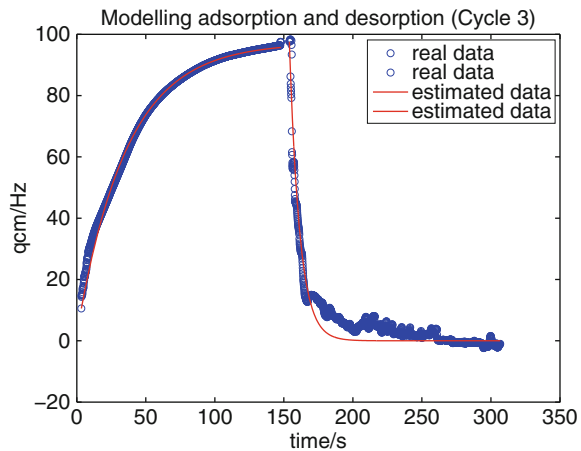


Fig. 7 Plot of estimated and observed data for cycle 3



6 Conclusion

In this paper, our aim is to find the parameters k_a and k_d of our models for given 3 cycles of data of Ruthenium polypridyl complex film, which is measured by QCM technique. We realized that the humidity desorption data of Ruthenium polypridyl complex film is noisy so we denoise the original data by appropriate wavelet families by using MATLAB[®] program. For each cycle 1, 2, and 3, we use haar, db3, and rbio1.5 wavelet families, respectively. We realized that adsorption concentration rate is in between 32.5 and 34.5. Moreover, the desorption concentration rate is found in between 0.1 and 0.2. We also conclude that the humidity adsorption data of spin coated 50 nm Ruthenium polypridyl complex film is smooth and the Langmuir model with disturbance term is appropriate to model.

References

1. Gregg, S.J., Sing, K.S.W.: Adsorption, Surface Area and Porosity. Academic Press (1982)
2. Erol, A., Okur S., Comba, B., Mermer, O., Arikan, M.C.: Humidity Sensing Properties of ZnO Nanodots Synthesized by Sol-Gel Process. *Sensors and Actuators B: Chemical* **145/1**, 174–180 (2010)
3. Hartmann, J., Auge, J., Lucklum, R., Rslar, S., Hauptmann, P., Adler, B., Dalcanale, E.: Supramolecular interactions on mass sensitive sensors in gas phases and liquids. *Sensors and Actuators B: Chemical* **34**, 305–311 (1996)
4. Kalchenko, V.I., Koshets, I.A., Matsas, E.P., Kopylov, O.N., Solovyov, A.V., Kazantseva, Z.I., Shirshov, Y.M.: Calixarene-based QCM sensors array and its response to 265 volatile organic vapours. *Journal of Materials Science* **20**, 73–88 (2002)
5. Kankare, J., Vinokurov I.A.: Kinetics of Langmuirian adsorption onto planar, spherical, and cylindrical surfaces. *Langmuir* **15**, 5591–5599 (1999)
6. Kapoor A., Ritter, J.A., Yang, R.T.: An Extended Langmuir model for adsorption of gas mixtures on heterogeneous surfaces. *Langmuir* **6**, 660–664 (1990)
7. Karpovich, D.S., Blanchard, G.J.: Direct measurement of the adsorption kinetics of Alkanethiolate Self-Assembled Monolayers on a microcrystalline gold surface. *Langmuir* **10**, 3315–3322 (1994)
8. Koshets, I.A., Kazantseva Z.I., Shirshov, Y.M., Cherenok, S.A., Kalchenko, V.I.: Calixarene films as sensitive coatings for QCM-based gas sensors. *Sensors and Actuators B: Chemical* **106**, 177–181 (2005)
9. Langmuir, I.: The constitution and fundamental properties of solids and liquids. *Journal of the American Chemical Society* **38**, 2221–2295 (1916)
10. Motulsky J.H., Christopoulos A.: Fitting Models to Biological Data Using Linear and Nonlinear Regression: a Practical Guide to Curve Fitting. Oxford University Press US (2004)
11. Motulsky J.H., Ransnas L.A.: Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *The FASEB Journal* **1**, 365–374 (1987)
12. Ocakoglu, K., Okur, S.: Humidity sensing properties of Novel Ruthenium Polypyridyl Complex. *Sensors and Actuators B: Chemical* **151/1**, 223–228 (2010)
13. Okur, S., Kus, M., Ozel, F., Aybek, V., Yilmaz, M.: Humidity Adsorption Kinetics of Calix [4] arene derivatives measured using QCM. *Talanta* **81/1-2**, 248–251 (2010)
14. Okur, S., Kus, M., Ozel, F., Aybek, V., Yilmaz, M.: Humidity Adsorption Kinetics of water soluble Calix [4] arene derivatives measured using QCM technique. *Sensors and Actuators B: Chemical* **145/1**, 93–97 (2010)
15. Sauerbrey, G.: Verwendung von schwingquarzen zur wägung dünner schichten und zur mikrowägung. *Z. Phys.* **155**, 206–222 (1959)
16. Seber G.A.F., Wild C.J.: Nonlinear Regression. Wiley (2003)
17. Su, P.G., Chang, Y.P.: Low Humidity sensor based on a quartz-crystal microbalance coated with polypyrrole/Ag/TiO₂ nanoparticles composite thin films. *Sensors and Actuators B: Chemical* **129**, 915–920 (2008)
18. Sun, Y.L., Wu, R.J., Huang, Y. C., Su, P.G., Chavali, M., Chen, Y.Z., Lin, C.C.: In situ prepared polypyrrole for low humidity QCM sensor and related theoretical calculation. *Talanta* **73**, 857–861 (2007)

Effects of Suspended Sediment on the Structure of Turbulent Boundary Layer

H.P. Mazumdar, S. Bhattacharya, and B.C. Mandal

1 Introduction

River flow is a class of turbulent boundary layer flow. It carries sediment. Suspended sediment is a portion of total sediment load carried by the rivers, and it plays a big role in morphological changes that occurs in rivers. Suspended sediment load is considered important in estimating the effects of land use changes and engineering practices in watercourses. Many areas of hydraulics and sediment transport require essentially knowledge of vertical velocity profile. Some insight into this problems have been gained from study of experimental results and evidence already presented by the profession [1,2,4,7,10,11,15]. In most of the above-mentioned papers, effects of the distribution of sediment concentration on the structure of turbulent boundary layer, specially the modification of the constants of the laws in the overlapping region and value of the wake parameter of the law of the wake are concerned.

Let us first analyze the structure of a single phase turbulent boundary layer as the results that to be drawn from such effort may subsequently be employed to find whether they could describe boundary layer structure with seeded particles. Further Coleman's [2] experimental data would be recalled for this purpose.

In the present analysis, it is admitted that the turbulent boundary layer may be split up [12] into two regions, namely: (I) the inner region where the law of the wall governs the flow and which is unaffected by outside manipulation (pressure changes, history, etc.) and (II) the outer region where the flow may be described by a law of the wake and velocity profiles are as such can take care of the outside manipulation of the turbulent boundary layer. Following Persen [12] a formal description on the

H.P. Mazumdar (✉)

Physics and Applied Mathematics Unit, Indian Statistical Institute,
203 B. T. Road, Kolkata-700108, India
e-mail: hpm@isical.ac.in

turbulent boundary layer will now be presented. Let us introduce the inner variables u^+, y^+ as

$$u^+ = \frac{u}{v_*}, \quad y^+ = \frac{yv_*}{\nu} \quad (1)$$

where u^+ is dimensionless velocity and y^+ is the dimensionless distance from the wall; u is the streamwise mean velocity; y is the distance from the wall; v_* is the friction velocity, defined by $v_*^2 = \tau_w/\rho$, τ_w being the wall shear stress; ρ and ν are the density and kinematic viscosity of the fluid.

One basic idea follows from dimensional consideration is that the law of the wall may be expressed in the form:

$$u^+ = f(y^+) \quad (2)$$

An objection to the law (2) is that it is too simple for describing the wake region (cf. [12]). Spalding [14] forwarded an analytical expression for $f(y^+)$ which led the law of the wall in the form:

$$y^+ = u^+ + A[\exp(\kappa u^+) - 1 - \kappa u^+ - (\kappa u^+)^2/2 - (\kappa u^+)^3/6 - (\kappa u^+)^4/24] \quad (3)$$

The formulation (3) has the advantage that it can be applied right from the wall. But Spalding's attempt to make the expression valid for the whole boundary layer with the choice of the constants κ and A , respectively, as $\kappa = 0.4$ and $A = 0.1108$ was too ambitious.

Persen [12] examined thoroughly the applicability of Spalding's formulation (3) against the voluminous data placed in the Stanford Conference (1968) on turbulent boundary layer and found the appropriate values of κ and A , respectively, as

$$A = 0.015, \quad \kappa = 0.53227, \quad (4)$$

which makes the Spalding formulation valid near to wall region only.

As discussed earlier, the turbulent boundary layer may assume to contain two regions, namely, (I). A 'near to wall region', which includes viscous sub-layer and wherein Spalding's formulation (3) is universally valid with the values of the constants κ and A , given in (4) and (II). A 'wake region' in which Persen proposed a formulation which allows a smooth matching of the velocity profile at the upper edge of wake region with the free stream velocity. In this concept, the joining point of the 'near to wall region' (inner region) with the wake region is mathematically well defined. Also, this has an advantage over the old concept of three regions (viscous sub-layer being one) where there are no well-defined conditions for their range of validity.

The establishment of Persen's theory depends on an experimentally supported relation between non-dimensional velocity $u^+ = \xi$ at the end point of the boundary layer and the corresponding non-dimensional distance from the wall $y^+ = y_0^+$. The point (ξ, y_0^+) at the edge of the boundary layer has been shown by Persen [12] to lie on a curve called the locus of ξ . The quantity is important on the point of view

that it may take care of manipulation from outside the boundary layer as well as the history of the boundary layer. It is worth mentioning that Persen's [12] theory has the special importance due to the fact that it is compatible with first principle of fluid mechanics. All these considerations lead us now to accept Persen's [12] theory and examine whether it is applicable to the open channel flow seeded with particles. In the present analysis, experimental data as measured by Coleman [2] for flow seeded with particles are brought into picture.

2 Coles' Wake Function (Coleman's Data)

Equation (3) being too simple for describing the wake region of the boundary layer and accordingly based on the idea of Coles [5] (3) should be replaced by

$$u^+ = f(y^+) + A(x)w(\eta) \quad (5)$$

where $A(x)$ is the amplitude function, $w(\eta)$ is the wake function and $\eta = y/\delta$, δ being the boundary layer thickness.

The wake function is generally defined as the difference between the measured data in the outer region of the boundary layer and values obtained from extension of logarithmic law in this region. The formulation proposed by Coles and Hirst [6] for the function $f(y^+)$ is

$$f(y^+) = \frac{1}{\kappa} \ln(y^+) + B \quad (6)$$

where

$$\kappa = 0.41, \quad B = 5 \quad (7)$$

This is the formulation valid for $y^+ > 50$. The method to be followed here will be somewhat different. Values of wake are determined as difference between the measured data in the outer region of the boundary layer and values obtained from (3) and (4) in that region. As $y^+ \rightarrow y_0^+$, $\eta \rightarrow 1$ one then obtains from (5) the relation

$$\xi = f(y_0^+) + A(x)w(1) \quad (8)$$

The maximum value w_{\max} and its position η_m at which it occurs are found by fitting a parabola through three points around the maximum value. Form now the wake function $w(\eta)$ is defined such that

$$w(\eta) = 1, \quad A(x) = w_{\max} \quad (9)$$

Effect of concentration on w_{\max} and η_m are shown in Figs. 1 and 2, respectively. It reveals that w_{\max} increases as sediment concentration increases while sediment concentration has more or less no effect on η_m .

Fig. 1 Variation of w_{\max} plotted as function of average sediment concentration

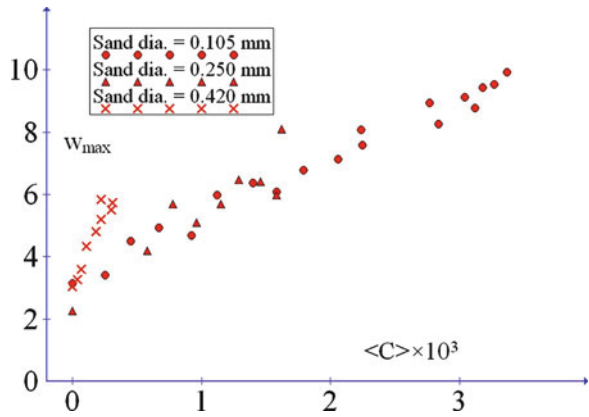
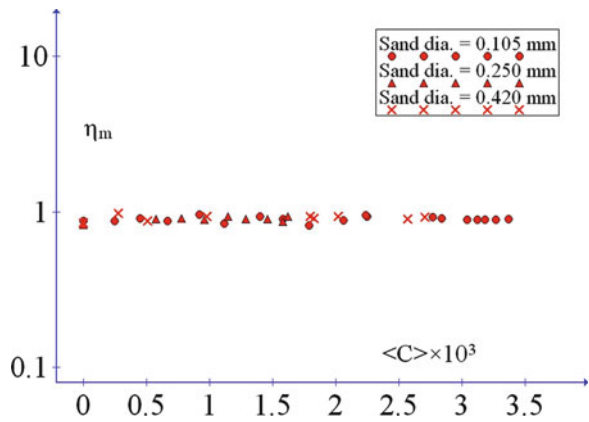


Fig. 2 Variation in the position η_m with average sediment concentration



3 Persen's Wake Law

Persen [12] replaced the Coles wake function with the expression

$$(u^+ - u_\infty^+)/(\xi - u_\infty^+) = \exp [-(y_0^+ - y^+)^2/\alpha^2] \quad (10)$$

where

$$\left. \begin{aligned} u &\rightarrow U_0 \text{ as } y \rightarrow \delta \\ u^+ &\rightarrow \frac{U_0}{v_*} = \xi \text{ as } y^+ \rightarrow \frac{\delta v_*}{v} = y_0^+ \end{aligned} \right\} \quad (11)$$

$u_\infty^+ = \text{constant}$, U_0 is the outer velocity and

$$\frac{1}{\alpha^2} = [\ln(\xi - u_\infty^+) - \ln(u_1^+ - u_\infty^+)]/(y_0^+ - y_1^+)^2$$

Here, (u_1^+, y_1^+) is the point where the law of the wall meets with the law of the wake. The boundary layer ends up at $y^+ = y_0^+$ and for which $u^+ = \xi$. Persen's law of the wake (10) has superiority over the Coles wake law as it exhibits a horizontal tangent at the outer edge ($u^+ = \xi$, $y^+ = y_0^+$) of the boundary layer and that applies also for the manipulated boundary layer (adverse pressure gradient etc.).

4 Colman Data: Open Channel Flow Seeded With Particles

The experiments of Coleman have been dully exposed to the profession through published papers (cf. [2, 4, 11]) and discussion (cf. [9] and Reply of Coleman [3]).

The experiments were performed in a re-circulating flume with a plexiglass channel 15 m long and 356 mm wide. The flume was supported on jacks so that the channel slope could be adjusted to maintain uniform flow. A pitot static tube on a vertical traverse mechanism was located at a position on the flume channel centerline 12 m from the channel entrance. The maximum outside diameter of the tube was 16 mm and the diameter of the impact leg opening was 3.2 mm. The impact leg could be isolated and could be used for taking suspended sediment samples. Uniform flow was maintained with constant discharge $0.064 \text{ m}^3/\text{s}$ and constant depth 169 mm with a standard deviation of 1.69 mm but with a systematic increase in sediment suspension. Sand particles of average diameters, namely 0.105 mm, 0.210 mm and 0.420 mm were used for seeding in the three series of experiments. No roughness elements were installed and no sand was allowed to deposit on the bed. Thus, any changes observed in the velocity profiles could be attributed to increase in suspended sediment only.

Water temperature varied within a narrow range around 23°C with a standard deviation of 1.2°C . So, value of viscosity and density of water at 23°C are considered throughout the analysis. The variation of kinematic viscosity is important in the vertical extent as it is dependent on the concentration distribution of the suspended sediments. The kinematic viscosity is modified for scaling purposes. We accept the relation between kinematic viscosity and concentration as given below [2, 8]:

$$\nu = \frac{\mu_w(1 + 2.5C + 6.25C^2 + 15.62C^3)}{\rho_w + (\rho_s - \rho_w)C} \quad (12)$$

where ν the kinematic viscosity of the sand water mixture, μ_w and ρ_w are the molecular viscosity and density of clear water, ρ_s is the density of sand and C is the local volumetric concentration which was expressed by Coleman in volume of sediment per unit volume of sediment water mixture.

In Coleman's [2] experiment, the aspect ratio (width: height) of the channel was about 2:1. Consequently, the influence of the side wall may be present in the centre plane and, accordingly velocity deep phenomenon was found to occur in this experiment.

Fig. 3 Wall shear stress as a function of average sediment concentration

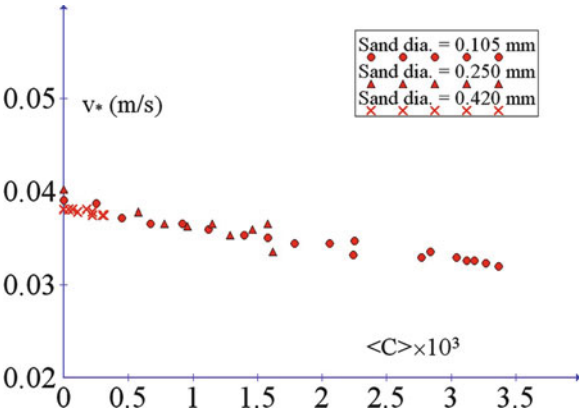
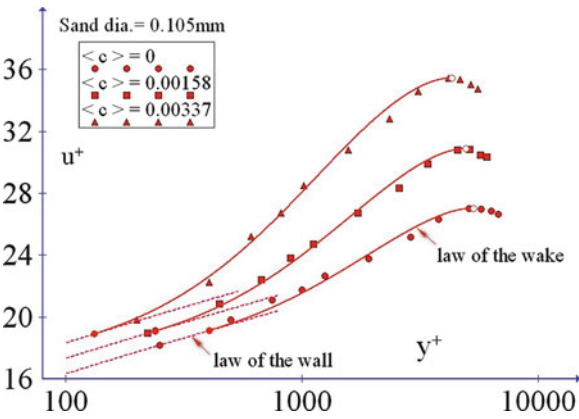


Fig. 4 Example on how the measured data and the theoretical curves correlate. The points (u_1^+, y_1^+) and (ξ, y_0^+) are shown as closed bold circle and open circle, respectively



The determination of wall shear stress or friction velocity v_* (m/s) is important as it is used for scaling purposes. The values of friction velocity v_* are taken from table published in Persen and Coleman [13]. Figure 3 shows the variation of wall shear stress with average sediment concentration $\langle C \rangle$. It is observed that wall shear stress is a decreasing function of concentration.

Once the outer edge condition ξ and y_0^+ are known u_1^+ and y_1^+ can be determined equating slope of (3) and (10) at the point u_1^+, y_1^+ . The values of u_∞^+ are allowed to float in the exercise of obtaining a best fit of Coleman’s [2] experimental data with the theoretical curve. The exercise leads to an appropriate value 40 for u_∞^+ .

In this case, ξ is the non-dimensional maximum velocity and y_0^+ is the position of ξ . The results of three profiles are shown in Fig. 4 where a comparison between the theoretical profile and measured data is shown for the series of experiments with 0.105 mm sand. The profiles have been selected for the case of no seeding to the case of maximum seeding. The comparison between the theoretical curve

Fig. 5 ξ and u_1^+ as function of average concentration

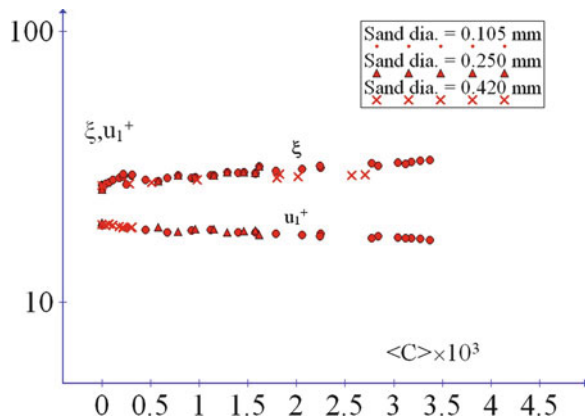
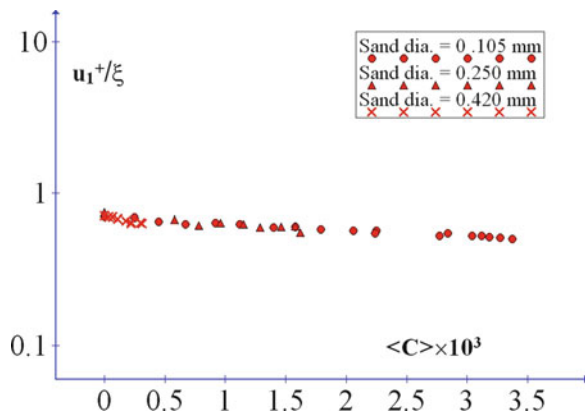


Fig. 6 The ratio between u_1^+ and ξ as a function of average concentration



and experimental data (Fig. 4) shows that most of the data lie in the wake region. Inspection of Fig. 4 indicates clearly that the portion of the near to wall region decreases with the increase of concentration.

For all cases correlation coefficient ' r ' is seen to be greater than 0.99, which justifies the use of the theoretical velocity profile derived for the single phase flow to the case of turbulent channel flow seeded with particles.

The values of ξ and u_1^+ are plotted against average concentration, $\langle C \rangle$ in Fig. 5. It reveals that ξ is an increasing function of concentration while u_1^+ is a decreasing function of concentration.

The values of the ratio u_1^+/ξ are plotted against average concentration in Fig. 6. It is a decreasing function of concentration. This indicates clearly that if one follows the path of velocity profile he will notice the shifting over of the flow within the boundary layer from the law of the wall to the law of the wake occurs early as average concentration increases i.e. contribution of wake region to total mass flux increases as average concentration increases.

Fig. 7 The values of y_1^+ and y_0^+ as function of average concentration

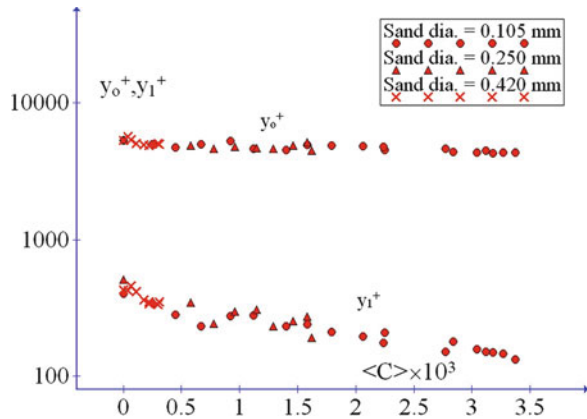
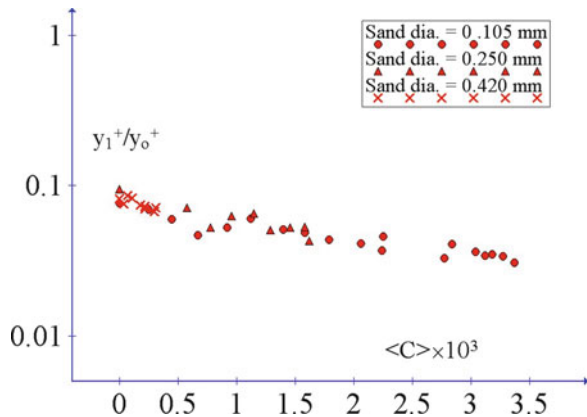


Fig. 8 The ratio between y_1^+ and y_0^+ as function of average concentration

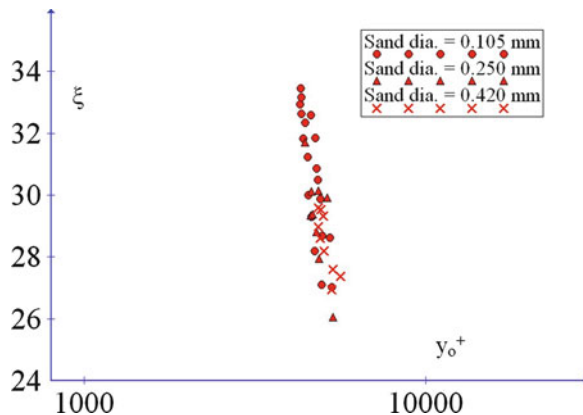


The values of y_1^+ and y_0^+ are plotted against $\langle C \rangle$ in Fig. 7. Both y_1^+ and y_0^+ decreases as concentration increases but at different rates. Figure 8 shows the ratio y_1^+/y_0^+ decreases as concentration increases. In fact, the ratio y_1^+/y_0^+ for each station is a measure of the non-dimensional quantity y/δ from the wall below which law of the wall (3), (4) governs the flow and above which law of the wake (10) takes over. The result may physically be interpreted that the near to wall region represents a decreasing portion of the total boundary layer as concentration increases.

The values of ξ are plotted against y_0^+ in Fig. 9. It is the locus of ξ under certain conditions. In other conditions, the locus of ξ will be different.

The values of the constant u_∞^+ turned out to be independent of the concentration distribution of the suspended sediment. The present examination shows that the influence of particle diameter on velocity profile appears to be insignificant.

Fig. 9 Semi-log plot of locus of ξ



5 Conclusion

We now discuss the results as follows:

- Flow field with sediment suspension are also describable through the inner variables.
- A model applicable for turbulent boundary layer, i.e. law of the wall (3), (4) and law of the wake (10) is found equally applicable for describing the channel flow seeded with particles. It is worth mentioning that in the present analysis, value of the viscosity as related to sediment concentration, i.e. relation (10) has been used.
- Shear velocity is a slowly decaying function of concentration.
- Maximum velocity of the flow, expressed in its non-dimensional form increases as sediment concentration increases, while, the non-dimensional wall distance at maximum velocity of the flow decreases as sediment concentration increases.
- With the increase of sediment concentration, the near to wall region is found to be a decreasing portion of the total boundary layer.
- With the increase of sediment concentration, the ratio between the mass flux in the wake region and the same occurs in the inner region increases.
- Maximum value of the wake (w_{\max}) increases with sediment concentration.
- Present investigation shows that sediment diameter has no significant influence on the velocity profile, but it has influence on sediment-carrying capacity of channel.

References

- Cioffi, F., Gallerano, F.: Velocity and concentration profiles of particles in a channel with movable and erodible bed. *J. Hydr. Res.* **29**, 387–401 (1991)
- Coleman, N. L.: (1981). Velocity profiles with suspended sediment. *J. Hydr. Res.* **19**(3), 211–229 (1981)
- Coleman, N. L.: Reply, *J. Hydr. Res.* **22**:4, 275–289 (1984)

4. Coleman, N. L.: (1986). Effects of suspended sediment on the open channel-channel velocity distribution. *Water Resources Research, AGU.* **22(10)**, 1377–1384 (1986)
5. Coles, D.E.: The law of the wake in the turbulent boundary layer. *J. Fluid Mech.* **1**, 191–226 (1956)
6. Coles, D. E., Hirst, E. A.: *Proc. Computation of Turbulent Boundary layers.* Stanford Univ. (1968)
7. Elata, C., Ippen, A. T.: (1961). The dynamics of open channel flow with suspensions of neutrally buoyant particles. Technical Report. Hydrodynamics Laboratory. MIT. **45**, (1961)
8. Graf, E. H.: *Hydraulics of sediment transport.* McGraw-Hill Book Co., Inc., New York (1971)
9. Gust, G.: Discussion: Velocity profiles with suspended sediment. *J. Hydr. Res.* **22:4**, 263–275 (1984)
10. Muste, M., Patel, V. C.: (1991). Velocity profiles for particles and liquid in open- channel flow with suspended sediment. *J. Hydr. Engr. ASCE.* **123(9)**, 742–751 (1991)
11. Parker, G., Coleman, N. L.: Simple model of sediment laden flows. *J. Hydr. Engr. ASCE.* **112(5)**, 356–374 (1986)
12. Persen, L. N.: The turbulent boundary layer and the closer problem. *Proc. AGARD Conf. on Turbulent Boundary Layer –Experiments, theory and modelling.* 17–1 to 17–14 (1974)
13. Persen, L. N., Coleman, N. L.: The influence of suspended particles on turbulent shear flow. *Turbulence Modification in Dispersed Multiphase Flow. FED.* **80**, 81–86 (1990)
14. Spalding, D. B.: A single formulae for the law of the wall. *J. Appl. Mech.*, **28**, Ser. E. 455–458 (1961)
15. Vanoni, V. A.: Transportation of suspended sediment by running water. *Trans. ASCE.* **111**, 67–133 (1946)

A Renormalization-Group Study of the Potts Model with Competing Ternary and Binary Interactions

Nasir Ganikhodjaev, Seyit Temir, Selman Uğuz, and Hasan Akin

1 Introduction

Consideration of spin models with multispin interactions has proved to be fruitful in many fields of physics, ranging from the determination of phase diagrams in metallic alloys and exhibition of new types of phase transition to site percolation. Systems exhibiting spatially modulated structures, commensurate or incommensurate with the underlying lattice, are of current interest in condensed matter physics [1]. Among the idealized systems for modulated ordering, the axial next-nearest-neighbor Ising (ANNNI) model, originally introduced by Elliot [2] to describe the sinusoidal magnetic structure of Erbium, and the chiral Potts model, introduced by Ostlund [3] and Huse [4] in connection with monolayers adsorbed on rectangular substrates, have been studied extensively by a variety of techniques. A particularly interesting and powerful method is the study of modulated phases through the measure-preserving map generated by the mean-field equations, as applied by Bak [5] and Jensen and Bak [6] to the ANNNI model. The main drawback of the method lies in the fact that thermodynamic solutions correspond to stationary but unstable orbits. However, when these models are defined on Cayley trees, as in the case of the Ising model with competing interactions examined by Vannimenus [10], it turns out that physically interesting solutions correspond to the attractors of the mapping. This simplifies the numerical work considerably, and detailed study of the whole phase diagram becomes feasible. Apart from the intrinsic interest attached to the study of models on trees, it is possible to argue that the results obtained on trees provide a useful guide to the more involved study of their counterparts on crystal lattices. The ANNNI model, which consists of an Ising model with nearest-neighbor interactions

N. Ganikhodjaev (✉)
Kulliyyah of Science, International Islamic University Malaysia,
25200 Kuantan, Malaysia
e-mail: gnasir@iiu.edu.my

augmented by competing next-nearest-neighbor couplings acting parallel to a single axis direction, is perhaps the simplest nontrivial model displaying a rich phase diagram with a Lifshitz point and many spatially modulated phases. There has been a considerable theoretical effort to obtain the structure of the global phase diagram of the ANNNI model in the $T - p$ space, where T is temperature and $p = -J_p/J_1$ is the ratio between the competing exchange interactions. On the basis of numerical mean-field calculations, Bak and von Boehm [7] suggested the existence of an infinite succession of commensurate phases, the so-called devil's staircase, at low temperatures. This mean-field picture has been supported by low-temperature series expansions performed by Fisher and Selke [8]. At the paramagnetic-modulated boundary analytic mean-field calculations show that the critical wave number varies continuously and vanishes at the Lifshitz point. A phase diagram of a model describes a morphology of phases, stability of phases, transitions from one phase to another and corresponding transitions line. A Potts model just as an Ising model on a Cayley tree with competing interactions has recently been studied extensively because of the appearance of nontrivial magnetic orderings (see [9–11, 14, 15, 19–22] and references therein). The Cayley tree is not a realistic lattice; however, its amazing topology makes the exact calculation of various quantities possible. For many problems, the solution on a tree is much simpler than on a regular lattice and is equivalent to the standard Bethe–Peierls theory [16]. On the Cayley tree one can consider two type of triple neighbors: prolonged and two-level (definitions see below). In the case of the Ising model with competing nearest-neighbor interactions J and prolonged next-nearest-neighbor interactions J_p Vannimenus [10] was able to find new modulated phases, in addition to the expected paramagnetic and ferromagnetic ones. From this result follows that Ising model with competing interactions on a Cayley tree is real interest since it has many similarities with models on periodic lattices. In fact, it has many common features with them, in particular the existence of a modulated phase, and shows no sign of pathological behavior – at least no more than mean-field theories of similar systems [10]. Moreover, detailed study of its properties was carried out with essentially exact results, using rather simple numerical methods. This suggest that more complicated models should be studied on trees, with the hope to discover new phases or unusual types of behavior. The important point is that statistical mechanics on trees involve nonlinear recursion equations and are naturally connected to the rich world of dynamical systems, a world presently under intense investigation [10]. The model (1) with $J_p = 0$ was considered in [20, 21] and proved that phase diagram of this model contains ferromagnetic and antiferromagnetic phases.

In this paper, we consider the Potts model with competing triple nearest-neighbor interactions. The Potts model [17] was introduced as a generalization of the Ising model to more than two components and encompasses a number of problems in statistical physics (see, e.g. [18]) recently. In [19], the phase diagram of the three states Potts model with nearest-neighbor interactions J and prolonged next-nearest-neighbors interactions J_p was described.

2 The Main Results

2.1 The Model

A Cayley tree Γ^k of order $k \geq 1$ is an infinite tree, i.e., a graph without cycles with exactly $k + 1$ edges issuing from each vertex. Let denote the Cayley tree as $\Gamma^k = (V, \Lambda)$, where V is the set of vertices of Γ^k , Λ is the set of edges of Γ^k . Two vertices x and y , $x, y \in V$ are called *nearest-neighbors* if there exists an edge $l \in \Lambda$ connecting them, which is denoted by $l = \langle x, y \rangle$. The distance $d(x, y)$, $x, y \in V$, on the Cayley tree Γ^k , is the number of edges in the shortest path from x to y . For a fixed $x^0 \in V$, we set

$$W_n = \{x \in V | d(x, x^0) = n\}, V_n = \{x \in V | d(x, x^0) \leq n\}$$

and L_n denotes the set of edges in V_n . The fixed vertex x^0 is called the 0-th level and the vertices in W_n are called the n -th level. For the sake of simplicity, we put $|x| = d(x, x^0)$, $x \in V$. Two vertices $x, y \in V$ are called *the next-nearest-neighbors* if $d(x, y) = 2$. Three vertices x, y , and z are called a triple of neighbors and they are denoted by $\langle x, y, z \rangle$, if $\langle x, y \rangle, \langle y, z \rangle$ are nearest neighbors. The triple of vertices x, y , and z is called *prolonged* if $x \in W_n, y \in W_{n+1}$ and $z \in W_{n+2}$ for some nonnegative integer n and is denoted by $\langle \widetilde{x}, y, z \rangle$. The triple of vertices $x, y, z \in V$ that are not prolonged is called *two-level* since $|x| = |z|$ and are denoted by $\langle x, \bar{y}, z \rangle$.

Below we will consider a semi-infinite Cayley tree Γ_+^2 of second order, i.e. an infinite graph without cycles with 3 edges issuing from each vertex except for x^0 which has only 2 edges.

For the three-state Potts model with spin values in $\Phi = \{1, 2, 3\}$, the relevant Hamiltonian with competing binary nearest-neighbor and ternary interactions has the form

$$H(\sigma) = -J_p \sum_{\langle x, y, z \rangle} \delta_{\sigma(x)\sigma(y)\sigma(z)} - J_1 \sum_{\langle x, y \rangle} \delta_{\sigma(x)\sigma(y)}, \quad (1)$$

where $J_p, J_1 \in \mathbb{R}$ are coupling constants and δ is the Kronecker symbol. Here, the generalized Kronecker's symbol $\delta_{\sigma(x)\sigma(y)\sigma(z)}$ is

$$\delta_{\sigma(x)\sigma(y)\sigma(z)} = \begin{cases} 1 & \text{if } \sigma(x) = \sigma(y) = \sigma(z) \\ 0 & \text{otherwise.} \end{cases}$$

Let $a = \exp(J_1/T); b = \exp(J_p/T)$, where T is the temperature.

The model (1) with $J_p = 0$ was considered in [12, 13, 20, 21] and proved that phase diagram of this model contains ferromagnetic and antiferromagnetic phases. Below we consider model (1) with $J_p \neq 0$, and describe its phase diagram.

2.2 Basic Equations

In order to produce the recurrent equations, we consider the relation of the partition function on V_n to the partition function on subsets of V_{n-1} . Given the initial conditions on V_1 , the recurrence equations indicate how their influence propagates down the tree. Let $Z^{(n)}(i, j)$ be a partition function on V_n with the configuration (i, j) on an edge $\langle x^0, x \rangle$, where $x \in W_1$ with $i, j = 1, 2, 3$ and $Z^{(n)}(i_1, i_0, i_2)$ be the partition function on V_n where the spin in the root x^0 is i_0 and the two spins in the proceeding ones are i_1 and i_2 , respectively. There are 27 different partition functions $Z^{(n)}(i_1, i_0, i_2)$ and the partition function $Z^{(n)}$ in volume V_n can be written as follows

$$Z^{(n)} = \sum_{i_1, i_0, i_2=1}^3 Z^{(n)}(i_1, i_0, i_2).$$

Then through a direct calculation one gets the following equalities:

$$\begin{aligned} Z^{(n)}(1, 1, 1) &= a^2 Z^{(n)}(1, 1) Z^{(n)}(1, 1), \\ Z^{(n)}(1, 1, 2) &= a Z^{(n)}(1, 1) Z^{(n)}(1, 2), \\ Z^{(n)}(1, 1, 3) &= a Z^{(n)}(1, 1) Z^{(n)}(1, 3), \\ Z^{(n)}(2, 1, 1) &= a Z^{(n)}(1, 2) Z^{(n)}(1, 1), \\ Z^{(n)}(2, 1, 2) &= Z^{(n)}(1, 2) Z^{(n)}(1, 2), \\ Z^{(n)}(2, 1, 3) &= Z^{(n)}(1, 2) Z^{(n)}(1, 3), \\ Z^{(n)}(3, 1, 1) &= a Z^{(n)}(1, 3) Z^{(n)}(1, 1), \\ Z^{(n)}(3, 1, 3) &= Z^{(n)}(1, 3) Z^{(n)}(1, 3), \\ Z^{(n)}(3, 1, 2) &= Z^{(n)}(1, 3) Z^{(n)}(1, 2), \\ Z^{(n)}(1, 2, 1) &= Z^{(n)}(2, 1) Z^{(n)}(2, 1), \\ Z^{(n)}(1, 2, 2) &= a Z^{(n)}(2, 1) Z^{(n)}(2, 2), \\ Z^{(n)}(1, 2, 3) &= Z^{(n)}(2, 1) Z^{(n)}(2, 3), \\ Z^{(n)}(2, 2, 1) &= a Z^{(n)}(2, 2) Z^{(n)}(2, 1), \\ Z^{(n)}(2, 2, 2) &= a^2 Z^{(n)}(2, 2) Z^{(n)}(2, 2), \\ Z^{(n)}(2, 2, 3) &= a Z^{(n)}(2, 2) Z^{(n)}(2, 3), \\ Z^{(n)}(3, 2, 1) &= Z^{(n)}(2, 3) Z^{(n)}(2, 1), \\ Z^{(n)}(3, 2, 2) &= a Z^{(n)}(2, 3) Z^{(n)}(2, 2), \\ Z^{(n)}(3, 2, 3) &= Z^{(n)}(2, 3) Z^{(n)}(2, 3), \\ Z^{(n)}(1, 3, 1) &= Z^{(n)}(3, 1) Z^{(n)}(3, 1), \\ Z^{(n)}(1, 3, 2) &= Z^{(n)}(3, 1) Z^{(n)}(3, 2), \\ Z^{(n)}(1, 3, 3) &= a Z^{(n)}(3, 1) Z^{(n)}(3, 3), \\ Z^{(n)}(2, 3, 2) &= Z^{(n)}(3, 2) Z^{(n)}(3, 2), \\ Z^{(n)}(2, 3, 1) &= Z^{(n)}(3, 2) Z^{(n)}(3, 1), \\ Z^{(n)}(2, 3, 3) &= a Z^{(n)}(3, 2) Z^{(n)}(3, 3), \\ Z^{(n)}(3, 3, 1) &= a Z^{(n)}(3, 3) Z^{(n)}(3, 1), \\ Z^{(n)}(3, 3, 2) &= a Z^{(n)}(3, 3) Z^{(n)}(3, 2), \\ Z^{(n)}(3, 3, 3) &= a^2 Z^{(n)}(3, 3) Z^{(n)}(3, 3). \end{aligned}$$

We can select only six variables $Z^{(n)}(1,1,1)$, $Z^{(n)}(2,1,2)$, $Z^{(n)}(3,1,3)$, $Z^{(n)}(1,2,1)$, $Z^{(n)}(2,2,2)$, $Z^{(n)}(3,3,3)$, and with the introduction of new variables

$$\begin{aligned} u_1^{(n)} &= \sqrt{Z^{(n)}(1,1,1)}, \quad u_2^{(n)} = \sqrt{Z^{(n)}(2,1,2)}, \\ u_3^{(n)} &= \sqrt{Z^{(n)}(3,1,3)}, \quad u_4^{(n)} = \sqrt{Z^{(n)}(1,2,1)}, \\ u_5^{(n)} &= \sqrt{Z^{(n)}(2,2,2)}, \quad u_6^{(n)} = \sqrt{Z^{(n)}(3,3,3)}, \end{aligned}$$

straightforward calculations show that

$$\begin{aligned} u_1^{(n+1)} &= a \left(bu_1^{(n)} + u_2^{(n)} + u_3^{(n)} \right)^2, \\ u_2^{(n+1)} &= \left(u_3^{(n)} + u_4^{(n)} + u_5^{(n)} \right)^2, \\ u_3^{(n+1)} &= \left(u_2^{(n)} + u_4^{(n)} + u_6^{(n)} \right)^2, \\ u_4^{(n+1)} &= \left(u_1^{(n)} + u_2^{(n)} + u_3^{(n)} \right)^2, \\ u_5^{(n+1)} &= a \left(u_3^{(n)} + u_4^{(n)} + bu_5^{(n)} \right)^2, \\ u_6^{(n+1)} &= a \left(u_2^{(n)} + u_4^{(n)} + bu_6^{(n)} \right)^2. \end{aligned} \quad (2)$$

The total partition function is given in terms of (u_i) by

$$Z^{(n)} = \left(u_1^{(n)} + u_2^{(n)} + u_3^{(n)} \right)^2 + \left(u_3^{(n)} + u_4^{(n)} + u_5^{(n)} \right)^2 + \left(u_2^{(n)} + u_4^{(n)} + u_6^{(n)} \right)^2. \quad (3)$$

Note that, for boundary condition $\bar{\sigma}_n(V \setminus V_n) \equiv 1$ we have $Z^{(n)}(2,1,2) = Z^{(n)}(3,1,3)$, and $Z^{(n)}(2,2,2) = Z^{(n)}(3,3,3)$, i.e., $u_2 = u_3$ and $u_5 = u_6$.

For discussing the phase diagram, the following choice of reduced variables is convenient:

$$\begin{aligned} x &= \frac{u_2 + u_4}{u_1 + u_6}, \\ y_1 &= \frac{u_1 - u_6}{u_1 + u_6}, \\ y_2 &= \frac{u_2 - u_4}{u_1 + u_6}. \end{aligned}$$

The variable x is just a measure of the frustration of the nearest–neighbor bonds and is not an order parameter like y_1 , and y_2 . Equations (2) yield:

$$\begin{aligned} x' &= \frac{1}{aD} [(2x + y_2 + 1)^2 + (y_1 + y_2)^2]; \\ y_1' &= \frac{2}{D} (2x + y_2 + b)(by_1 + y_2); \\ y_2' &= -\frac{2}{aD} (2x + y_2 + 1)(y_1 + y_2); \end{aligned} \quad (4)$$

where

$$D = (b + 2x + y_2)^2 + (by_1 + y_2)^2.$$

Below we use numerical methods to study its detailed behavior.

2.3 Morphology of the Phase Diagram

It is convenient to know the broad features of the phase diagram before discussing the different transitions in more detail. This can be achieved numerically in a straightforward fashion. The recursion relations (4) provide us the numerically exact phase diagram in $(T/J_1, -J_p/J_1)$ space. Let $T/J_1 = \alpha$, $-J_p/J_1 = \beta$ and, respectively, $a = \exp(\alpha^{-1})$, $b = \exp(-\alpha^{-1}\beta)$. Starting from initial conditions

$$\begin{aligned} x^{(1)} &= \frac{1 + a^2}{(a^3b^2 + a)}, \\ y_1^{(1)} &= \frac{a^2b^2 - 1}{a^2b^2 + 1}, \\ y_2^{(1)} &= \frac{1 - a^2}{(a^3b^2 + a)}, \end{aligned}$$

that corresponds to boundary condition $\bar{\sigma}^{(n)} \equiv 1$, one iterates the recurrence relations (4) and observes their behavior after a large number of iterations. In the simplest situation, a fixed point (x^*, y_1^*, y_2^*) is reached. It corresponds to a paramagnetic phase if $y_1^* = 0, y_2^* = 0$ or to a ferromagnetic phase if $y_1^*, y_2^* \neq 0$.

Secondary, the system may be periodic with period p , where case $p = 2$ corresponds to antiferromagnetic phase and case $p = 4$ corresponds to so-called antiphase, that denoted $< 2 >$ for compactness. Finally, the system may remain aperiodic. The distinction between a truly aperiodic case and one with a very long period is difficult to make numerically. Below we consider periodic phases with period p where $p \leq 12$. All periodic phases with period $p > 12$ and aperiodic phase we will consider as modulated phase. The resultant phase diagram is shown that in Fig. 1.

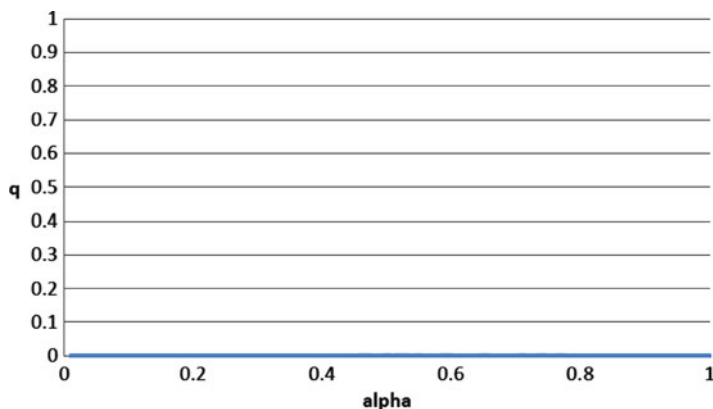


Fig. 2 Variation of the wavevector q versus, for $\beta = 0.2$

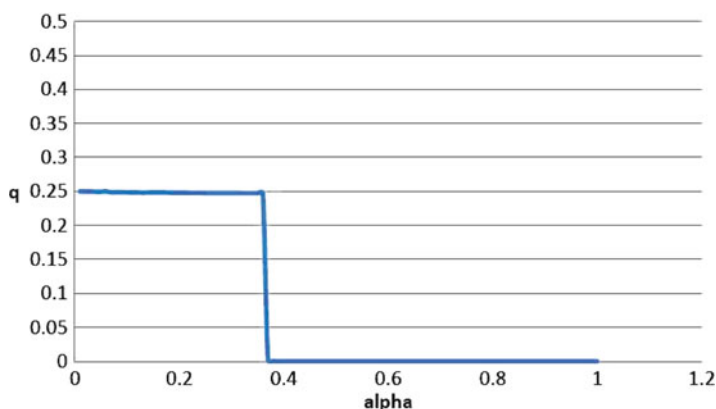


Fig. 3 Variation of the wavevector q versus, for $\beta = 0.8$

3 Conclusion

In [19] proved that the phase diagram of Potts model with competing nearest-neighbor and prolonged next-nearest-neighbor interactions consists of five phases: ferromagnetic, paramagnetic, modulated, antiphase, and paramodulated phases. We have found the phase diagram of the Potts model with competing prolonged ternary and binary nearest-neighbor interactions on the Cayley tree of second order and show that it consists of three phases only: ferromagnetic, paramagnetic, and antiphase with period 4. Thus for considered model, one can reach periodic phase with period 4 only, i.e., the set of modulated phases has simplest structure.

Acknowledgements The work is supported by The Scientific and Technological Research Council of Turkey-TUBITAK (Project No: 109T678).

References

1. P. Bak., Rep. Prog. Phys., 45, 587-629 (1982).
2. R.J. Elliott, Phys. Rev. 124, 346-353 (1961).
3. S. Ostlund, Phys. Rev. B, 24, 398-405 (1981).
4. D.A. Huse, Phys. Rev. B, 24, 5180-5194, 1981.
5. P. Bak, Phys. Rev. Lett., 46, 791-794 (1981).
6. M.H. Jensen and P. Bak, Phys. Rev. B, 27, 6853-6868 (1983).
7. P. Bak and J. von Boehm, Phys. Rev. B, 21, 5297-5308 (1980).
8. M.E. Fisher and W. Selke, Phys. Rev. Lett., 44, 1502-1505 (1980).
9. S. Coutinho, W. A. M. Morgado, E. M. F. Curado, L. da Silva, Physical Review B, 74(9): 094432-1-7 (2006).
10. J. Vannimenus, Z. Phys. B, 43, 141-148 (1981).
11. Mariz A.M., Tsallis C. and Albuquerque E.L., Journal of Statistical Physics, 40, pp. 577-592. (1985).
12. Monroe J.L., Journal of Statistical Physics, V. 67, pp. 1185-2000 (1992).
13. Monroe J.L., Physics Letters A, V. 88, pp. 80-84 (1994).
14. C.R. da Silva and S. Coutinho, Physical Review B, 34, pp. 7975-7985 (1986).
15. S. Inawashiro, C.J. Thompson, Physics Letters , 97A, 245-248 (1983).
16. Katsura, S., Takizawa, M., Prog. Theor. Phys., 51, 82-98 (1974).
17. R.B. Potts, Proc. Cambridge Philos. Soc., 48, 106 (1952).
18. F. Y. Wu, Rev. Mod. Phys., 54, 235-268 (1982).
19. N.N. Ganikhodjaev, F.M. Mukhamedov, C.H. Pah, Physics Letters A, 373, 33-38 (2008).
20. Ganikhodjaev N.N., Temir S. and Akin H., Cubo. A Mathematical Journal., 7, pp. 37-48 (2005).
21. Ganikhodjaev N.N., Akin H. and Temir S., Turkish Journal of Mathematics, 31, pp. 229-238 (2007).
22. Ganikhodjaev N.N., Temir S. and Akin H., Journal of Statistical Physics, 137, pp. 701-715. (2009).

The Mechanical Properties of CaX_6 ($\text{X} = \text{B}$ and C)

Sezgin Aydin and Mehmet Şimşek

1 Introduction

Graphite intercalation compounds (GICs) are produced by placing foreign atoms between two-dimensional sheets of graphite [1], and due to these sheets they possess interesting two-dimensional properties [2]. Therefore, these compounds have attracted a considerable interest and their physical properties were studied in a wide framework [3–5]. In alkali metal GICs, two-dimensional high conductivity is appeared through charge transfer from s-electron of intercalated atoms to carbon 2p π band of the graphene layers [2]. However, some of the alkali-metal GICs are superconductors (CaC_6) and generally their superconducting temperature is low (11.5 K) [6].

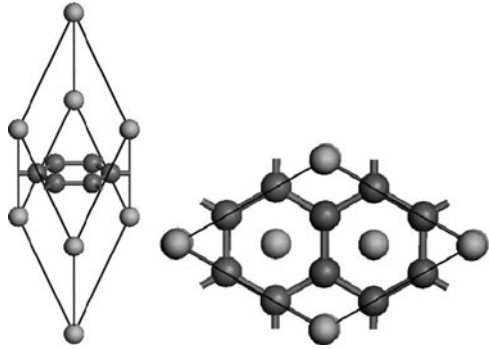
In this study, structural, electronic and mechanical properties (such as bulk modulus, shear modulus, Young modulus and hardness) of CaC_6 compound were investigated by using first-principle calculations within density functional theory. Replacing carbon atom to boron atom in the structure, it was studied how change the physical properties of the structure.

Crystal structure of CaX_6 compound is shown in Fig. 1 [7]. There is one Ca atom and six carbon atoms in the unit cell. Space group is $R\bar{3}m$ (No:166).

S. Aydin (✉)

Department of Physics, Gazi University, Teknikokullar, 06500, Turkey
e-mail: sezginaydin@gazi.edu.tr

Fig. 1 Crystal structure of rhombohedral CaC_6 . *Light grey and dark grey spheres stand for calcium and carbon atoms, respectively*



2 Computational Method

Our calculations are based on the first-principle plane-wave calculations. All calculations were performed by using CASTEP simulation package [8] within the density functional theory. In calculations, all the atomic coordinates and unit cell parameters were relaxed for minimization of Broyden, Fletcher, Goldfarb, Shanno (BFGS) scheme. The Vanderbilt ultrasoft pseudopotential [9] was used to model the ion–electron interactions, and exchange–correlation effects were treated within the generalized gradient approximation (GGA) by the Perdew–Burke–Ernzerhof (PBE) [10, 11] exchange correlation functional. The plane wave cut-off energy of 500 eV was employed, and the special k-points were generated by Monkhorst–Pack scheme, and they were chosen as $10 \times 10 \times 10$. For the convergence, the ultra-fine setup of software package was chosen, i.e. all calculations were converged in following qualities together (a) when the maximum ionic Hellman–Feynman force was below $0.01 \text{ eV}/\text{\AA}$, (b) maximum displacement was below $5.0 \times 10^{-4} \text{ \AA}$, (c) maximum energy change was below $5.0 \times 10^{-6} \text{ eV/atom}$ and (d) maximum stress was below 0.02 GPa.

After optimization of the unit cell, cohesive energy is calculated, $E_{\text{coh}} = (E_{\text{Total}} - E_{\text{ica}} - 6E_{\text{iX}})/n$, which is a key parameter to discuss energetic stability of the structure, where E_{Total} is total energy of the unit cell, E_{ica} and E_{iX} are energies of an isolated Ca atom and X atom, respectively. n is the total number of atoms in the unit cell.

Elastic constants are determined by using stress-strain method [12], and mechanical properties such as Bulk modulus, shear modulus and Young modulus are calculated as functions of elastic constants within Reuss–Voigt–Hill approximation [13]. For hexagonal/rhombohedral crystals, there are five independent elastic constants; c_{11} , c_{33} , c_{44} , c_{12} and c_{13} . Reuss (R) and Voigt (V) bulk modulus and shear modulus are given as,

$$\begin{aligned}
B_V &= \frac{1}{9} [2(c_{11} + c_{12}) + 4c_{13} + c_{33}], \\
G_V &= \frac{1}{30} [M + 12c_{44} + 12c_{66}], \\
B_R &= C^2/M, \\
G_R &= \frac{5}{2} (C^2 c_{44} c_{66}) / [3B_V c_{44} c_{66} + C^2 (c_{44} + c_{66})],
\end{aligned}$$

where the abbreviations are $C^2 = (c_{11} + c_{12})c_{33} - 2c_{13}^2$, $M = c_{11} + c_{12} + 2c_{33} - 4c_{13}$, and $c_{66} = (c_{11} - c_{12})/2$. While Reuss value is minimum limit, Voigt value is maximum limit for quantity. And, mechanical stability criteria are

$$c_{44} > 0, \quad c_{11} > |c_{12}|, \quad (c_{11} + 2c_{12})c_{33} > 2c_{13}^2.$$

Other hand, hardness is one of the important mechanical properties of a material, and can be calculated by Simunek's method [14]. In this method, hardness of a material is defined as

$$\begin{aligned}
H &= \frac{C}{\Omega} n \left[\prod_{i,j=1}^n N_{ij} S_{ij} \right]^{1/n} e^{-\sigma f_e}, \\
f_e &= 1 - \left[k \left(\prod_{i=1}^k e_i \right)^{1/k} / \sum_{i=1}^k e_i \right]^2,
\end{aligned}$$

where C and σ are constants, Ω is volume of unit cell, n is denoted the number of different bond type, k corresponds to the number of atoms in the system, and the number N_{ij} counts inter-atomic bonds in unit cell. S_{ij} is bond strength of individual bonds (d_{ij}) between atoms i and j defined as, $S_{ij} = \sqrt{e_i e_j} / (n_i n_j d_{ij})$; and $e_i = Z_i / R_i$, where Z_i is the valance electron number and R_i is atomic radius of the atom i . In calculation, $C = 1,450$ and $\sigma = 2.8$ values are used and the atomic radii for different elements are taken from Pearson text book [15].

3 Results and Discussion

The calculated structural parameters, elastic constants (c_{ij}), bulk modulus (B), shear modulus (G) and Young modulus (E) are listed in Table 1. Cohesive energies of both compounds are negative; thus, they are energetically stable. Cohesive energy of CaC_6 is smaller than that of CaB_6 . Therefore, we can expect that CaC_6 is more stable than CaB_6 . The calculated structural parameters for CaC_6 agree well with the literature, but calculated bulk modulus in this study is higher than the value of 103 GPa in [16].

Table 1 Calculated structural parameters, elastic constants (c_{ij}), bulk modulus (B), shear modulus (G) and Young modulus

	CaC ₆	CaB ₆
a (Å)	5.142	5.333
	5.170 [7]	
α (°)	49.40	56.00
	49.55 [7]	
V (Å ³)	71.939	97.354
E _{coh} (eV/atom)	−8.512	−5.763
c ₁₁	645	184
c ₃₃	84	88
c ₄₄	78	42
c ₁₂	88	132
c ₁₃	87	49
c ₆₆	278	26
B	147	90
	103 [16]	
G	122	35
E	287	93

However, all mechanical properties of CaC₆ listed in Table 1 are higher than those of CaB₆. This is an expected result, because CaC₆ is more stable than CaB₆. From calculated elastic constants, it was shown that CaC₆ and CaB₆ are mechanically stable. Other words, CaB₆ can be crystallized in CaC₆-type structure with lower mechanical properties.

Calculated band structure and density of states for CaX₆ compounds are shown in Figs. 2 and 3, respectively. It is shown from Figs. 2 and 3 that CaC₆ and CaB₆ have metallic character. From Fig. 3, X atoms are dominant on DOS. B p-orbitals possess more density of states than C p-orbitals. The hybridization between Ca s-orbitals and C p-orbitals is stronger than that of B-p orbitals.

Finally, hardness which is one of the important mechanical properties, is investigated. Micro-hardness for CaC₆ and CaB₆ is calculated as 13.4 GPa and 8.8 GPa, respectively. CaC₆ is harder than CaB₆. For CaX₆ compounds, it was shown that X–X bonds are stronger than Ca–X bonds. It is concluded that graphene layer in the structures plays important role on the mechanical properties.

In conclusion, the structural, electronic and mechanical properties of CaX₆ (X = B and C) were investigated by using first-principle calculations. From calculated cohesive energies and elastic constants, both compounds are energetically and mechanically stable. Our calculations showed that CaB₆ can be crystallized in CaC₆-type structure. However, graphene layer in the structures plays important role on the electronic and mechanical properties. From hardness analysis, CaX₆ compounds are hard material (not superhard).

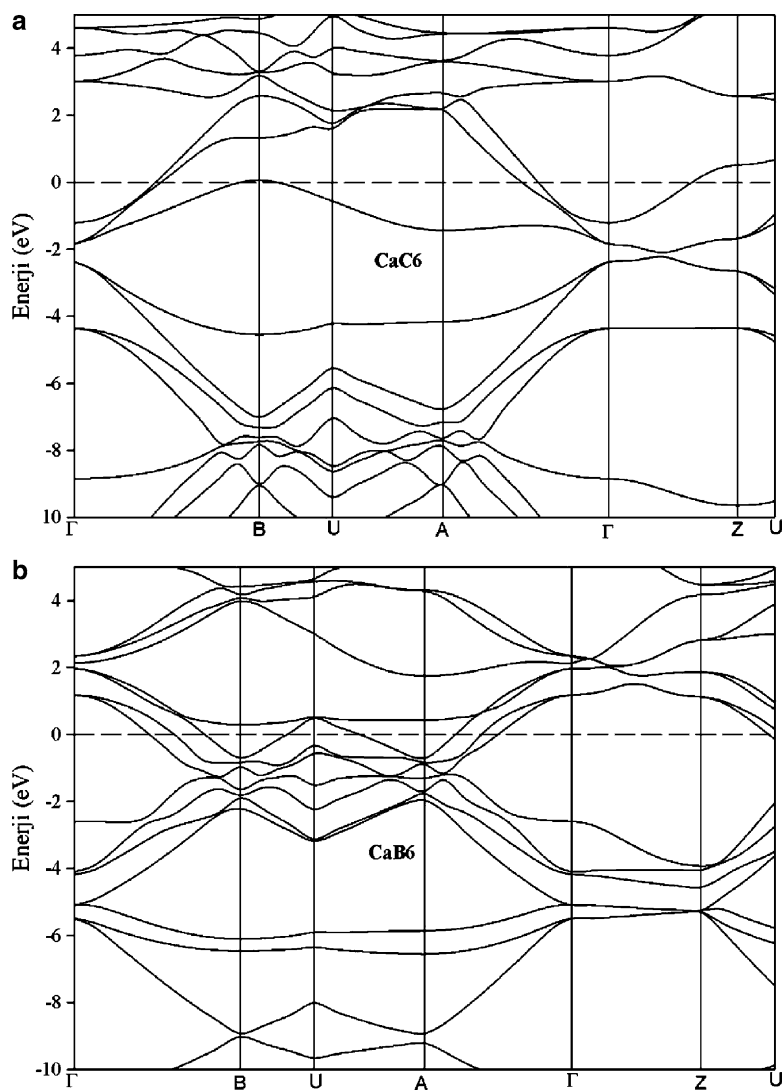


Fig. 2 Calculated band structure for (a) CaC_6 and (b) CaB_6

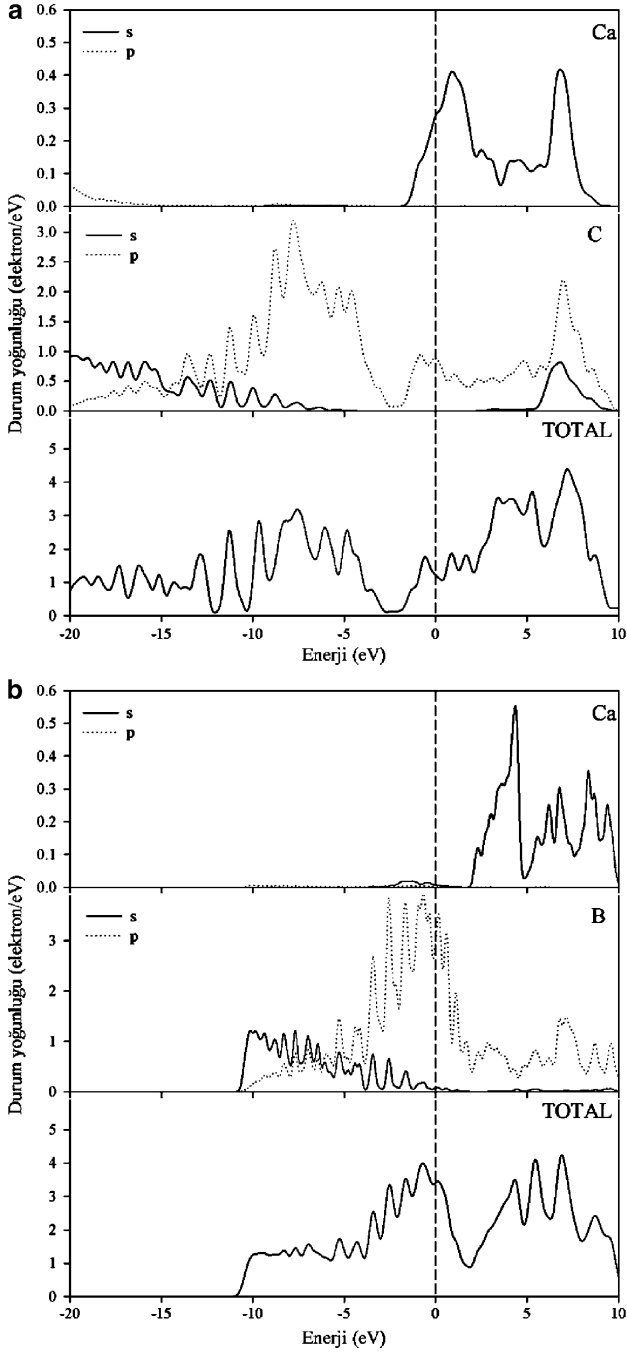


Fig. 3 Calculated partial density of states for (a) CaC_6 and (b) CaB_6

References

1. R. Singh, S. Prakash, Carbon 48, 1341–1344 (2010)
2. H. Okazaki et al., Physica C 469, 1041–1044 (2009)
3. M.S. Dresselhaus, D. Dresselhaus, Adv. Phys. 51, 1 (2002)
4. U. Mizutani et al., Phys. Rev. B 17, 3165 (1978)
5. S.L. Molodtsov et al., Phys. Rev. B 53, 16621 (1996)
6. T.E. Weller et al., Nat. Phys. 1, 39 (2005)
7. N. Emery et al., Solid State Sci 10, 466–470 (2008)
8. M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, M. C. Payne, J. Phys.: Cond. Matt., 14, 2717 (2002)
9. D. Vanderbilt, Phys. Rev. B, 41, 7892 (1990)
10. J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, Phys. Rev. B 46, 6671 (1992)
11. J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett., 77, 3865 (1996)
12. S.Q. Wu, Z.F. Hou, Z.Z. Zhu, Solid State Comm. 143, 425–428 (2007)
13. Zhi-Jian Wu, Er-Jun Zhao, Hong-Ping Xiang, Xian-Feng Hao, Xiao-Juan Liu, and J. Meng, Phys. Rev. B 76, 054115 (2007)
14. A. Simunek, Phys. Rev. B 75, 172108 (2007)
15. W. B. Pearson, The Crystal Chemistry and Physics of Metals and Alloys (Wiley, New York, p.151, Table 4-4 (1972)
16. J. S. Kim et al., Phys. Rev. B 74, 214513 (2006)

The System Design of an Autonomous Mobile Waste Sorter Robot

Ahmet Mavus, Sinem Gozde Defterli, and Erman Cagan Ozdemir

1 Introduction

Recycling is receiving increased attention in the environmental debate. This makes the first step of recycling, which is classification of waste materials, much more important. Manual sorting is obviously slow, tedious and unhealthy. Therefore, development of automatic techniques is inevitable. In their development, multidisciplinary approaches are applied because of great variety of parameters influencing the operation such as size, chemical properties, physical properties, etc.

First, while doing research for the general information about automatic waste sorting, some patents and commercial products are found in which more or less, the same techniques are applied. These existing automatic waste sorters are big plants including giant machines contrary to our case, which is a really little sorter. The sorting techniques used in these plants in a small-size sorter can be listed as follows: (for non-ferrous metals) eddy-current method, inductive sensors, electromagnetic and dual energy X-ray transmission sensor; (for ferrous metals) magnets; (for plastics) air classifiers, hydro cyclones, and float/sink baths, IR spectroscopy, selective dissolution, triboelectric charging; (for glasses) pneumatic separator. In addition to separation techniques how waste materials will be handled in operation, how they place in containers are crucial topics to search on. Moreover, energy management and automatic control systems are also important subjects to be investigated.

In this study, our design project is about recyclable waste sorting problem which is the subject of 2010 ASME Student Design Contest [1]. The aim is to design, build, and test a system capable of rapidly and accurately sorting the four waste

A. Mavus (✉)

Middle East Technical University, Department of Mechanical Engineering,
Ankara, 06531 Turkey

e-mail: ahmetmavus@hotmail.com

materials into distinct waste containers. This system must operate autonomously and be capable of both material identification and waste handling. A semi-rigid waste container (containing twelve waste products) is provided, specifically [1]:

1. Three empty plastic bottles $D = 75 \text{ mm } (\pm 20 \text{ mm})$, $L = 220 \text{ mm } (\pm 20 \text{ mm})$
2. Three empty aluminum cans $D = 65 \text{ mm } (\pm 20 \text{ mm})$, $L = 120 \text{ mm } (\pm 20 \text{ mm})$
3. Three empty steel containers $D = 75 \text{ mm } (\pm 20 \text{ mm})$, $L = 110 \text{ mm } (\pm 20 \text{ mm})$
4. Three empty glass containers $D = 60 \text{ mm } (\pm 20 \text{ mm})$, $L = 95 \text{ mm } (\pm 20 \text{ mm})$

1.1 Sorting and Working Principles

In the structure of the design, it is supposed to identify, handle and sort some kinds of wastes such as ferrous–non ferrous wastes, glass, and plastics. In order to do that, we use a kind of robot which has only one arm, one long hopper, one inductive sensor, one capacitive sensor, one magnetic switch, one encoder and Programmable Logic Controller (PLC) for control mechanism. These parts are connected to the main machine, and there are four bins for each type of waste. First, the capacitive sensor defines whether a waste exists or not. Then the inductive sensor and magnetic switch are used for detecting non-ferrous and ferrous wastes. The encoder is used for sorting glass and plastic bottles, since the capacitive gives sign “on”, but the inductive sensor and the magnetic switch give “off” signal, at this time encoder is used for counting steps while translational motion of robot passing the waste. The lengths of the plastic bottles are much larger than the lengths of glass bottles, so this remarkable difference can be used for sorting these bottles by counting step signals from encoder and separating them into different containers. Therefore, sorting logic is defined as follows:

Case 1: Capacitive sensor is “on”, inductive sensor is “on” but magnetic switch is “off” concludes that the waste is *non-ferrous metal*.

Case 2: Capacitive sensor is “on”, inductive sensor is “on” but magnetic switch is “on” concludes that the waste is *ferrous metal*.

Case 3: Capacitive sensor is “on”, inductive sensor is “off” but magnetic switch is “off” and encoder gives larger value than the limit concludes that the waste is *plastic material*.

Case 4: Capacitive sensor is “on”, inductive sensor is “off” but magnetic switch is “off” and encoder gives smaller value than the limit concludes that the waste is *glass material*.

The hopper is used for arranging wastes, the arm is for picking the wastes (cf. Fig. 1), and these four bins are lying through the translational motion of robot. At the beginning of the process, the wastes are poured onto hopper by hand. After finishing the pouring of the waste process, the robot starts to work.

In the beginning, the capacitive sensor gives the information that there is a waste here by giving signal “on”. At the same time, the inductive sensor and magnetic

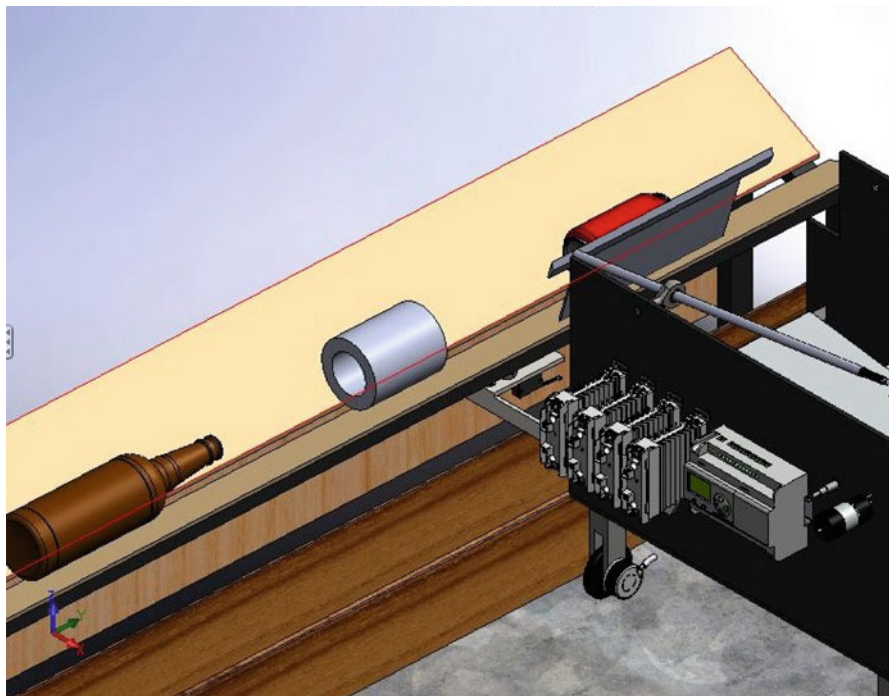


Fig. 1 The wastes on the V-shaped hopper and the robot

switch detect non-ferrous and ferrous metals. If inductive sensor gives “on” but magnetic switch gives “off”, this means that a non-ferrous waste exists here; on the other hand, both the inductive sensor and magnetic switch give “on” signals, this means that there is a ferrous waste here. After all wastes are detected, the robot arm begins to collect wastes, and put them into correct bin by the help of the sloped surface driven by linear actuator and geometric design of the robot. All wastes are collected by the robot arm. The robot arm picks these wastes into the machine one by one. At the end, the wastes are put into their containers, and the operation of the machine is completed.

2 Properties of the Designed System

First of all, this machine is considered as a robot and mainly, the system has 3 degrees of freedom. One of them is the translational motion of the robot along the hopper. The second one is related to the arm mechanism to be used for the process of taking the wastes from hopper to the inside of the robot to separate. The third one is the translational motion of the turning block mechanism for the sloped surface driven by the linear actuators.

2.1 *The Degrees of Freedom*

If it is needed to be more precise, a detailed explanation can be given about the degrees of freedom of the mechanism listed above. First, the robot moves to the right or to the left by the aid of a strong direct current (DC) motor. The necessary torque is calculated afterwards. Specifically, this torque is calculated by putting the friction coefficient of the ground and weight of the robot (all parts inside the robot plus the part on the robot and the skeleton) into the perspective. According to this torque value, a suitable DC motor is selected.

The other important point of the arm mechanism of the robot is that the selection of the dimensions of the links of the inverted slider-crank mechanism as an arm. These links of the mentioned mechanism are selected for mainly carrying the waste with the maximum weight. However, in addition to that, selected links should not have contact with the sloped surface. Since the wastes are taken one by one, the heaviest parts are the ferrous wastes (steel parts), which are dominant in this process considering the durability of the mechanism. The third degree of freedom is about the motion of turning block mechanism of the sloped surface. This motion is used as sorting glass and plastic wastes. When the capacitive sensor gives signal as “on”, but the inductive sensor and the magnetic switch give signal as “off”, this means that the type of waste recognized is either plastic or glass. Since it is known that the length of the plastic wastes are much larger than the length of glass wastes, this remarkable difference is used for sorting by counting the step signals given from the encoder which is placed on the shaft of the driven wheel of the robot. So, the process of sorting the plastic and glass bottles are done by this principle.

2.2 *The Hopper Design*

The hopper design is also crucial for the system. Since it carries all the wastes, it should have the necessary strength. Moreover, its length is determined by adding the maximum lengths of the wastes plus a clearance value so that when they are poured to the hopper, they are arranged along the length of the hopper and they do not overlap. This is very important when identification process is put into the perspective. However, it should also be noted that its length may be 2 m maximum. Hence, overlap actually is inevitable.

For the hopper, the following conditions have to be considered:

- (a) It should be wide enough not to let waste to fall down during pouring.
- (b) It should be in a geometry not disturbing the arm.
- (c) It should have enough alignment to guide waste in appropriate manner.

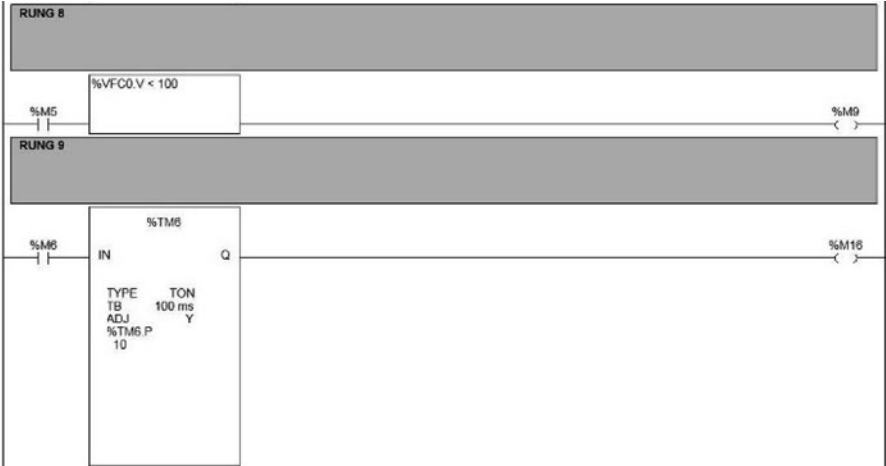


Fig. 2 A part of case definitions in the ladder diagram of PLC algorithm

2.3 The Control Mechanism

The control of the slider crank arm mechanism, driven wheel DC motor, arm motor, encoders, linear actuator is achieved by PLC. PLCs provide ease and flexibility of control based on programming and executing simple logic instructions (often in ladder diagram form) [2].

For the inductive sensor and capacitive sensor which are used in our mechanism, the range of sensor is not significant since its head is on the underneath of the long hopper part. Therefore, accuracy may be increased. The key point is actually when using sensors, motors, encoder and linear actuator, computer interaction should be achieved. To be more specific, our system includes a main computer and all the motors, sensors, encoder, magnetic switch and linear actuator are controlled by the main controller that is PLC. In order to connect motors and sensors to the computer, PLC is easy to implement and allows us to do any changes in the algorithm. For this reason, these kinds of control systems or derivatives can be used to obtain accurate results.

PLC algorithm is written by first considering the cases like if capacitive sensor is “on”, the inductive sensor is “on” but the magnetic switch is “off”, the case is that the waste is non-ferrous material such as called “M2”. After defining the cases as mentioned above, the actions are taken into consideration in the second part of the algorithm. All these cases are defined in the ladder diagram where a part of it can be seen in Fig. 2.

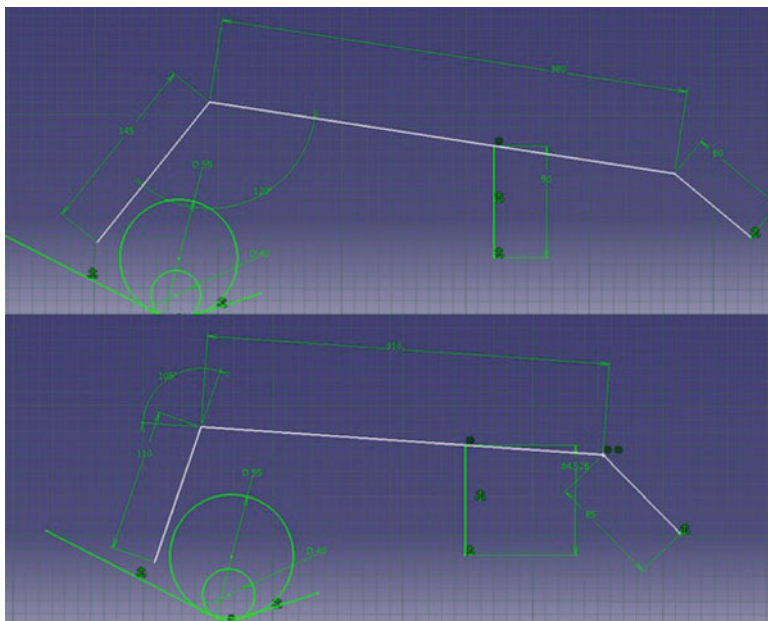


Fig. 3 Geometric design of the inverted slider crank arm mechanism

2.4 *The Motions of the Inverted Slider Crank Arm Mechanism*

The slider crank arm mechanism has two different motions such as inward and outward motion. These types of motions are done according to the detection of the waste material. If the waste material is plastic, the arm mechanism does outward motion and if the waste material is non-ferrous and glass, the arm does the inward motion to the robot. When the material of the waste is detected as ferrous, then arm does not move and the ferrous waste stays on the hopper since the weight of ferrous wastes can damage the arm mechanism, which is undesirable.

For the arm, the items given below have to be satisfied:

- It should, in a position in its turn, lay within the given geometric restrictions (see Fig. 3).
- It should cover whole waste range, not skipping some small wastes.
- It should not, in a position in its turn, intersect with other components.

2.5 *Motor Selection*

The choice of the suitable motors is also vital. The most critical motor is the one located under the base plate to drive the robot. It is chosen by considering the total

torque acting on the wheels. The strongest motor is the driving motor as stated at the beginning of the mechanism. There is also other motor namely, the DC motor actuating the arm mechanism, this motor does not have to apply huge torques as the main driving motor since the weight of the wastes is very small when compared to the total weight of the robot.

3 Engineering Calculations of the Geometric Design

3.1 The Inverted Slider Crank Arm Mechanism

The arm mechanism (as seen in Fig. 3) is the most crucial part of the robot since it is used for taking the detected wastes and sending them to the correct container. For the arm of the robot, four-bar and inverted slider crank are preferred mechanisms and the inverted slider crank is the first candidate since its motion curve seems more appropriate [3].

In Excel, as seen in Table 1 and Fig. 4, parametric model of inverted slider crank is created. By changing the values of parameters, a suitable motion curve is generated. Since there is no limitation about the design of the long hopper, it is achieved according to the design requirements of the arm mechanism.

3.2 Turning Block Mechanism for the Sloped Surface Driven by the Linear Actuator

This mechanism is used for separating glass bottles into their specified container and driven by a linear actuator which is controlled by PLC according to taken data from sensors and encoders. If the control logic, which is mentioned before for the glass bottles is satisfied, then linear actuator becomes active and turning block mechanism for sloped surface moves toward to the long V-shaped hopper. The arm mechanism takes the glass on the sloped surface and the glass bottle moves along the surface and falls into the corresponding container. The geometric design

Table 1 Position analysis of arm mechanism in Excel

		Links	x (mm)	y (mm)	Prismatic joint	
x_{fixed}	345.4	–	0	0	–354.995	97.007
y_{fixed}	70.4	A_0	40.358	–110.884	–318.793	79.995
Crank	118	A	–345.4	70.4	–335.805	43.793
Slider	495	B_0	–407.639	99.648	–372.007	60.806
Coupler arm	105	B	–488.274	32.396	–354.995	97.007
Force angle α_3	65	C	–457.556	58.016	–	–

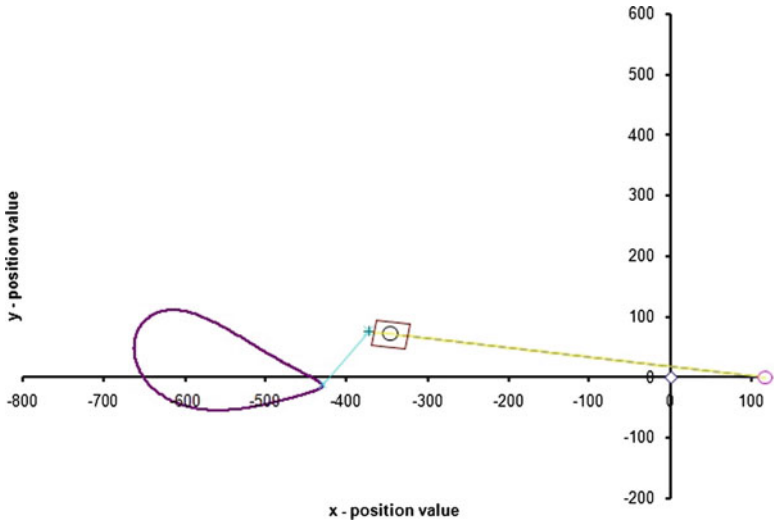


Fig. 4 Kinematic modelling of the arm mechanism

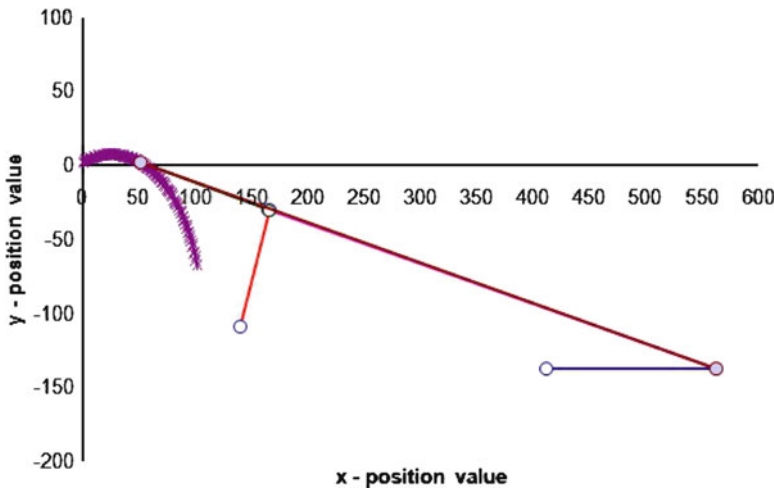


Fig. 5 Visualization of the turning block mechanism after kinematic synthesis

of turning block mechanism for the sloped surface is achieved according to hopper and arm mechanism’s design in order to get a high accuracy. Thus, the motion analysis of turning block mechanism for the sloped surface is done in Excel Macro and presented in Fig. 5. Moreover, Fig. 6 shows the side view of the turning block mechanism in the robot.

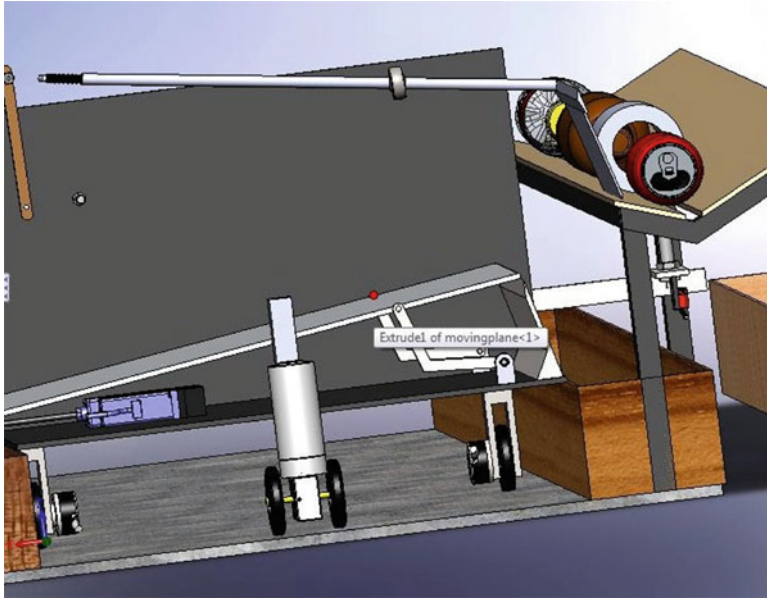


Fig. 6 Side view of the turning block mechanism

3.3 Shaft Design of Wheels

Considering the construction materials and the control components of the robot, the total weight is calculated a 300 N. Focusing on all five wheels of the robot, the static analysis gives that the most critical force exerts on the wheel shaft is 90 N. Also, the shaft design is performed according to this critical force value.

$$F_w = 90 \text{ N}, \quad d_s = 5.25 \text{ mm}, \quad L = 19.56 \text{ mm}, \quad f = 0.5, \quad d_w = 62 \text{ mm}.$$

The material properties are taken as: AISI 1080

$$S_{ut} = 615.4 \text{ MPa}, \quad S_y = 375.8 \text{ MPa}, \quad E = 200 \text{ GPa}.$$

(S_{ut} : Ultimate Tensile Stress, S_y : Yield Stress, E : Modulus of Elasticity)

After determining the torque, moment and direct force on each shaft on the wheels, bending and shear stresses are evaluated. Von Misses Stresses are evaluated to find alternating and mean components of the shear and bending stresses. Moreover after determining the design factors (size, temperature, reliability, and so on), the process for finding safety factor begins. The corresponding formulas are given in the following equations [4].

$$\sigma_{a1} := \sqrt{\sigma_{abending}^2 + 3\tau_a^2} \quad (1)$$

$$\tau_m := 16 \frac{T_w}{\pi d_s^3} \quad (2)$$

$$\sigma_m := 0 \quad (3)$$

$$\sigma_{m1} := \sqrt{\sigma_m^2 + 3\tau_m^2} \quad (4)$$

$$S_{e1} := 0.5S_{ut} \quad (5)$$

The endurance strength limit is calculated as:

$$S_e = k_a \cdot k_c \cdot k_d \cdot k_f \cdot S_{e1} \quad (6)$$

$$S_e = 1.853 \times 10^2 \text{ MPa} \quad (7)$$

$$n := \frac{1}{\left(\frac{\sigma_{a1}}{S_e} + \frac{\sigma_{m1}}{S_y} \right)} \quad (8)$$

At the end, it is defined that the safety factor, which is denoted by n , should be at least 2 and after calculations are done we obtained that $n = 2.542$.

Hence, this design for shafts of wheels is safe enough to satisfy the needs of the design.

3.4 The Design of Structural Members

The structural members carry sweeper motor for arm mechanism, spherical bearing connection, PLC controller box and relays. Since the exact location for the arm connection point to the structural members should be determined in detail, more durable body frame is satisfied and space minimization is provided, it is better to use special designed laser cut 5 mm thick sheet metal.

4 Conclusion

Before deciding to use this design, it is thought about lots of different designs workability. At the beginning of the concept selection, there are lots of alternative concept designs. First, it is tried to use some mechanical ways to separate the wastes. But in order to use these mechanical ways, huge amount of space for the machine is needed and large amount of energy is required for running these mechanical systems. According to these design specifications, the space is strictly limited, and it is only permitted to use dry cells as energy sources in the system so it should

be better to give up thinking of usage of mechanical ways to separate the wastes. Second, it is tried to use some piston arrangement to separate the wastes but in this design, it is needed lots of small pistons, and compressors to run these pistons, also the system needs lots of energy to run these pistons. It is known that dry cells are not powerful enough to run these compressors, also compressors occupy large amount of space. Another disadvantage of that system is that these pistons are hard to find, and it is too expensive. Finally, we tried to find a system which occupies small amount of space and also economic. In order to fulfill these conditions, we decided to use a kind of robot.

In a nut shell, the robot has some specialties. These specifications are weight, time, energy, user friendliness, accuracy and space occupation [5]. It is chosen that the materials of the system with considering these specifications. There are a few motors in this system. These motors should not be heavy, should occupy small amount of space, and should consume small amount of energy and should satisfy torque and revolution requirements. Also, the process has to be completed in five minutes so that the robot should make the process quick. Our system has to be user friendly. In our design, the user will only pour the waste, push the button, and wait until the wastes separated. The system has to occupy small amount of space. A robot can easily fulfill all of these specifications. Further effort is going to be made in order to apply for patent of this design project and to provide automation of the robot for industrial needs. Since this autonomous mobile waste sorter robot design provides an innovation to the sorting problem in recycling industry, we hope that this new and original project is going to be patented.

Acknowledgements This work is partially supported by the Scientific and Technical Research Council of Turkey.

References

1. ASME (2010) Problem Description. In: The 2010 ASME Student Design Competition - EARTH SAVER: Autonomous Material Sorter. <http://files.asme.org/asmearg/Events/Contests/DesignContest/18197.pdf>.
2. Warnock IG (1988) Programmable controllers: Operation and Application. Prentice Hall, London, UK
3. Soylemez E (1999) Mechanisms. Middle East Technical University Press Cooperation, Ankara, Turkey
4. Budynas RG, Nisbett JK (2006) Shigley's Mechanical Engineering Design, 8th edition. McGraw-Hill, Boston MA
5. Dieter GE (2000) Engineering Design, 3rd edition. McGraw Hill, Boston MA

Nomenclature

d :	Diameter (mm)
d_s :	Diameter of the Shaft (mm)
d_w :	Diameter of the Wheel (mm)
L :	Length (mm)
A :	Cross Sectional Area (mm ²)
E :	Modulus of Elasticity (GPa)
S_y :	Yield Strength (MPa)
S_t :	Tensile Strength (MPa)
S_{ut} :	Ultimate Strength (MPa)
S_e :	Endurance Limit (MPa)
n :	Safety Factor
k :	Yield Strength Modification Factor
k_a :	Surface Condition Modification Factor
k_c :	Load Modification Factor
k_d :	Temperature Modification Factor
k_f :	Miscellaneous Effect Modification Factor
F :	Force (N)
F_w :	Force on the Wheel (N)
f :	Friction Coefficient
M :	Moment ($N.m$)
T :	Torque ($N.m$)
σ :	Tensile Stress (MPa)
σ_m :	Mean Tensile Stress (MPa)
σ_a :	Alternating Tensile Stress (MPa)
τ :	Shear Stress (MPa)
τ_m :	Mean Shear Stress (MPa)
τ_a :	Alternating Shear Stress (MPa)

Evidence of the Wave Phase Coherence for Freak Wave Events

Alexey Slunyaev

1 Introduction

Irregular waves arise in many important physical problems and are the subject of investigation. The study of irregular wave dynamics and statistics is relevant for correct physical understanding and for practical applications as well. Sea waves are an example of inherently stochastic waves. They are often understood as a combination of quasi-sinusoidal waves with independent random uniformly distributed phases (the Gaussian sea). If waves were linear and random, they would possess the Gaussian probability distribution function due to the central limit theorem.

The difference between the Gaussian sea approximation and the real sea results in recognizing the problem of *freak wave* or *rogue wave* phenomenon, see reviews [12, 19, 20]. The attempts to integrate the wave nonlinearity effect into the statistical models have been undertaken many times with a certain success. However, each time these endeavors employ the condition of weak nonlinearity, which is questionable when applied to freak wave events. Meanwhile, obtaining the statistical description of freak waves for the case of a given wave energy spectrum is the cherished aim of the researchers.

Kinetic approach is conventional and well established for the study of random wave spectrum evolution. The kinetic theory is weakly nonlinear, uses closure assumptions, thus, is eventually capable of describing only near-Gaussian processes, and it disregards wave phases. The stochastic approach employs dynamical models, which resolve the wave phases. Then the relations between the spectral and

A. Slunyaev (✉)

Institute of Applied Physics, Nizhny Novgorod, 603950, Russia

Keele University, Keele, ST5 5BG, UK

e-mail: Slunyaev@hydro.appl.sci-nnov.ru

statistical wave characteristics may be established. Instead of computing the kinetic equations, the stochastic approach has become very popular due to computer power progress and building large and well-equipped experimental facilities.

The stochastic modeling requires definition of the initial wave field. Typically, the initial condition is defined in the form of wave fields obeying some spectrum with random uniformly distributed phases. A great number of numerical and laboratory experiments prove that this initial condition undergoes strong evolution at the initial stage, what changes the average spectrum of the waves, see [5–7, 14, 16, 21–25, 27–29, 32] among others. Thus, the wave fields at this stage cannot be considered statistically stationary. Limited fetches (due to the limitation of the laboratory facility or short numerical simulations) may prevent the achievement of the stationary state at all.

The most striking nonlinear effect, which is now believed able to cause rogue waves is the Benjamin-Feir instability (otherwise, the side-band or the modulational instability, see for instance [17, 36]), which leads to the generation of intense wave groups from uniform wave trains in deep-water conditions. This effect for unidirectional waves has been confirmed many times by means of numerical simulations, and also by laboratory measurements. Meanwhile, it is known to be weakened and even cancelled, when broad-band waves or random waves are concerned.

The Benjamin-Feir index (BFI) was introduced by Onorato et al. [22] and Janssen [16] to measure the strength of the nonlinear self-modulation effects for a given wave energy spectrum. This compound spectral parameter is in agreement with its dynamic counterpart, which follows from the weakly nonlinear weakly modulated theory for surface waves (the framework of the nonlinear Schrödinger equation, NLS). Then the BFI corresponds to the similarity parameter of the NLS equation, otherwise, the soliton number.

The NLS equation is a unique mathematical model due to the property of integrability. Its solutions have been suggested to describe real freak waves in the ocean [2, 13, 15, 26]. The so-called *breather* solutions of the NLS equation are actually solitary waves interacting with other background waves. The similarity of the breather solutions and the large-wave events observed in numerical simulations has been pointed out many times [8, 9, 15]. The existence of long-living strongly nonlinear wave groups similar to the envelope solitons has been reported recently for numerical simulations of fully nonlinear equations for hydrodynamics [10, 30].

The framework of the NLS equation for unidirectional waves requires the conditions of weak nonlinearity and narrow spectrum. In the general case, the solitary-like patterns are supposed to show themselves as coherent wave structures, what implies nonzero correlation between the Fourier modes. The dynamics of a single four-wave resonance quartet was studied within the framework of the Zakharov equation by means of the analytic solution in [18, 33]; in [33] ensembles of the wave quartets were considered as well and compared versus the results of the kinetic approach. It was shown, in particular, that initially random phases develop a significant coherence in the course of evolution. Specific spectrum shape was suggested by [1] as a possible early-warning criterion for the rogue wave danger.

In this paper, we deal with the deep-water limit, and the unidirectional case of surface sea waves, which is the most investigated. The wave trains are generally supposed narrow-band; the physics is governed by the free wave components, while the bound waves (Stokes corrections) under the narrow-band assumption may be trivially obtained on the basis of the free waves.

The coherence might be revealed in the dynamics of the strongly interacting components (when they may be singled out) such as resonance quartets or soliton-like wave groups. We show that the wave coherence can be revealed globally in a stochastic wave field. A more detailed description of the background of the present study is given in [31].

2 Stochastic Numerical Simulations of Modulationally Unstable Wave Fields

In this section, we summarize the results of numerical simulations, performed within the frameworks of the nonlinear Schrödinger equation, its high-order generalization (the Dysthe equation with the exact deep-water linear dispersion law taken into account [11, 34]), and the fully nonlinear simulations of the Euler equations in conformal variables [35]. The algorithms are briefly described in [30]. Nonbreaking waves are considered; 100 wave ensembles are used for the averaging.

In all cases, the initial wave realizations are defined in the form of a linear superposition of Stokes waves with random phases, similar to those described in [28], which obey the Gaussian spectrum. The carrier wavenumber is chosen $k_0 \approx 1.78 \text{ m}^{-1}$. Due to the deep-water conditions, the mean wave period is estimated through the linear dispersion law as $T_0 = 1.5 \text{ s}$. The NLS model is used to solve waves with a moderate initial steepness, $k_0 \eta_{\text{rms}} \approx 0.042$ (where η_{rms} is the root-mean-square surface wave displacement defined on the basis of the free wave component) for different initial spectrum widths. The fully nonlinear simulations are performed for one initial spectrum width, $\nu/k_0 \approx 0.076$, where ν is defined as the second moment of the average spectrum. In terms of the introduced parameters, the BFI may be defined as [16]

$$\text{BFI} = 2\sqrt{2}k_0^2 \frac{\eta_{\text{rms}}}{\nu}, \quad (1)$$

and the characteristic nonlinear time we define as

$$T_{\text{nl}} = \frac{1}{\omega_0 k_0^2 \eta_{\text{rms}}^2}. \quad (2)$$

In the course of evolution the spectrum width, ν and the BFI evolve. The variable BFI for the simulations is shown in Fig. 1 as the function of the scaled time. The results of the NLS simulations are represented by solid lines, and the results of the fully nonlinear simulations are represented by dashed lines.

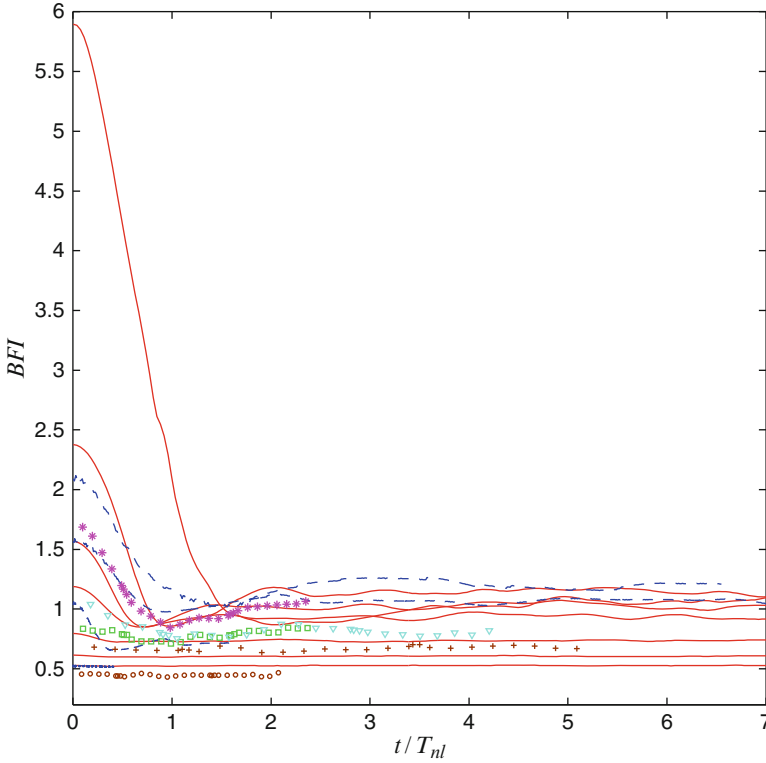


Fig. 1 The temporal dependences of the BFI versus the scaled time for numerical simulations (lines) with different initial spectrum widths and wave intensities, and for the laboratory measurements [29] (symbols) with different spectral widths and shapes

The results of laboratory measurements [29] are given in Fig. 1 with symbols. In the laboratory experiments, the spatial wave evolution was considered and surface elevation time series were retrieved at different distances. The distances are recomputed to the corresponding times in Fig. 1 supposing that the waves propagate with the group velocity of the carrier wave.

It is evident that the curves corresponding to the two sets of numerical simulations and the results of laboratory measurements well agree when represented in the scaled variables. For sufficiently long times kinds of steady states are achieved in all the cases.

The Alber theory for narrow-band weakly nonlinear random waves [3, 16] predicts the cancellation of the Benjamin-Feir instability effect for $BFI < 1$. It may be seen in Fig. 1 that this threshold describes well the qualitative difference in the evolution of the BFI. The wave fields with initially large values of the BFI tend to the state, which seems to be marginally stable. The wave fields with small values of the BFI remain practically unchanged during the evolution.

3 The Evidence of Phase Coherence

The spectral phases are implied by the concept of the Gaussian sea to be uniformly distributed. Indeed, the phase distribution observed in our simulations may be considered uniform (see details in [31]). However, the nonlinear wave phases are obviously not fully independent as it is in the linear approximation.

The phase coherence supports existence of coherent wave structures. Some exact solutions of the NLS equations are discussed in this context in [1, 31]. When many nonlinear coherent structures are present in the wave field, the wave dynamics is supposed to be quite complicated.

We apply the following correlation function to make the coherence between the Fourier phases evident:

$$\begin{aligned}
 R(\delta, t) &= \frac{R_1}{R_2}, \\
 R_1 &= \left| \sum_{n=1}^N S_n(k_0 + \delta) S_n(k_0 - \delta) S_n^*(k_0 + \delta') S_n^*(k_0 - \delta') \right|, \\
 R_2 &= \sum_{n=1}^N \left| S_n(k_0 + \delta) S_n(k_0 - \delta) S_n^*(k_0 + \delta') S_n^*(k_0 - \delta') \right|, \\
 \delta &= \frac{2\pi}{L} m, \quad \delta' = \frac{2\pi}{L} (m + 1), \quad m \geq 0.
 \end{aligned} \tag{3}$$

In (3) k_0 is the dominant wavenumber, δ and δ' specify the wavenumber offsets according to the wavenumber discretization (integer m counts the spectral nodes), L is the computational domain length. The summation is performed over all N realizations.

The results of the computation of the autobicoherence function (3) for the conditions $k_0 \eta_{\text{rms}} \approx 0.042$, $v/k_0 \approx 0.076$ (and $\text{BFI} \approx 1.56$) are presented in Fig. 2 by the colour intensity. Different models are simulated: the NLS equation (Fig. 2a), the Dysthe equation (Fig. 2b), and the Euler equations in conformal variables (Fig. 2c). The horizontal axis shows the normalized time, the vertical axis represents the wavenumber offset. The temporal dependence of the average spectrum width v is given over the diagram (the solid line), and the spectrum shape at $t = 0$ for $k \geq k_0$ is shown to the right of the diagram (the line with circles); both are given for the reference.

The case of the NLS equation simulation is shown in Fig. 2a. It is obvious that the initial condition corresponds to an insignificant correlation (dark area near $t = 0$). But with time the correlation for a sufficiently large offset quickly grows up to the unity. While the Fourier modes which are sufficiently far from the spectral peak are thus shown to be absolutely correlated, the most energetic area of the spectrum close to the spectral peak is uncorrelated (at least with respect to the autobicorrelation function (3)).

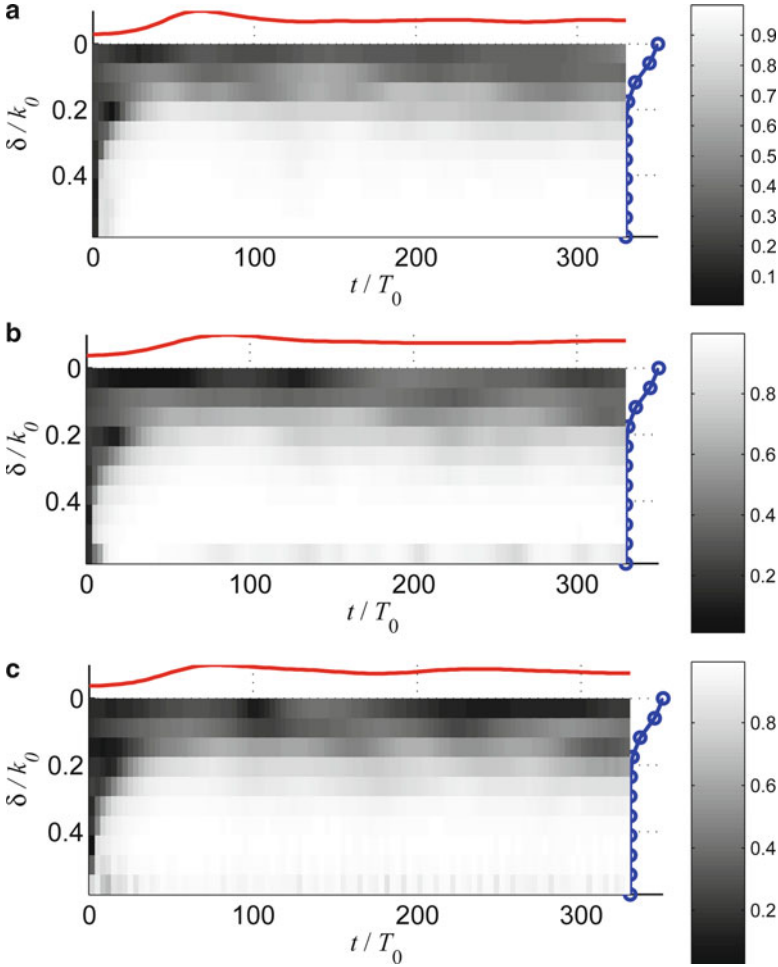


Fig. 2 Diagrams of the correlation estimator $R(\delta, t)$ shown by the color intensity. The temporal dependence of the average spectrum width is given over the diagram (*the solid line*) for the reference, and the spectrum shape at $t = 0$ is shown to the right of the diagram (*the line with circles*) for the reference. The initial condition for simulation $\text{BFI}(t = 0) = 1.56$ is computed in different frameworks: the NLS equation (a), the Dysthe model (b), and the fully nonlinear simulations (c)

The obtained result is not an artifact of the NLS approximation, but is confirmed in the simulations of the Dysthe model (Fig. 2b) as well as the fully nonlinear equations (Fig. 2c). Some difference between the diagrams Fig. 2a–c may be noticed, but are not significant.

Steeper wave conditions ($k_0 \eta_{\text{rms}} = 0.056$) are concerned in Fig. 3 for the same initial spectrum width $\nu/k_0 \simeq 0.076$ ($\text{BFI} = 2.08$). The waves are simulated by means of the fully nonlinear model. Although the correlation picture in Fig. 3 is less sharp than in Fig. 2, the level of the phase coherence is again very high. It is

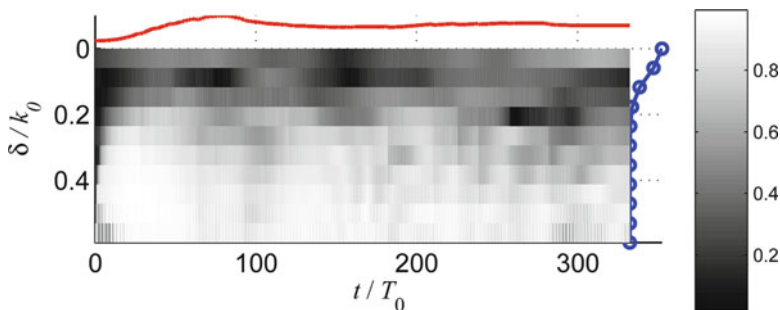


Fig. 3 The diagram of the correlation estimator $R(\delta, t)$ shown by the color intensity. The temporal dependence of the average spectrum width is given over the diagram (*the solid line*) for the reference, and the spectrum shape at $t = 0$ is shown to the right of the diagram (*the line with circles*) for the reference. The fully nonlinear simulation with $\text{BFI}(t = 0) = 2.08$ is reported

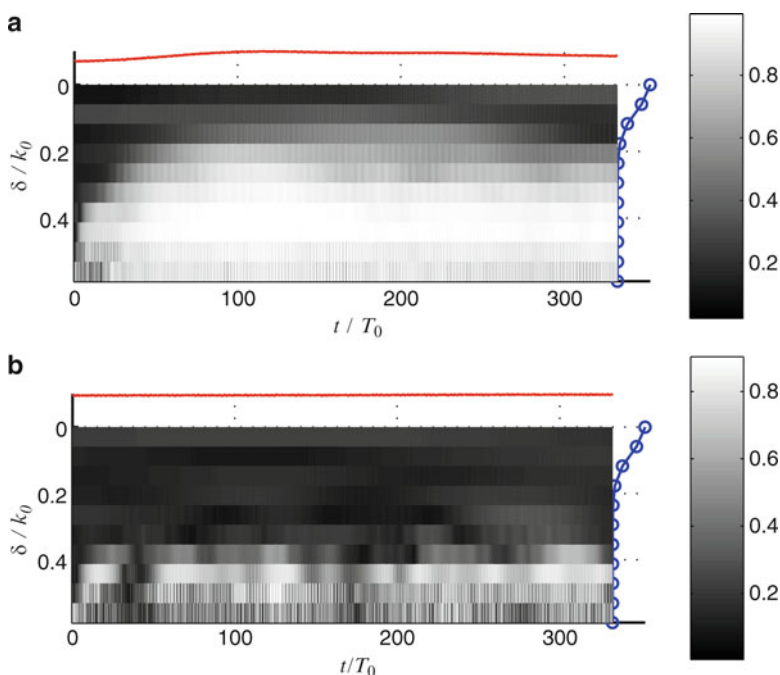


Fig. 4 Diagrams of the correlation estimator $R(\delta, t)$ shown by the color intensity. The temporal dependence of the average spectrum width is given over the diagram (*the solid line*) for the reference, and the spectrum shape at $t = 0$ is shown to the right of the diagram (*the line with circles*) for the reference. The fully nonlinear simulations with $\text{BFI}(t = 0) = 1.04$ (**a**), and $\text{BFI}(t = 0) = 0.52$, (**b**) are reported (different wave intensities)

also significant that in the course of evolution about one half of the realizations resulted in very steep waves; these simulations were not taken into account after the steep event occurrence. Thus, in a certain sense, less coherent wave fields compose the statistical data at longer times.

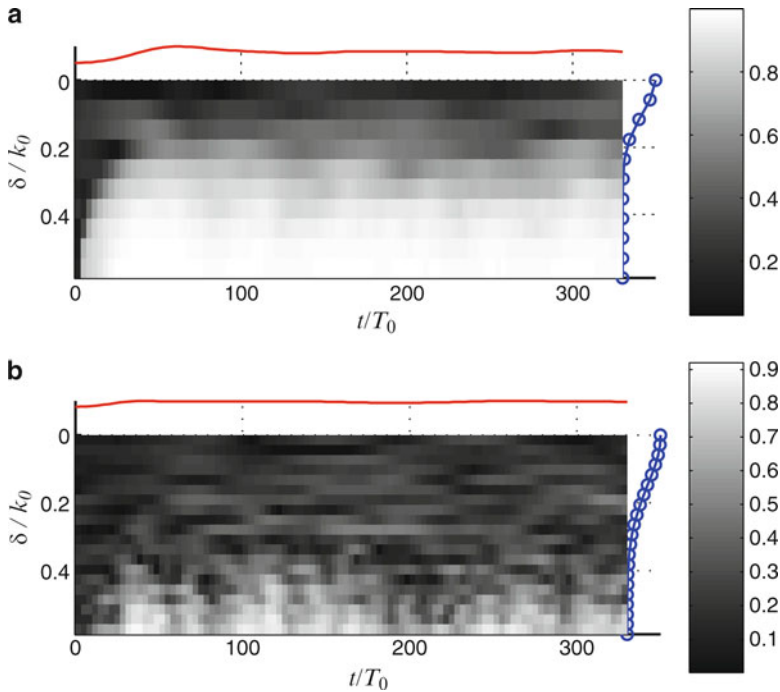


Fig. 5 Diagrams of the correlation estimator $R(\delta, t)$ shown by the colour intensity. The temporal dependence of the average spectrum width is given over the diagram (*the solid line*) for the reference, and the spectrum shape at $t = 0$ is shown to the right of the diagram (*the line with circles*) for the reference. The NLS equation simulations with $\text{BFI}(t = 0) = 1.19$ (a), and $\text{BFI}(t = 0) = 0.79$ (b) are reported (different spectrum widths)

Different wave amplitudes with the same initial spectrum width are considered in Fig. 4 within the fully nonlinear framework with the initial values of the BFI equal to 1.04 and 0.52 (Fig. 4a,b, respectively). The level of the correlation and the interval of wave numbers, where the coherence is revealed are noticeably smaller for the case displayed in Fig. 4b in comparison with Fig. 4a.

In Fig. 5 wave fields with the same root-mean-square surface displacement but different initial spectrum widths are considered. The typical wave amplitude is modest, and the NLS equation is solved to obtain the results. Similar to Fig. 4, the initial BFIs were chosen equal to 1.19 and 0.79 in Fig. 5a,b correspondingly. Again, the coherence is much less evident, if $\text{BFI} < 1$ (Fig. 5b), than if $\text{BFI} > 1$ (Fig. 5a).

A single example based on a laboratory experiment data is given in Fig. 6. The intense irregular wave groups were generated by the wavemarker at one wall of the tank, and then they propagate, see [29]. The initial spectrum had a rectangular shape, and the initial BFI was greater than 1, thus occurrence of large waves associated with rogue events was observed for the distance about 70 m. It is very well seen in the figure that there is a bright spot between $50 \text{ m} < x < 100 \text{ m}$, what manifests a

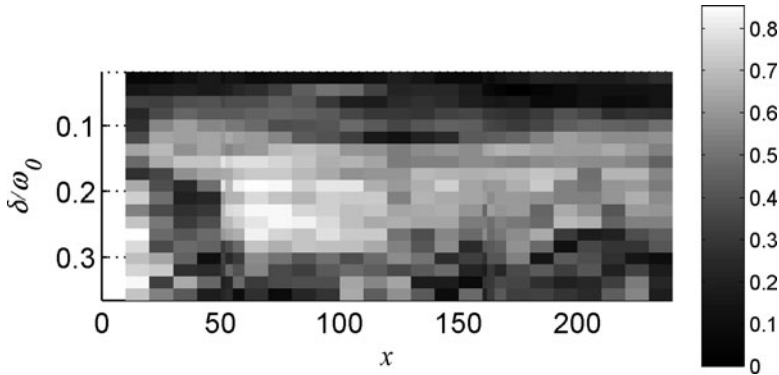


Fig. 6 The correlation estimator R shown by the colour intensity for laboratory wave modeling. The initial condition ($x = 0$) is characterized by $\text{BFI}(t = 0) = 1.71$ ($\eta_{\text{rms}} = 0.03$ m, the wave period is about 1.5 s, and the typical wave length is about 3.5 m; the frequency spectrum width is $\nu_{\omega}/\omega_0 = 0.045$). Coordinate x shows the distance from the wave maker in meters

strong coherence between spectral harmonics. It goes without saying that laboratory data is more difficult for analysis due to the instrumental imperfectness and natural presence of noise in the data. However, the estimator (3) turns out able to capture the coherence.

4 Conclusion

The stochastic approach to the study of intense sea waves, and in particular the freak or rogue waves, has become quite popular during the recent years. The irregular waves are usually defined as a linear superposition of Stokes waves with random phases following the conventional concept of a Gaussian or near-Gaussian sea. Although the random phase assumption is acknowledged to be violated due to nonlinearity, the obtaining of the phase correlation functions for realistic waves from general principles is a hard task. The four-wave interactions are the most efficient for the deep-water case, but the resonance wave quartets interact among each other [18, 33]; moreover, quasi-resonance wave interactions obviously have to be taken into account [4, 33] to obtain a reliable result.

Nonresonant interactions lead to the appearance of phase-locked modes in the Fourier spectrum. The free wave components (they may support exact resonant or near resonant wave modes) are considered governing the physics of nonlinear waves. The bound wave components (otherwise the nonlinear Stokes corrections) at multiple frequencies/wavenumbers are naturally observed in experiments. They are an indicator of occurrence of steep waves. The bound waves are coherent waves, but in the spectrum they are far from the spectral peak, and may be accounted by modern models.

The Benjamin-Feir (modulational) instability has been shown to be a nonlinear effect which increases the probability of freak waves in the deep-water narrow-band case. The onset of the Benjamin-Feir nonlinear instability is well controlled by the Benjamin-Feir index, BFI. Initially unstable wave fields tend to a marginally stable state. The self-modulation effect becomes apparent through the occurrence of large-amplitude coherent wave groups.

In this note, we employ the empirically written estimator for the wave coherence, complied with exact model solutions of the nonlinear Schrödinger equation as well as with the general comprehension of the resonance wave quartet nature, see details in [31]. We focus the attention on the coherence in the Fourier space, which corresponds to harmonics, phase-locked due to the nonlinearity. These modes indicate the presence of the mentioned above large-amplitude coherent wave groups (responsible for freak events) rather than individual steep waves.

The estimator, which is actually a kind of an autocorrelation function, turns out to be efficient to reveal wave coherence in all cases, when the self-modulation effects are significant. The deep-water frequency spectrum is twice narrower than the wavenumber spectrum, and, thus, probably even a stronger wave coherence might be observed in the frequency spectrum. Indeed, a preliminary analysis corroborated the ability of the suggested autocorrelation estimator to reveal the coherence in laboratory-measured time series of the surface displacement.

We conclude that in the case of modulationally unstable wave fields the wave correlation is quite significant and can be revealed in rather close vicinity to the most energetic spectral area. This part of spectrum cannot be considered as a superposition of independent waves with random phases. The phase coherence should be taken into account when describing the nonlinear waves accurately.

Acknowledgements The author thanks A. Sergeeva for providing with Fig. 6 based on laboratory results, and with the laboratory experiment data displayed in Fig. 1.

The research is supported by grants MK-6734.2010.5 and RFBR 11-02-00483, 11-05-00216, and has received funding from the European Community's Seventh Framework Programme FP7-PEOPLE-2009-IIF under grant agreement No 254389.

References

1. Akhmediev, N., Ankiewicz, A., Soto-Crespo, J.M., Dudley, J.M.: Rogue wave early warning through spectral measurements? *Phys. Lett. A* **375**, 541–544 (2011).
2. Akhmediev, N., Ankiewicz, A., Taki, M.: Waves that appear from nowhere and disappear without a trace. *Phys. Lett. A* **373**, 675–678 (2009).
3. Alber, I.E.: The effects of randomness on the stability of two-dimensional wavetrains. *Proc. Roy. Soc. Lond. A* **363**, 525–546 (1978).
4. Annenkov, S.Yu., Shrira, V.I.: Role of non-resonant interactions in the evolution of nonlinear random water wave fields. *J. Fluid Mech.* **561**, 181–207 (2006).
5. Annenkov, S.Y., Shrira, V.I.: Evolution of kurtosis for wind waves. *Geophys. Res. Lett.* **36**, L13603 (2009).
6. Annenkov, S.Y., Shrira, V.I.: “Fast” nonlinear evolution in wave turbulence. *Phys. Rev. Lett.* **102**, 024502 (2009).

7. Chalikov, D.: Freak waves: Their occurrence and probability. *Phys. Fluids* **21**, 076602 (2009).
8. Clamond, D., Francius, M., Grue, J., Kharif, C.: Long time interaction of envelope solitons and freak wave formations. *Eur. J. Mech. B / Fluids* **25**, 536–553 (2006).
9. Dyachenko, A.I., Zakharov, V.E.: Modulation instability of stokes wave \rightarrow freak wave. *JETP Lett.* **81**, 255–259 (2005).
10. Dyachenko, A.I., Zakharov, V.E.: On the formation of freak waves on the surface of deep water. *JETP Lett.* **88**, 307–310 (2008).
11. Dysthe, K.B.: Note on a modification to the nonlinear Schrödinger equation for application to deep water waves. *Proc. Roy. Soc. London A* **369**, 105–114 (1979).
12. Dysthe, K., Krogstad, H.E., Muller, P.: Oceanic rogue waves. *Annu. Rev. Fluid Mech.* **40**, 287–310 (2008).
13. Dysthe, K.B., Trulsen, K.: Note on breather type solutions of the NLS as a model for freak-waves. *Physica Scripta T* **82**, 48–52 (1999).
14. Dysthe, K.B., Trulsen, K., Krogstad, H.E., Socquet-Juglard, H.: Evolution of a narrow-band spectrum of random surface gravity waves. *J. Fluid. Mech.* **478**, 1–10 (2003).
15. Henderson, K.L., Peregrine, D.H., Dold, J.W.: Unsteady water wave modulations: fully nonlinear solutions and comparison with the nonlinear Schrodinger equation. *Wave Motion* **29**, 341–361 (1999).
16. Janssen, P.A.E.M.: Nonlinear four-wave interactions and freak waves. *J. Phys. Oceanogr.* **33**, 863–884 (2003).
17. Johnson, R.S.: A modern introduction to the mathematical theory of water waves. Cambridge Univ. Press (1997).
18. Kartashova, E., Raab, C., Feurer, Ch., Mayrhofer, G., Schreiner, W.: Symbolic Computation for Nonlinear Wave Resonances. In: Pelinovsky, E., Kharif, C. (eds.) *Extreme Waves*, pp. 95–126. Springer (2008).
19. Kharif, C., Pelinovsky, E.: Physical mechanisms of the rogue wave phenomenon. *Eur. J. Mech. B / Fluids* **22**, 603–634 (2003).
20. Kharif, C., Pelinovsky, E., Slunyaev, A.: *Rogue Waves in the Ocean*. Springer-Verlag, Berlin Heidelberg (2009).
21. Mori, N., Onorato, M., Janssen, P.A.E.M., Osborne, A.R., Serio, M.: On the extreme statistics of long-crested deep water waves: Theory and experiments. *J. Geophys. Res.* **112**, C09011 (2007).
22. Onorato, M., Osborne, A.R., Serio, M., Bertone, S.: Freak waves in random oceanic sea states. *Phys. Rev. Lett.* **86**, 5831–5834 (2001).
23. Onorato, M., Osborne, A.R., Serio, M.: Extreme wave events in directional, random oceanic sea states. *Phys. Fluids* **14**, L25–L28 (2002).
24. Onorato, M., Osborne, A.R., Serio, M., Cavaleri, L.: Modulational instability and non-Gaussian statistics in experimental random water-wave trains. *Phys. Fluids* **17**, 078101 (2005).
25. Onorato, M., Waseda, T., Toffoli, A., Cavaleri, L., Gramstad, O., Janssen, P.A., Kinoshita, T., Monbaliu, J., Mori, N., Osborne, A.R., Serio, M., Stansberg, C.T., Tamura, H., Trulsen, K.M.: Statistical properties of directional ocean waves: the role of the modulational instability in the formation of extreme events. *Phys. Rev. Lett.* **102**, 114502 (2009).
26. Osborne, A.R., Onorato, M., Serio, M.: The nonlinear dynamics of rogue waves and holes in deep water gravity wave trains. *Phys. Lett. A* **275**, 386–393 (2000).
27. Shemer, L., Sergeeva, A.: An experimental study of spatial evolution of statistical parameters in a unidirectional narrow-banded random wavefield. *J. Geophys. Res.* **114**, C01015 (2009).
28. Shemer, L., Sergeeva, A., Slunyaev, A.: Applicability of envelope model equations for simulation of narrow-spectrum unidirectional random field evolution: experimental validation. *Phys. Fluids* **22**, 016601 (2010).
29. Shemer, L., Sergeeva, A., Liberzon, D.: Effect of the initial spectral shape on spatial evolution of the statistics of unidirectional nonlinear random waves. *J. Geophys. Res.* **115**, C12039 (2010).
30. Slunyaev, A.V.: Numerical simulation of “limiting” envelope solitons of gravity waves on deep water. *JETP* **109**, 676–686 (2009).

31. Slunyaev, A.V.: Freak wave events and the wave phase coherence. *Europ. Phys. J. Special Topics* **185**, 67–80 (2010).
32. Socquet-Juglard, H., Dysthe, K.B., Trulsen, K., Krogstad, H.E., Liu, J.: Probability distributions of surface gravity waves during spectral changes. *J. Fluid Mech.* **542**, 195–216 (2005).
33. Stiassnie, M., Shemer, L.: On the interaction of four water-waves. *Wave Motion* **41**, 307–328 (2005).
34. Trulsen, K., Kliakhandler, I., Dysthe, K.B., Velarde, M.G.: On weakly nonlinear modulation of waves on deep water. *Phys. Fluids* **12**, 2432–2437 (2000).
35. Zakharov, V.E., Dyachenko, A.I., Vasilyev, O.A.: New method for numerical simulation of a nonstationary potential flow of incompressible fluid with a free surface. *Eur. J. Mech. B/Fluids* **21**, 283–291 (2002).
36. Zakharov, V.E., Ostrovsky, L.A.: Modulation instability: the beginning. *Physica D* **238**, 540–548 (2009).

Quantum Mechanical Treatment of the Lamb Shift Without Taken into Account the Electric Charge

Voicu Dolocan, Andrei Dolocan, and Voicu Octavian Dolocan

1 Introduction

The Lamb shift [7] is a small difference in energy between two energy levels $^2S_{1/2}$ and $^2P_{1/2}$ of the hydrogen atom by an amount now known to be $E/h = 1,057.864\text{MHz}$. This result is in contradiction with the Dirac and Schrödinger theory which shown that the states with the same n and j quantum numbers but different l quantum numbers ought to be degenerate. The effect is explained by the theory of quantum electrodynamics [1,5,9], in which the electromagnetic interaction itself is quantized. It is assumed that the ground state of the electromagnetic field is not zero, but rather the field undergoes “vacuum fluctuations” that interact with the electron. The contributions to this effect come from the vacuum polarization, electron mass renormalization, and anomalous magnetic moment. Often the “vacuum” is a “refuge” for speculations in science.

By using a Hamiltonian of interaction between fermions based on the coupling through flux lines [3], we have found an equivalent expression for the Coulomb energy of interaction [4], on the form $\alpha\hbar c/R$, where α is the fine structure constant.

In the interaction picture, the effective Hamiltonian is given by [3,4]

$$\begin{aligned}
 H_I^{\text{eff}} &= H_{I1}^{\text{eff}} + H_{I2}^{\text{eff}} \\
 H_{I1}^{\text{eff}} &= \hbar \sum_{\mathbf{q}, \mathbf{q}_0, \mathbf{k}} |\mathbf{g}_{\mathbf{q}_0}|^2 \frac{\omega_{\mathbf{q}}}{(\varepsilon_{\mathbf{k}} - \varepsilon_{\mathbf{k}-\mathbf{q}})^2 - \omega_{\mathbf{q}}^2} \times \left(a_{\mathbf{q}} a_{\mathbf{q}_0}^+ a_{\mathbf{q}'_0}^+ a_{\mathbf{q}'}^+ + a_{\mathbf{q}}^+ a_{\mathbf{q}_0} a_{\mathbf{q}'_0}^+ a_{\mathbf{q}'} \right) \\
 &\quad c_{\mathbf{k}-\mathbf{q}, \sigma}^+ c_{\mathbf{k}', \sigma'}^+ c_{\mathbf{k}, \sigma} c_{\mathbf{k}'-\mathbf{q}, \sigma'} \\
 H_{I2}^{\text{eff}} &= 2\hbar \sum_{\mathbf{q}, \mathbf{k}} 2 |\mathbf{g}_{\mathbf{q}}|^2 \frac{1}{(\varepsilon_{\mathbf{k}} - \varepsilon_{\mathbf{k}-\mathbf{q}}) - \omega_{\mathbf{q}}} a_{\mathbf{q}}^+ a_{\mathbf{q}} c_{\mathbf{k}-\mathbf{q}, \sigma}^+ c_{\mathbf{k}-\mathbf{q}, \sigma}
 \end{aligned} \tag{1}$$

V. Dolocan (✉)

Faculty of Physics, University of Bucharest, Bucharest, Romania

e-mail: dolocan_voicu@yahoo.com

The expectation value of the energy of H_{I1}^{eff} (1) is

$$E_{\text{int}} = \hbar \sum_{\mathbf{q}, \mathbf{q}_0, \mathbf{k}} 2 |g_{\mathbf{q}_0}|^2 \frac{\omega_{\mathbf{q}}}{(\varepsilon_{\mathbf{k}} - \varepsilon_{\mathbf{k}-\mathbf{q}})^2 - \omega_{\mathbf{q}}^2} (n_{\mathbf{q}} + 1) (n_{\mathbf{q}_0} + 1) \times n_{\mathbf{k}-\mathbf{q}, \sigma} n_{\mathbf{k}, \sigma'} \quad (2)$$

and the expectation value of the energy of the Hamiltonian H_{I2} (1) is

$$E_{I2} = 4\hbar \sum_{\mathbf{q}, \mathbf{q}_0, \nu'} |g_{\mathbf{q}_0}|^2 \frac{1}{(\varepsilon_{\nu'} - \varepsilon_{\nu}) - \omega_{\mathbf{q}}} (n_{\mathbf{q}} + 1) n_{\mathbf{k}-\mathbf{q}} \quad (3)$$

where

$$g_{\mathbf{q}_0} = \frac{\hbar D}{8N^2 m R \left(\rho_o + \frac{DR}{c^2} \right)} \frac{\mathbf{q} \mathbf{q}'}{\omega_{\mathbf{q}} \omega_{\mathbf{q}_0}} \sum_n e^{i\mathbf{q}_0 \cdot \mathbf{z}_n} \quad (4)$$

D is a coupling constant, \mathbf{q}, \mathbf{q}' are the wave vectors associated with the bosons of the connecting field, \mathbf{q}_0 is the wave vector associated with the oscillations of the electron, and \mathbf{k}, \mathbf{k}' are the wave vectors of the electrons. $\omega_{\mathbf{q}}, \omega_{\mathbf{q}_0}$ are the classical oscillation frequencies, $a_{\mathbf{q}}^+$ and $a_{\mathbf{q}}$ are the boson creation and annihilation operators, respectively, $c_{\mathbf{k}\sigma}^+$ and $c_{\mathbf{k}\sigma}$ are the creation and annihilation operators for electrons, \mathbf{k} is the wave vector of an electron, $n_{\mathbf{q}}$ is the occupation number for bosons and $n_{\mathbf{k}}$ is the occupation number for fermions. We assume $n_{\mathbf{q}}, n_{\mathbf{q}_0} = 0, n_{\mathbf{k}}, n_{\mathbf{k}-\mathbf{q}} = 1$.

If instead of the Fröhlich fraction

$$\frac{\omega_{\mathbf{q}}}{(\varepsilon_{\mathbf{k}} - \varepsilon_{\mathbf{k}-\mathbf{q}})^2 - \omega_{\mathbf{q}}^2} \quad (5a)$$

we use the fraction

$$\frac{-1}{(\varepsilon_{\mathbf{k}} - \varepsilon_{\mathbf{k}-\mathbf{q}}) - \omega_{\mathbf{q}}} \quad (5b)$$

then for $\rho_o = 0$ (mass less interacting field), (2) becomes

$$E_I = \frac{\hbar^3 c^4}{32m^2 R^4} \sum_{\mathbf{q}, \mathbf{q}_0, \mathbf{k}} \frac{(\mathbf{q} \cdot \mathbf{q}_0)^2}{\omega_{\mathbf{q}}^2 \omega_{\mathbf{q}_0}^2} \left| \sum_n e^{i\mathbf{q}_0 \cdot \mathbf{R}_n} \right|^2 \frac{-1}{\varepsilon_{\mathbf{k}} - \varepsilon_{\mathbf{k}-\mathbf{q}} - \omega_{\mathbf{q}}} \quad (2a)$$

Now we apply this equation to a system of two electrons at \mathbf{R}_1 and \mathbf{R}_2 acting on the vacuum of the less mass boson field. In this case $\sum_n |e^{i\mathbf{q}_0 \cdot \mathbf{z}_n}|^2 = 2(1 + \cos(\mathbf{q}_0 \cdot \mathbf{R}))$, $\varepsilon_{\mathbf{k}} = \hbar k^2 / 2m$, and

$$\omega_{\mathbf{q}} = cq, \omega_{\mathbf{q}_0} = \hbar q_o^2 / 2m,$$

$$\mathbf{R} = \mathbf{R}_1 - \mathbf{R}_2.$$

Further, if we assume $(\epsilon_{\mathbf{k}} - \epsilon_{\mathbf{k}-\mathbf{q}}) \ll \omega_{\mathbf{q}}$, we write

$$\sum_{\mathbf{q}} \frac{(\mathbf{q} \cdot \mathbf{q}_0)^2}{\omega_{\mathbf{q}}^3 \omega_{\mathbf{q}_0}^2} = \left(\frac{2m}{\hbar} \right)^2 \frac{1}{q_0^2 c^3} \frac{\Omega}{(2\pi)^2} \int_0^\pi \cos^2 \alpha \sin \alpha d\alpha \int_0^{q_0} q dq = \left(\frac{2m}{\hbar} \right)^2 \frac{R^3}{9\pi c^3}$$

and

$$\begin{aligned} 2 \sum_{\mathbf{q}_0} [1 + \cos(\mathbf{q}_0 \cdot \mathbf{R})] &= 2 \sum_{\mathbf{q}_0} 1 + 2 \sum_{\mathbf{q}_0} \cos(\mathbf{q}_0 \cdot \mathbf{R}) \\ &= 2 + 2 \frac{\Omega}{(2\pi)^2} \int_0^{0.76\pi/R} dq_0 q_0^2 \int_0^\pi d\theta \cos(q_0 R \cos \theta) \sin \theta = 3.028 \end{aligned}$$

We have considered $N = 1$, $\sum_{\mathbf{k}} 1 = 1$ and $\Omega = 4\pi R^3/3$. The upper limit of the integrals over q_0 appears from the requirement

$$\frac{4\pi R^3/3}{(2\pi)^3} \times 4\pi \frac{q_m^3}{3} = 1$$

The interaction energy (2a) becomes

$$E_I \approx 2 \times 7.24 \times 10^{-3} \frac{\hbar c}{R} \approx 2 \times \frac{1}{137} \frac{\hbar c}{R} = 2 \frac{e^2}{4\pi\epsilon_0 R}$$

The factor 2 appears because we have considered the two nearest neighbours of an electron. The interaction energy between two electrons is

$$E_I = \alpha \frac{\hbar c}{R} \quad (6)$$

where $\alpha = 1/137$ is the fine structure constant. This is an equivalent expression for the Coulomb's law. Expression (6) is obtained from (2a), containing the fraction (5b), which is valid for the interaction between the like charges acting in a mass less boson field. In this case, the flux lines of the two particles do not interfere, the two particles absorb bosons from the ambient space, and move apart from one another. In the case where the two particles have opposite charges, then in a mass less boson field, the interaction energy is given by (2a) where the fraction (5b) is substituted by fraction (5a) and the interaction energy (6) becomes negative (in front of the term from the right hand side is a negative sign). The explanation for this is as follows. Fröhlich obtained the fraction (5a) by dividing fraction (5b) into two parts

$$\frac{1}{2} \frac{1}{\epsilon_{\mathbf{k}} - \epsilon_{\mathbf{k}-\mathbf{q}} - \omega_{\mathbf{q}}} - \frac{1}{2} \frac{1}{\epsilon_{\mathbf{k}'} + \omega_{\mathbf{q}} - \epsilon_{\mathbf{k}'+\mathbf{q}}}$$

one part for absorption and the other for emission of a boson. When the two charges have an opposite sign, there is a continuity of the field lines from one particle to the other, one particle absorbs and the other emits a boson, so that an attraction

is assured. Also, an attraction may be assured between the like charges, when the connecting field is a massive field; likewise, in this case the interaction energy is given by (2).

When a particle has a charge Q , that is Q/e electronic charges, the term from the right hand side of (6) is multiplied by $Q_1 Q_2 / e^2$, because in this case we must define $|\Psi|^2 = Q/e$ (the number of electronic charges per particle).

2 Lamb Shift in the Three-Dimensional Space

From (3) and (4) one obtains

$$E_{I2} = \frac{\hbar^3 D^2}{16m^2 R^2 \left(\rho_o + \frac{DR}{c^2}\right)^2} \frac{4\pi V}{(2\pi)^3} \int_{mc/\hbar}^{\infty} q_o^2 dq_o \frac{4\pi V}{(2\pi)^3} \times \int_0^{q_o} q^2 dq \frac{q^2 q_o^2}{\omega_q^2 \omega_{q_o}^2} \sum_{v'} \frac{1}{\varepsilon_{v'} - \varepsilon_v - \omega_q} \quad (7a)$$

We recognize that there is also a shift for free states. For free electrons (7a) becomes

$$E_{I2f} = \frac{\hbar^3 D^2}{8m^2 R^2 \left(\rho_o + \frac{DR}{c^2}\right)^2} \frac{4\pi V}{(2\pi)^3} \int_{mc/\hbar}^{\infty} q_o^2 dq_o \frac{4\pi V}{(2\pi)^3} \times \int_0^{q_o} q^2 dq \frac{q^2 q_o^2}{\omega_q^2 \omega_{q_o}^2} \frac{-1}{\omega_q} \quad (7b)$$

One gets the physical energy shift by subtracting expression (7b) from (7a). The expression renormalized in this way is

$$\begin{aligned} \delta E = E_{I2} - E_{I2f} &= \frac{\hbar^3 D^2 V^2}{64\pi^4 m^2 R^2 \left(\rho_o + \frac{DR}{c^2}\right)^2} \int_{mc/\infty}^{\infty} q_o^2 dq_o \\ &\times \int_0^{q_o} q^2 dq \frac{q^2 q_o^2}{\omega_q^2 \omega_{q_o}^2} \sum_{v'} \left(\frac{1}{\varepsilon_{v'} - \varepsilon_v - \omega_q} + \frac{1}{\omega_q} \right) \\ &= \frac{\hbar^3 D^2 (4\pi R^3/3)^2}{64\pi^4 m^2 R^2 \left(\rho_o + \frac{DR}{c^2}\right)^2} \int_{mc/\hbar}^{\infty} q_o^2 dq_o \int_0^{q_o} q^2 dq \\ &\times \frac{q^2 q_o^2}{\omega_q^3 \omega_{q_o}^2} \sum_{v'} \frac{\varepsilon_{v'} - \varepsilon_v}{\varepsilon_{v'} - \varepsilon_v - \omega_q} \end{aligned} \quad (8)$$

Now we assume $\rho_o = 0$, $\omega_{\mathbf{q}} = cq$, $\omega_{\mathbf{q}_o} = \hbar q_o^2 / 2m_{\mathbf{q}} = cq$, and $\varepsilon_{v'} - \varepsilon_v \ll \omega_{\mathbf{q}}$. Therefore

$$\delta E = -\frac{\hbar R^2}{9\pi^2} \int_{mc/\hbar}^{\infty} dq_o \int_0^{q_o} dq \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) \quad (9)$$

The expectation values of the energy $\hbar\varepsilon_v = E_v$ are determined from the Schrödinger equations

$$\begin{aligned} H_o \Psi_v &= \left(-\hbar^2 \frac{\nabla^2}{2m} - \frac{Z\alpha\hbar c}{r} \right) \Psi_v = E_v \Psi_v \\ H \Psi_{v'} &= \left[-\hbar^2 \frac{\nabla^2}{2m} - Z\alpha\hbar c \left(\frac{1}{r} - \frac{1}{2} (\delta r)^2 \nabla^2 \left(\frac{1}{r} \right) \right) \right] \Psi_{v'} = E_{v'} \Psi_{v'} \end{aligned} \quad (10)$$

Ψ_v is the nonrelativistic wave function in the hydrogen like atom. $(\delta r)^2 = s^2$, where s^2 is the mean square value of s_l [3,4]

$$s^2 = \frac{1}{R^2} \frac{\hbar^2}{4m^2 \omega_{\mathbf{q}_o}^2} (2n_{\mathbf{q}_o} + 1) \quad (11)$$

Even in the lowest state ($n_{\mathbf{q}_o} = 0$) the oscillator has a finite amplitude with a finite probability. In this case, $s^2 = (1/R^2 q_o^4)$. By using that

$$\nabla^2 \left(\frac{1}{r} \right) = -4\pi \delta(\mathbf{r})$$

and by using (6) may be written

$$\begin{aligned} \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) &= -2 \times \frac{1}{2} Z\alpha c \frac{1}{R^2 q_o^4} \int \Psi_{vl}^*(\mathbf{r}) \times \nabla^2 \left(\frac{1}{r} \right) \Psi_{vl}(\mathbf{r}) d\mathbf{r} \\ &= -\frac{4\pi Z\alpha c}{R^2 q_o^4} |\Psi_{vl}(0)|^2 \delta_{l0} = -\frac{Z\alpha c}{2a_o^5 q_o^4} \delta_{l0} \end{aligned} \quad (12)$$

A factor 2 appears because of the two values of the electron spin. For p orbitals, the nonrelativistic wave function vanishes at the origin, so there is no energy shift. But for s orbitals there is some finite value at the origin

$$\Psi_{2s}(0) = \left(\frac{1}{8\pi a_o^3} \right)^{1/2}$$

where we have denoted $R \equiv a_0$, the Bohr radius. By substituting (12) in (9), we have

$$\delta E_{2o} = \frac{Z\alpha \hbar c}{18\pi^2 a_o^3} \int_{mc/\hbar}^{\infty} \frac{dq_o}{q_o^4} \int_0^{q_o} dq = \frac{Z\alpha \hbar^3}{36\pi^2 m^2 c a_o^3} \quad (13)$$

For hydrogen atom, $Z = 1$, this shift is about 6.42×10^{-25} J, which correspond to a frequency of 970 MHz. Further we consider the contribution to the Lamb shift of the interaction terms from the fine-structure Hamiltonian in according to Dirac theory [2]. The first term of interaction is the usual spin-orbit coupling

$$H'_2 = \frac{1}{2m^2 c^2} \frac{1}{r} \frac{dV}{dr} \mathbf{L} \cdot \mathbf{S} = \frac{1}{2m^2 c^2} \frac{Z\alpha \hbar c}{r^3} \times \frac{1}{2} (\mathbf{J}^2 - \mathbf{L}^2 - \mathbf{S}^2) \quad (14)$$

and the second term of interaction is the Darwin term due to the nonlocalized interaction between the electron and the field

$$H'_3 = \frac{\hbar^2}{8m^2 c^2} \nabla^2 V = -\frac{\hbar^2}{8m^2 c^2} \frac{2Z\alpha \hbar c}{r^3} \quad (15)$$

In the case of the spin-orbit coupling

$$\begin{aligned} \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) &= 2 \times \frac{1}{2m^2 c^2} \frac{Z\alpha \hbar^2 c s^2}{2} \int \Psi_{vljm_j}^*(\mathbf{r}) \times \nabla^2 \left(\frac{1}{r^3} \right) \mathbf{L} \mathbf{S} \Psi_{vljm_j} d\mathbf{r} \\ &= \frac{6Z\alpha \hbar^2 s^2}{m^2 c} \times \left[j(j+1) - l(l+1) - \frac{3}{4} \right] \int \Psi_{vl}^*(\mathbf{r}) \frac{1}{r^5} \Psi_{vl}(\mathbf{r}) d\mathbf{r} \\ &= \frac{6Z\alpha \hbar^2 s^2}{m^2 c} \left[j(j+1) - l(l+1) - \frac{3}{4} \right] \left\langle \frac{1}{r^5} \right\rangle \end{aligned} \quad (16)$$

By substituting (16) in (9), one obtains

$$\delta E_{\mathbf{LS}}^{21} = -\frac{Z\hbar^3 \alpha}{3\pi^2 m^2 c} \frac{\hbar^2}{m^2 c^2} \left\langle \frac{1}{r^5} \right\rangle_{21} \times \begin{cases} 1 \text{ for } j = 1 + 1/2 \\ -2 \text{ for } j = 1 - 1/2 \end{cases} \quad (17)$$

By using the radial wave function of the electron in the hydrogen atom

$$R_{21}(r) = \frac{1}{\sqrt{3}} \left(\frac{Z}{2a_o} \right)^{3/2} \left(\frac{Zr}{a_o} \right) \exp(-Zr/a_o)$$

and

$$\left\langle \frac{1}{r^5} \right\rangle_{21} = \int_{\lambda_c}^{\infty} |R_{21}|^2 \frac{dr}{r^3}$$

one obtains

$$\left\langle \frac{1}{r^5} \right\rangle_{21} = \frac{0.15}{a_o^5}$$

and $\delta E_{\text{LS}}^{21} = 1.2 \times 10^{-28} \text{J}$, which correspond to a frequency of $\sim 0.18 \text{MHz}$. In the above expression, $\lambda_C = h/mc$ is the Compton wave length. The spin-orbit coupling contribution is zero for s electrons. Next we consider the Darwin term (15). We write

$$\begin{aligned} \hbar \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) &= -2 \times \frac{\hbar^2}{8m^2c^2} \frac{1}{2} \alpha s^2 \hbar c \times \int \Psi_{vl}^*(\mathbf{r}) \nabla^4 \left(\frac{1}{r} \right) \Psi_{vl}(\mathbf{r}) \\ &= \frac{3\hbar^3 \alpha}{m^2c} \frac{1}{R^2 q_o^4} \left\langle \frac{1}{r^5} \right\rangle \end{aligned} \quad (18)$$

and

$$\delta E_{21}^{\text{Darwin}} = \frac{\hbar^3 \alpha}{3\pi^2 m^2 c} \left\langle \frac{1}{r^5} \right\rangle_{21} \int_{mc/\hbar}^{\infty} \frac{dq_o}{q_o^5} \int_0^{q_o} dq = \frac{\hbar^3 \alpha}{6\pi^2 m^2 c} \frac{0.15}{a_o^5} \frac{\hbar^2}{m^2 c^2} = 2.9 \times 10^{-29} \text{J} \quad (19)$$

The Darwin contribution to the $2s$ level is

$$\delta E_{20}^{\text{Darwin}} = \frac{\hbar^3 \alpha}{3\pi^2 m^2 c} \frac{\hbar^2}{m^2 c^2} \left\langle \frac{1}{r^5} \right\rangle_{20} \quad (20)$$

By using the radial wave function of the electron in the hydrogen atom

$$R_{20} = 2 \left(\frac{1}{2a_o} \right)^{3/2} \left(1 - \frac{r}{2a_o} \right) \exp(-r/2a_o)$$

one obtains

$$\left\langle \frac{1}{r^5} \right\rangle_{20} = \int_{h/mc}^{\infty} |R_{20}|^2 \frac{dr}{r^3} = \frac{101.392}{a_o^5}$$

and $\delta E_{20}^{\text{Darwin}} = 2 \times 10^{-26} \text{J}$. In the last equations, we have considered $Z = 1$. The contribution of the Darwin term is of the order of 60.2MHz . By adding the Darwin contribution to the contribution (13), one obtains $\nu = 1,030.2 \text{MHz}$, which is close to the experimental value of the Lamb shift.

3 The Fine Structure Constant in the Two-Dimensional Space

In the two-dimensional space may be written

$$\left| \sum_l e^{i\mathbf{q}_0 \cdot \mathbf{R}_l} \right|^2 \delta_{R_{12}, R_o} = 2 \left[1 + \frac{1}{2\pi} \int_0^{2\pi} e^{iq_o R_o \cos \theta} d\theta \right] = 2[1 + J_o(q_o R_o)] \quad (21)$$

where $J_o(x)$ is the Bessel function of the first kind, $\mathbf{R}_{12} = \mathbf{R}_1 - \mathbf{R}_2$. Further,

$$\begin{aligned} \sum_{\mathbf{q}} \frac{(\mathbf{q} \cdot \mathbf{q}_o)^2}{\omega_{\mathbf{q}}^2 \omega_{\mathbf{q}_o}^2} &= \left(\frac{2\mu}{\hbar} \right)^2 \sum_q \frac{q^2 q_o^2 \cos^2 \alpha}{q_o^4 q^3 c^3} \left(\frac{2\mu}{\hbar} \right)^2 \times \sum_q \frac{\cos^2 \alpha}{q_o^2 q c^3} \\ &= \frac{S}{(2\pi)^2} \left(\frac{2\mu}{\hbar} \right)^2 \frac{1}{c^3 q_o^2} \int_0^{2\pi} \cos^2 \alpha d\alpha \int_0^{q_o} dq = \frac{\pi R^2}{(2\pi)^2} \left(\frac{2\mu}{\hbar} \right)^2 \frac{\pi}{c^3 q_o} \\ &= \left(\frac{2\mu}{\hbar} \right)^2 \frac{R^2}{4c^3 q_o} \end{aligned} \quad (22a)$$

Next,

$$\begin{aligned} \sum_{\mathbf{q}, \mathbf{q}_o} \frac{(\mathbf{q} \cdot \mathbf{q}_o)^2}{\omega_{\mathbf{q}}^3 \omega_{\mathbf{q}_o}^2} [1 + J_o(q_o R_o)] &= \left(\frac{2\mu}{\hbar} \right)^2 \frac{R^2}{4c^3} \frac{S}{(2\pi)^2} \times 2\pi \int_0^{2/R} \frac{1 + J_o(q_o R_o)}{q_o} q_o dq_o \\ &= \left(\frac{2\mu}{\hbar} \right)^2 \frac{R^2}{4c^3} \frac{\pi R^2}{2\pi} \frac{1}{R} \left[2 + \int_0^2 J_o(q_o R_o) d(q_o R_o) \right] \\ &= \left(\frac{2\mu}{\hbar} \right)^2 \frac{R^3}{8c^3} [2 + 1.426] \end{aligned} \quad (22b)$$

The interaction energy in the two-dimensional space becomes

$$E_I^{2D} = -\frac{\hbar^3 c^4}{16\mu^2 R^4} \frac{4\mu^2}{\hbar^2} \frac{R^3}{8c^3} \times 3.426 = -2 \times 0.053 \frac{\hbar c}{R} = -2 \times \frac{\alpha_{2D} \hbar c}{R} \quad (23)$$

where the fine structure constant in the two-dimensional space is

$$\alpha_{2D} = 0.053 = 7.26\alpha \quad (23a)$$

where α is the fine structure constant in the three-dimensional space. It appears that in the two-dimensional space the Coulomb interaction is approximately 7 times stronger than in the three-dimensional space. Here $1/\mu = 1/M_p + 1/m$, where μ

is the reduced mass of the proton and electron in the hydrogen atom. We consider $\mu \approx m$, the electron mass. The electron radius in the hydrogen atom in the two-dimensional space is

$$a_n^{2D} = \frac{\hbar}{\alpha_{2D}\mu c} \left(n - \frac{1}{2} \right)^2 \quad (24)$$

where n is the principal quantum number. The Bohr radius a_1^{2D} is 29 times larger than the Bohr radius in the three-dimensional space. The increasing of binding energy of the electron in the hydrogen atom is

$$\frac{E_n^{2D}}{E_n^{3D}} = \frac{\alpha_{2D}^2}{\alpha^2} \frac{n^2}{\left(n - \frac{1}{2} \right)^2} \quad (25)$$

For the ground state, the binding energy in the two-dimensional space is 200 times larger than that in the three-dimensional space. It results that if should be a two-dimensional space, in this space the matter should be more condensate than in the three-dimensional space. We specify that the two-dimensional hydrogen atom was studied in the past by many authors [6, 8, 10, 11].

4 Lamb Shift in the Two-Dimensional Space

In the two-dimensional space the expression (8) for the Lamb shift becomes

$$\delta E = \frac{\hbar^3 D^2}{16m^2 R^2 \left(\rho_o + \frac{DR}{c^2} \right)^2} \frac{S^2}{(2\pi)^2} \int_{mc/\hbar}^{\infty} q_o dq_o \times \int_0^{q_o} q dq \frac{q^2 q_o^2}{\omega_{\mathbf{q}}^2 \omega_{\mathbf{q}_o}^2} \sum_{v'} \frac{\varepsilon_{v'} - \varepsilon_v}{(\varepsilon_{v'} - \varepsilon_v) - \omega_{\mathbf{q}}} \quad (26)$$

No, we assume $\rho_o = 0$, $\omega_{\mathbf{q}} = cq$, $\omega_{\mathbf{q}_o} = \hbar q_o^2/2m$, $\varepsilon_{v'} - \varepsilon_v \ll \omega_{\mathbf{q}}$, so that

$$\delta E = -\frac{\hbar}{16} \int_{mc/\hbar}^{\infty} \frac{dq_o}{q_o} \int_{mc/\hbar}^{q_o} \frac{dq}{q} \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) \quad (26a)$$

We write

$$\begin{aligned} \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) &= 2 \times \frac{1}{2} Z \alpha_{2D} c \frac{1}{R^2 q_o^4} \times \int \Psi_{vl}^*(\mathbf{r}) \nabla^2 \left(\frac{1}{r} \right) \Psi_{vl}(\mathbf{r}) d\mathbf{r} \\ &= \frac{2Z\alpha_{2D}c}{R^2 q_o^4} \left\langle \frac{1}{r^3} \right\rangle_{vl} \end{aligned}$$

For the 3D case, since $\nabla^2 V(\mathbf{r}) = -4\pi\delta(\mathbf{r})$, δE is nonzero only for the s electron. However, this is no longer true for the 2D hydrogen atom, this term is nonzero for all the electrons. By using that

$$\begin{aligned}\nabla^2 \left(\frac{1}{r} \right) &= \frac{2}{r^3}: \quad \Psi_{2o}(\mathbf{r}) = \frac{\beta_2}{\sqrt{3}} e^{-\beta_2 r/2} (1 - \beta_2 r) \times \frac{1}{\sqrt{\pi}} e^{i|m|\phi} = R_{2o}(\mathbf{r}) \frac{e^{i|m|\phi}}{\sqrt{2\pi}} \\ \Psi_{21}(\mathbf{r}) &= \frac{\beta_2^2 r}{\sqrt{6}} e^{-\beta_2 r/2} \frac{e^{i|m|\phi}}{\sqrt{2\pi}} = R_{21}(\mathbf{r}) \frac{e^{i|m|\phi}}{\sqrt{2\pi}}; \\ \beta_2 &= \frac{4}{3} \frac{mc}{\hbar} \alpha_{2D}\end{aligned}\tag{27}$$

For 2s level, we have

$$\left\langle \frac{1}{r^3} \right\rangle_{20} = \frac{\beta_2^2}{3} \times 0.091 \beta_2 = 0.03 \beta_2^3$$

and therefore

$$\begin{aligned}\delta E_{2s}^{2D} &= \frac{\hbar}{16} \frac{2Z\alpha_{2D}c}{(a_2^{2D})^2} \times 0.03 \beta_2^2 \int_{mc/\hbar}^{\infty} \frac{dq}{q} \int_{mc/\hbar}^q \frac{dq_o}{q_o^5} = \frac{3.75 \times 10^{-3}}{4} \frac{\alpha_{2D}\hbar c}{(a_2^{2D})^2} \beta_2^3 \\ &\times \left\{ \int_{mc/\hbar}^{\infty} \frac{dq}{q^5} - \left(\frac{\hbar}{mc} \right)^4 \int_{mc/\hbar}^{2/R} \frac{dq}{q} \right\} \\ &= 4.38 \times 10^{-4} \alpha_{2D}^6 mc^2 \left\{ \frac{1}{4} - \ln \left(\frac{8}{9} \alpha_{2D} \right) \right\} = 1.98 \times 10^{-25} \text{J}\end{aligned}\tag{28}$$

which correspond to a frequency of 300 MHz. For 2p level

$$\left\langle \frac{1}{r^5} \right\rangle_{21} = \frac{0.193}{6} \beta_2^5$$

and $\delta E_{2p}^{2D} = 5.28 \times 10^{-5} \text{J}$ which corresponds to a frequency of $\sim 800 \text{MHz}$.

For the spin-orbit coupling we have

$$\sum_{v'} (\varepsilon_{v'} - \varepsilon_v) = 2 \frac{\hbar}{m^2 c^2} \frac{\alpha_{2D}\hbar c}{(a_2^{2D})^2} \times [j(j+1) - l(l+1) - 3/4] \left\langle \frac{1}{r^5} \right\rangle_{21}\tag{29}$$

and

$$\delta E_{\text{LS}}^{2p} = 0.0065 \alpha_{2D}^8 mc^2 = 3.31 \times 10^{-26} \text{J}\tag{30}$$

which corresponds to a frequency of 50 MHz.

For the Darwin term we have

$$\hbar \sum_{v'} (\varepsilon_{v'} - \varepsilon_v) = \frac{3\hbar^3 \alpha_{2D}}{m^2 c} \frac{1}{R^2 q_o^4} \left\langle \frac{1}{r^5} \right\rangle \quad (31)$$

and

$$\delta E_{2D}^{\text{Darwin}} = \frac{\hbar^5 \alpha_{2D}^3}{27 m^4 c^3} \left\langle \frac{1}{r^5} \right\rangle \quad (32)$$

so that

$$\delta E_{2D,2s}^{\text{Darwin}} = 8.27 \times 10^{-4} \alpha_{2D}^8 m c^2 = 4.22 \times 10^{-27} \text{J}$$

which corresponds to a frequency of 6.3 MHz, and

$$\delta E_{2D,2p}^{\text{Darwin}} = 5 \times 10^{-3} \alpha_{2D}^8 m c^2 = 2.55 \times 10^{-26} \text{J}$$

which corresponds to a frequency of 38 MHz. The total shift in the two-dimensional space is of the order of 500 MHz. This value of the shift is some smaller than that in the three-dimensional space.

5 Conclusions

We have presented a theory of the Lamb shift without taking into account the electron charge. This appears as a natural result of the equivalent expression for the Coulomb's interaction energy, $\alpha \hbar c / R$, just derived from our Hamiltonian of interaction. This gives rise to the questions: may be the electron taken off the charge, or the electron is an indestructible charge? What is this the charge? We specify that in 1947 Hans Bethe was the first to explain the Lamb shift in the hydrogen spectrum and thus laid the foundation for the modern development of quantum electrodynamics. Neither the mass nor the charge of the electron or any other charged particle can actually be calculated in QED—they have to be assumed. In our theory, it is not necessary to assume a priori the charge of the electron. Also, we have found that in the two-dimensional space, the matter is more condensate than in the three-dimensional space. In another paper, we extend these results to the interaction between nucleons via mass less bosons and massive particles [12].

References

1. Bethe H. A. (1947), The electromagnetic shift of energy levels, *Phys. Rev.* 72, 339–341.
2. Bransden B. H. and C. J. Joachain (1983), *The Physics of Atoms and Molecules*, Longman Group Ltd.

3. Dolocan A., V. O. Dolocan and V. Dolocan (2005), A new Hamiltonian of interaction for fermions, *Mod. Phys. Lett. B*19, 669–681.
4. Dolocan V., A. Dolocan and V. O. Dolocan (2010), Quantum Mechanical Treatment of the Electron-Boson Interaction Viewed as a Coupling through Flux Lines. Polarons, *Int. J. Mod. Phys. B*24, 479–495; On the Chemical Bindings in Solids, *Rom. J. Phys.* 55, 153–172.
5. Greiner W. and J. Reinhardt (1994), *Quantum Electrodynamics*, Springer-Verlag, Berlin.
6. Guo S. H., X. L. Yang, F.T. Chen, K. W. Wong and W. W. Y. Ching (1991), Analytic Solution of a Two-Dimensional Hydrogen Atom. I. Nonrelativistic Theory, *Phys. Rev. A*43, 1197–1205.
7. Lamb E., Jr. and R. C. Retherford (1947), Fine Structure of the Hydrogen Atom by a Microwave Method, *Phys. Rev.* 72, 241–243.
8. Taut M. (1995), Two-dimensional Hydrogen in a Magnetic Field, *J. Phys. A*28, 2081–2085.
9. Welton T. A. (1948), Some Observable Effects of the Quantum-Mechanical Fluctuations of the Electromagnetic Field, *Phys. Rev.* 74, 1157–1167.
10. Yang X. L., S. H. Guo, F. T. Chan, K. W. Wong an, W. Y. Ching (1991), Analytic Solution of a two Dimensional Hydrogen Atom. II. Relativistic Theory, *Phys. Rev. A*43, 1186–1196.
11. Zaslow B. and M. E. Zandler (1967), Two-Dimensional analog to the Hydrogen Atom, *Am. J. Phys.* 35, 1118–1119.
12. Dolocan V., V. O. Dolocan and A. Dolocan (to be published), On the interaction between fermions via mass less bosons and massive particles.

Determining the Climate Zones of Turkey by Center-Based Clustering Methods

Fidan M. Fahmi, Elçin Kartal, Cem İyigün, Ceylan Yozgatligil,
Vilda Purutcuoğlu, İnci Batmaz, Murat Türkeş, and Gülser Köksal

1 Introduction

The increase in atmospheric concentrations of so-called greenhouse gasses such as carbon dioxide (CO₂) and methane (CH₄) has resulted in greenhouse effect and global warming. It is not surprising to see the effect of these changes also in Turkey. Recent analyses of Turkey's climate data reveal that minimum and maximum temperatures are increasing [5]. On the contrary, observed temporal distribution patterns of winter rainfalls are significantly changing while moderate decreases in the total amount of rainfall have also been recorded. As a result of global warming, alterations in the weather patterns and the existence of extreme events can be considered as important indicators of the change in climate zones. To reduce the future vulnerability, it is important to examine the changes and new climate zones (if any) so that appropriate strategies can be developed and precautions can be taken accordingly.

In determining the climate zones, different methods and variables have been used in the literature. The most popular one is the classification method called Koeppen and Thorntwaite [6]. This method is a quantitative one that directly determines the climate types but the rules used in classification are subjectively extracted. Therefore, relatively objective clustering approaches can be applied instead for the same purpose (e.g. [7]). To exemplify, in the study of Ünal et al.'s, Ward's method is proposed to be used as a hierarchical clustering technique.

On the other hand, in literature, temperature and precipitation are treated as the main research variables in this kind of study because of their influence on the distribution of flora and human activities. In our study, temperature measures

F.M. Fahmi (✉)

Department of Statistics, Middle East Technical University, Ankara, Turkey
e-mail: fidantelaferli@yahoo.com

Table 1 Variables studied and corresponding number of clusters and sample size

Temperature variables	Number of clusters	Sample size
Average	4	61
Minimum	4	65
Maximum	5	65
Average minimum	4	65
Average maximum	4	65
Minimum average	4	65
Maximum Average	5	65

obtained from the Turkish State Meteorological Service stations in the time period 1950–2006 are examined by using two different center-based clustering techniques, which are k-means and fuzzy k-means.

2 Data

The data used in identifying the climate zones of Turkey are temperature measures including minimum and maximum temperature, average temperature, and average of maximum and minimum temperature, minimum and maximum of average temperature. In this data set, the measurements are monthly records belonging to the period 1950–2006. Although 270 Turkish State Meteorological Service stations provide climate data, only 65 stations are used within the 1950–2006 period. The main reason of using such a small number of stations is that the eliminated stations had missing values for more than two successive years. Mean imputation is applied for the missing values recorded for months within a year. For each temperature variable, the number of stations (i.e. sample size) used are listed in Table 1.

3 The Methodology

Clustering is the partitioning of a data set into subsets in such a way that the data in each subset share some common trait—often proximity—according to a distance measure. We assume that our data $D = \{x_{ij}\}$, where $i = 1, 2, \dots, n, j = 1, 2, \dots, p$, consist of p features measured on n independent units. For a given data set, D , and an integer, $K, 2 \leq K \leq N$, the clustering problem is to partition D into K disjoint clusters. In other words, $D = C_1 \cup C_2 \cup \dots \cup C_K$, with $C_j \cap C_k = \phi$ if $j \neq k$ and $C_j \neq \phi$ for $j \in \{1, \dots, K\}$. Thus, each cluster is consisting of points that are similar in some sense, and points of different clusters are dissimilar. Here, similarity means to be close in the sense of a distance $d(\mathbf{x}, \mathbf{y})$ between any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Note here that the number of clusters denoted by K should be determined beforehand. However, determining the “right” number of clusters is an important issue to be dealt with in clustering. Also note that different distance measures such as Euclidean

distance, Manhattan distance, and Mahalanobis distance can be used to measure the similarity between any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$.

There are many different clustering methods available in the literature [3]. In this study, well-known and easily applied ones, called center-based clustering techniques, were applied as a first attempt.

3.1 Center-Based Clustering Methods

Center-based clustering algorithms construct clusters by using the distances of data points from the cluster centers. The best-known and most commonly used center-based algorithm is the k-means algorithm [2], which minimizes the objective

$$\sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - c_k\|^2, \quad (1)$$

where c_k is the centroid of the k th cluster. The steps of the algorithm are as follows:

Step 0: Initialization: Given the data set D and an integer K , $2 \leq K \leq N$, select K initial centers $\{c_k\}$.

Step 1: Compute the distances $d(\mathbf{x}_i, \mathbf{c}_k)$, $i = 1, \dots, N$; $k = 1, \dots, K$, between the data points and cluster centers, and partition the data set by assigning each data point to a cluster whose center is the nearest.

Step 2: Recompute the cluster centers as the cluster mean points.

Step 3: If the centers have not changed, stop; else go to Step 1.

A major challenge in cluster analysis is to determine the correct or natural number of clusters. It may be possible to provide insight into the number of clusters in the data by using the classical “elbow method.” Figure 1 shows a typical graph of evaluation measure (SSE for our case) for k-means clustering method applied for minimum temperature versus the number of clusters employed. As one can see, the SSE shown below in (2) decreases monotonically as the number of clusters increases.

$$SSE = \sum_{k=1}^K \sum_{x_i \in C_k} d^2(c_i, x), \quad (2)$$

where c_i represents the center of the i th cluster.

We can try to find the natural number of clusters in a data set by looking for the number of clusters at which there is a knee in the plot [4]. In Fig. 1, there is a distinct knee in the SSE for k-means clustering applied for the minimum temperature when the number of cluster is equal to four. For the other variables, the optimal number of clusters listed in Table 1 is determined by using the same approach.

Several variants of k-means algorithm have been reported in the literature [1]. Some of them attempt to select a good initial partition so that the algorithm is more

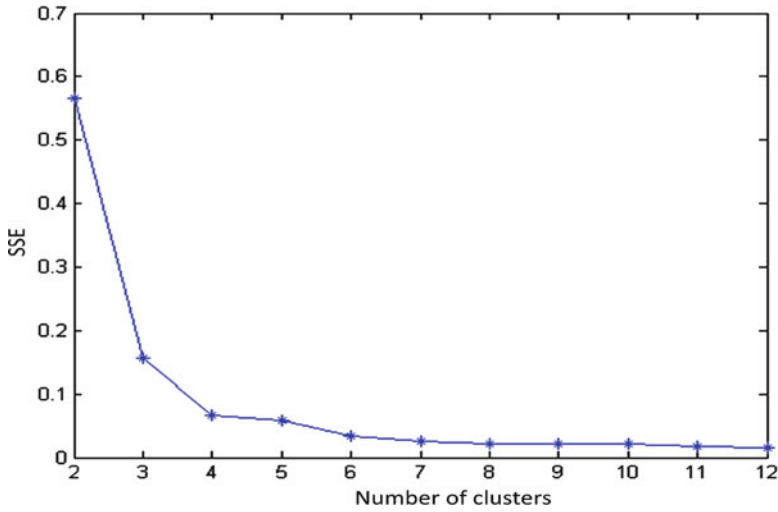


Fig. 1 Plot of SSE versus the number of clusters, K , for the minimum temperature

likely to find the global minimum value. The k-means algorithm can be adapted to soft clustering. A well-known center-based algorithm for soft clustering is the so-called fuzzy k-means algorithm in which the objective function to be minimized is

$$\sum_{i=1}^N \sum_{k=1}^K u_{ik}^m d_{ik}^2 = \sum_{i=1}^N \sum_{k=1}^K u_{ik}^m \|x_i - v_k\|^2. \quad (3)$$

Here, u_{ik}^m is the membership function of $x_i \in C_k$, and typically satisfy (4); $m > 1$ is a real number known as fuzzifier.

$$\sum_{k=1}^K u(x, C_k) = 1, \quad \text{and} \quad u(x, C_k) \geq 0 \quad \text{for all} \quad k = 1, \dots, K. \quad (4)$$

The same optimal number of clusters which were determined by using elbow approach for the k-means clustering method (in Table 1) is used for fuzzy k-means clustering; therefore, the assignment of the stations to the clusters can be compared for both methods. The clustering results obtained by using both methods are displayed in graphs like in Figs. 2 and 3. By using these graphs, clusters are examined, and the stations that fall into different clusters from their neighbors are determined, and the reasons are discussed. Additionally, all of the graphs obtained for each variable are compared. Clustering methods were also applied to the multivariate data set containing all temperature variables. The optimal number of clusters for this application is found to be five by using the elbow method for the k-means clustering approach and this number of cluster is also used for fuzzy k-means clustering method. As a result, the graphs obtained by using the k-means and the fuzzy k-means clustering methods are displayed in Figs. 2 and 3, respectively.

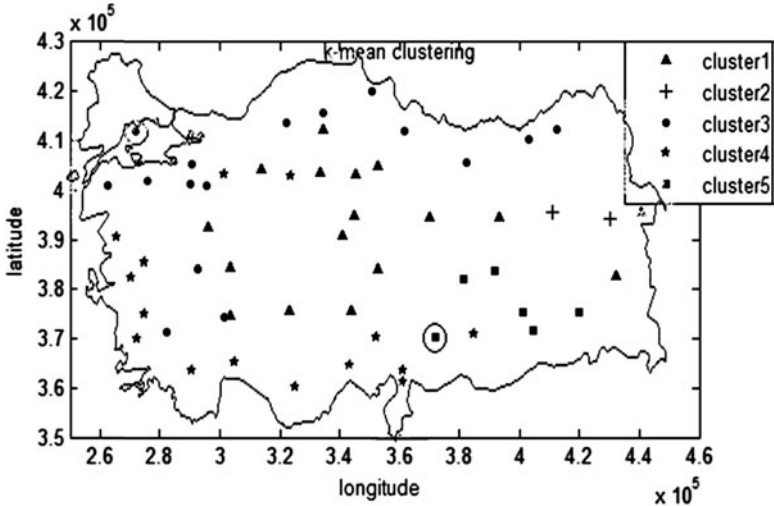


Fig. 2 k-means clustering graph

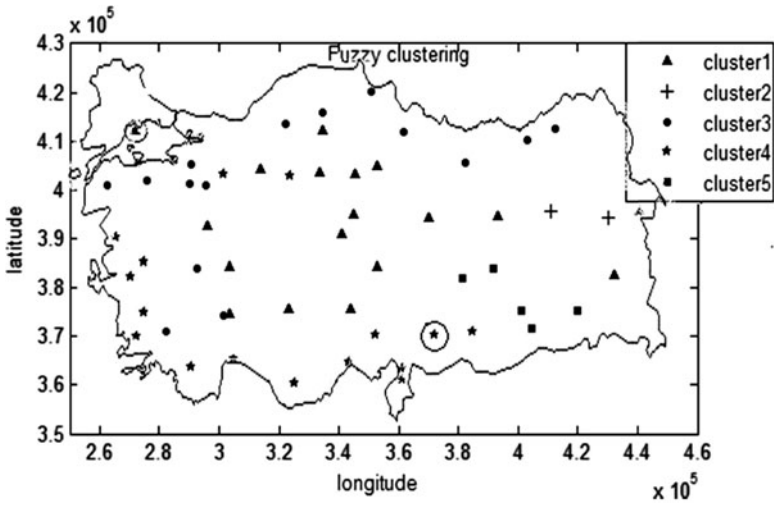


Fig. 3 Fuzzy k-means clustering graph

In both graphs, similar partitioning structures are obtained except two stations indicated by circles. This is not surprising, however, since the membership values of these two stations (i.e. Lüleburgaz and İslahiye) for two clusters are very close to each other.

4 Conclusion and Future Studies

There have been seven recognized climate zones of Turkey for many years. Black sea, Marmara, Aegean, Mediterranean, central Anatolian, eastern Anatolian, and southeastern Anatolian regions. Note here that in addition to climate differences, social and economic variables are also considered in identifying these zones. In this study, however, the climate zones of Turkey are reexamined by using a mathematical methodology of center-based clustering analysis. As a result of applying k-means and fuzzy k-means clustering methods on Turkish data, different clusterings results are obtained for each temperature variable. Moreover, when the multivariate cluster analysis is applied, five clusters are obtained over Turkey, and the partition of regions is found to be similar to the two methods studied, except two stations. As can be seen in Figs. 2 and 3, the Aegean region has the same properties as the Mediterranean and Southeastern Anatolian regions. In addition, in the Marmara region, Black Sea's climate effect become dominant except Trakya station, which is similar to the central Anatolian region. Moreover, two stations have different climate characteristics within the Eastern Anatolian region.

Because the clustering of the temperature variables does not consider the spatial properties of the stations, resulting clusters may not represent the correct partitioning. To increase the efficiency of the method bisecting k-means, which is less susceptible to initialization problems, can be applied. This method is an extension of the k-means algorithm that is based on simple idea: to obtain K clusters, split the set of all points into two clusters, select one of these clusters to split, and so on, until K clusters have been produced [4]. Actually, our main goal is to cluster the stations by considering their spatio and temporal properties. In order to accomplish our aim, as an initial step, bisecting and hierarchical clustering will be applied, and compared with k-means and fuzzy k-means results.

Acknowledgements This study has been supported by Middle East Technical University, Turkey, under the project number BAP-2008-01-09-02. Authors would like to thank all other NINLIL research group members for their support and contribution to this study.

References

1. Anderberg, M.R.: Cluster Analysis for Cluster Applications. Academic Press Inc., New York, NY (1973)
2. Hartigan, J.: Clustering Algorithms. Wiley and Sons Inc., New York, NY (1975)
3. Jain, A.K., Murty, M.M., Flynn, P.J.: Data clustering: a review. *ACM Computing Surveys*. **31**(3), 264–323 (1999)
4. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education Inc., Boston, MA (2006)
5. Türkeş, M.: Spatial and temporal analysis of annual rainfall variations in Turkey. *International Journal of Climatology*. **16**, 1057–1076 (1996a)

6. Türkeş, M.: Observed changes in maximum and minimum temperatures in Turkey. *International Journal of Climatology*. **16**, 463–477 (1996b)
7. Ünal, Y., Kindap, T., Karaca, M.: Redefining the climate zones of Turkey using cluster analysis. *International Journal of Climatology*. **23**, 1045–1055 (2003)

The Determination of Rainy Clouds Motion Using Optical Flow

O. Raaf and A. Adane

1 Introduction

For several years, rain forecasting in the short time limit has been a topic of meteorologist interest. To account for these phenomena, most meteorological networks are now equipped with automatic measurement systems and remote sensing tools [12]. In these conditions, different geostationary satellites used for meteorological observations, turn around the Earth in the equatorial plane since the last 30 years. A network of rain gauges gives the punctual and direct measures of rain rate on the ground so that it is possible to interpolate in order to reconstitute the complete rainy field. In this case, this operation is not very reliable when rain presents a large spatial variability.

In this case, radars, lidars, and sodars are other remote sensing systems used to collect meteorological data in the large-scale characterizing the air motions and cloud formations in the atmosphere [6, 9]. The radars equipping the meteorological stations are especially designed to detect rainy clouds. The radar equipment represents rain cells with: position, shape, intensity, dimensions, and displacement [9, 13].

Motion estimation from an image sequence is used in several applications. In robotics, it can identify and anticipate changes in the position of objects. In video compression, it allows the fullest possible understanding where the temporal redundancy of the sequence and information describing an image using the surrounding images. In meteorology, it allows the detection and forecasting of rainy clouds including those which are dangerous and predicts their motion.

O. Raaf (✉)

Faculty of Electronics and Computer Science, University of Science and Technology
Houari Boumediene (U.S.T.H.B.), PO Box 32, El Alia, Bab Ezzouar 16111, Algiers, Algeria
e-mail: rf_ouarda@yahoo.fr

The movement, in a sequence of images is visible through changes in spatial distribution of a photometric variable between two successive images, such as luminance, brightness, or reflectivity. This latter is the variable used in meteorological images. With the use of nonlinear algorithms, signal processing and modern mathematics have been opened to the study of such a random phenomenon.

In this context, we are interested in this chapter for using optical flow in the detection and extraction of the velocity of rainfall from radar images. Specifically, the part “low level” which provides local information on the speed as a field of velocity is treated. Given these objectives, this chapter will be organized as follows: first the database of the images used in this study is presented. Then, a formulation of the optical flow equation used to calculate the velocity fields in artificial images and weather radar images is given. After that, an application of the method on images from the databank is illustrated. Finally, the results of the application are discussed and interpreted.

2 Data Bank

The images to be studied in this chapter were collected with meteorological radar ASWR 81 (Algerian Service Weather Radar) installed in the city of Maghras in Setif ($36^{\circ}11\text{N}$, $05^{\circ}25\text{E}$, altitude 1,730 m) in the north-east of Algeria. This region is mostly known under the name of high plateaus. In Fig. 1, we present an example of a radar image taken in December 26th, 2004 at 17:30, where the radar is in the middle of the picture. In this figure, the fixed echoes reflected from mountains are clearly apparent.

Indeed, at the west of Setif we can discern an orange band due to the presence of regional mountain ranges known as Djurdjura where the highest elevation corresponds to Lalla Khadidja Mountain (2,308 m). At the south and at the north of the radar we can observe a bluish zone corresponding respectively to Bibans and Babor ranges. The resolution of this image is 1 km par pixel in PPI (Plan Position Indicator) representation. Its format is 512×512 pixels with 16 reflectivity levels. We can also see in this figure rain echoes represented as a quasi-homogeneous structure, while fixed echoes are scattered in small cells of nonhomogeneous intensity. The use of meteorological radars is delicate because of the earth relief, which produces masque effects that degrades noticeably rain echoes detection and reduces radar performance. It is therefore essential to remove this clutter to be able to follow the evolution of the precipitation echoes [4].

In order to eliminate the ground echoes that constitute a parasitic signal in the forecasting and the quantification of rain cells, a mask method combined with a shape detection is used [11]. To filter the radar images, templates representing the obstacles of the Earth surface are constructed and then employed to hide the ground clutter in these images. An image of templates is therefore obtained by averaging a large number of clear sky images. To improve the filtering method, an algorithm

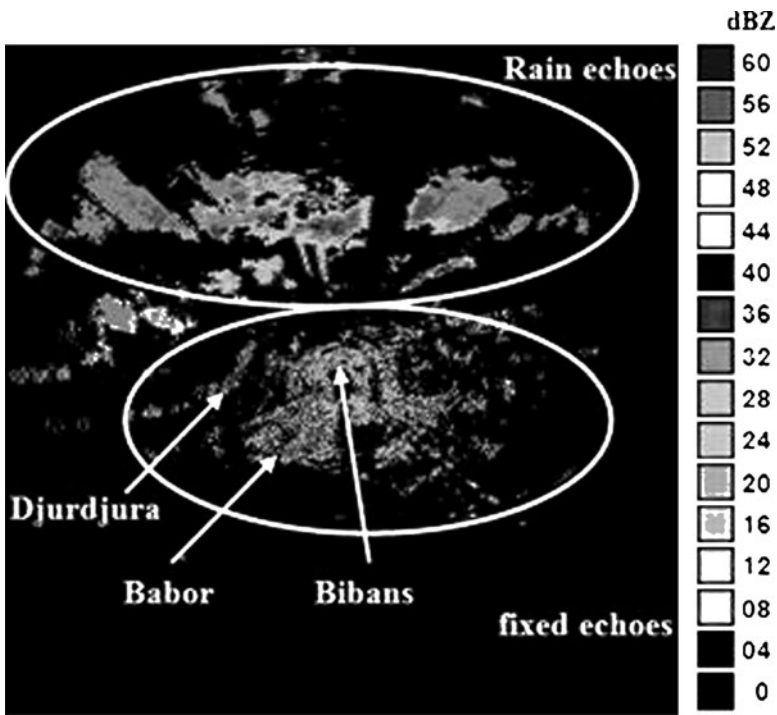


Fig. 1 Radar images collected at Setif during periods of bad weather

of pattern recognition has been developed. It consists in detecting the shape of the radar echoes and recording only those which have their surface greater than 30 km^2 . This algorithm operates by exploring the image, pixel by pixel, using an analysis window of 3×3 pixels. Figure 2 exhibits the images of rainfall echoes obtained by applying the templates and pattern recognition filters to the original images given in Fig. 1. The images so filtered show that the rainfall cells are properly reconstituted and all the undesirable ground echoes have practically been eliminated.

3 Identification and Tracking of Rainfall Clouds

3.1 Structure of Rain Clouds

The precipitation fields detected by weather radar are composed of a multitude of water droplets and ice crystals (with diameters lying between $0.1\mu\text{m}$ and 5 mm). The resulting clouds are heterogeneous structures classified into two main

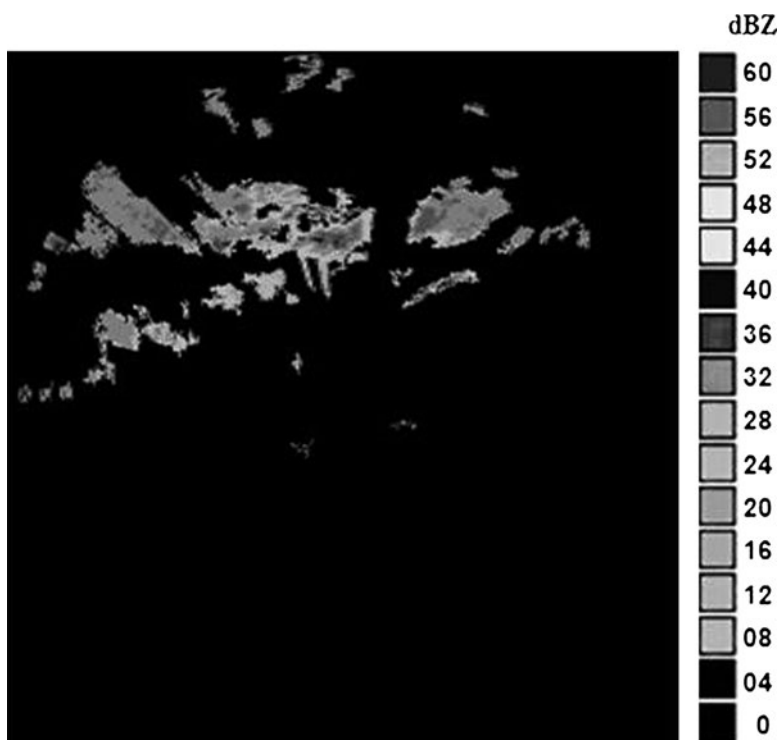


Fig. 2 Images of rainfall echoes for the regions of Setif, filtered using template and pattern recognition based-algorithm

categories: stratified clouds and convective cells. Briefly, stratiform clouds arise from the meeting of masses of warm and dry air with cool and wet atmospheric layers. These clouds essentially caused by advection, spread horizontally over large areas and then, yield rain of low intensity and long duration. Convective clouds are generated by subsidence due to the rise of strong and warm air masses and their condensation during the descent of cold air currents. Owing to the convective motion, these clouds extend mainly in the vertical direction and can even reach the top of troposphere. Most of them are dense and produce strong rainfall during short periods of time. Among the convective clouds, cumulonimbuses are known to be submitted to important atmospheric instabilities. Consequently, they frequently give rise to big thunderstorms and strong showers. Radar images can be processed so as to identify the rainy clouds, follow their motion, and get a good estimate of the intensity of precipitations watering the terrestrial surface.

3.2 Classification of Rain Clouds

The precipitation fields observed on the radar images are made up of several groups of cells cloud resulting from the agglomeration of hydrometeors in varied size and shape. Each cell is cloud-shaped variable and evolves separately for a period of life ranging from one to eight hours. The cells spread rain for most of the roughly 50–1,000 km² and are centered on a core of reflectivity to a greater or lesser extent. To characterize a cell and monitor cloud changes from one image to another, the radar images have been previously screened decomposed into elementary images each containing a single cell cloud. The algorithm of pattern recognition described in Sect. 2, enables us to separate a rainfall cell from the rest of the cloud cover, index it in the radar image and calculate its surface, gravity centre reflectivity and its direction.

In order to evaluate the precipitation fields more efficiently for each cell, and to know their evolution in time, a method of processing radar images presented hereafter, is implemented using mainly optical flow.

4 Motion Estimation Using Optical Flow

4.1 Optical Flow Equation

Motion estimation is based on the fact that each point on an image has a fixed intensity. It is therefore possible to estimate the motion of a point by superposing two consecutive images using the method of optical flow [1, 10]. A sequence of images can be represented by its brightness function $I(p, t)$. The hypothesis of brightness conservation says that a physical point of the image sequence does not vary with time, which gives:

$$I(p, t) = I(p + V(p) \delta t, t + \delta t), \quad (1)$$

where $p(x, y)$ is one point of the image and $V(p) = V(u, v)$ is the velocity vector at point p and time t , with u and v are, respectively, the velocity along x and y .

Equation (1) leads to the cancelation of the difference between the successive two images. Assuming that $I(x, y, t)$ is a continuous function together with its time derivative, the assumption of conservation of the luminance is given by:

$$\frac{dI}{dt} = \frac{\delta I}{\delta t} + \nabla I \cdot V = 0. \quad (2)$$

Where: $\nabla I = \left[\frac{\delta I}{\delta x}, \frac{\delta I}{\delta y} \right]$ is the spatial gradient of I .

Several models have been used to solve the equation of optical flow such as the model of translation, the parametric model, and the nonparametric model. In here we choose to use the nonparametric differential methods to determine the velocity field seen in a sequence of radar images.

4.2 Differential Method

The differential methods belong to the techniques commonly used for the calculation of optical flow in image sequences. Their advantages include the reduction in the complexity of calculations commonly used in the matching methods while increasing the range of measurable displacements. They can be classified into two categories: global methods such as the technique of Horn and Schunck, and local methods as the Lucas–Kanade approach and that based on Gabor filters. These methods are based on the equation of the apparent motion (ECMA) issue to Taylor development of the equation of the intensity $I(x, y, t)$ described by:

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\delta I}{\delta x} dx + \frac{\delta I}{\delta y} dy + \frac{\delta I}{\delta t} dt + o(\varepsilon), \quad (3)$$

where ε is the term of higher order of the first order Taylor development of I which tends to 0 when dt tends to 0. Considering that the quantities dx , dy , and dt is sufficiently small and that (1) is respected, we have:

$$uI_x + vI_y + I_t = 0, \quad (4)$$

I_x , I_y , I_t : spatiotemporal derivative of intensity at point $p(x, y)$ at time t .

To solve this equation by respecting the conditions of validity, two solutions are proposed in this study: the Lucas–Kanade method and Gabor filters.

4.2.1 Lucas and Kanade Method

The method proposed by Lucas and Kanade is a local approach based on the assumption that the motion is uniform over a region of the image or that the optical flow is constant over Ω neighborhood centered on the pixel that we want to calculate the displacement. We are therefore led to minimize the following functional [2, 8]:

$$E = \sum_{\Omega} [\nabla I \cdot V(p_i) + I_t]^2, \quad (5)$$

where Ω is a window around the point at we wish to determine the displacement.

In this method, each calculation is done on a small window in parallel independently of other windows. The solution is to make an estimate in the sense of least squares. In matrix notation the solution is:

$$V = (A^T A)^{-1} A^T b. \quad (6)$$

With: $A = \begin{bmatrix} I_{x1} & I_{y1} \\ I_{x2} & I_{y2} \\ \vdots & \vdots \\ I_{xn} & I_{yn} \end{bmatrix}, \quad b = \begin{bmatrix} I_{t1} \\ I_{t2} \\ \vdots \\ I_{tn} \end{bmatrix}.$

It may happen that the matrix $A^T A$ is ill conditioned (almost zero determinant and therefore not invertible). This is because the pixel is in an area where the luminance intensity is constant with zero gradient (problem opening) leading to the estimate in the sense of least squares of V to become aberrant. We can remedy these problems by using a regularization technique. This is done by using:

$$E = \sum_{\Omega} [\nabla I \cdot V(p_i) + I_t]^2 + \alpha V^2. \quad (7)$$

With α (adjustable) representing the regularity of the solution. This equation is always linear and can yield to the following:

$$V = (A^T A + \alpha J)^{-1} A^T b, \quad (8)$$

where J is the identity matrix.

The developed optical flow equations are valid only if the displacement is small, so we have to use the multi-resolution developed in the algorithm of Horn and Schunck to solve them. On the other hand, it is possible to further refine the results at each level of the pyramid minimizing the gap between two successive frames and running again the algorithm after moving one of two images according to the latest calculated field velocity. If the algorithm is stable, the method converges and therefore:

$$V^l = V^{l-1} + \eta^l, \quad (9)$$

with $V^0 = 0$ and $\eta^l = (A^T A + \alpha J)^{-1} A^T b_l$.

If L iterations are necessary to achieve convergence, the final solution is the last calculated vector and is given by:

$$V = \sum_{l=1}^L V^l. \quad (10)$$

4.2.2 The Gabor Filters Approach

The technique involves estimating the optic flow locally based on sub-band decomposition of the image sequence. The estimate was made by Gabor filters space by combining information from filter banks positioned at different orientations to robustly estimate the motion. Gabor filters are widely used in computer vision. This success is due both to their spatial and spectral properties. Bruno [3] proposed a novel approach based on the differential method for local estimation of optical flow from Gabor filter banks. It is based on projecting the optical flow equation (21.4) into a pixel image on a bank of N Gabor filters. We obtain an over determined system of N equations that allows us to calculate the two velocity components of the pixel considered.

Assuming that V is constant over the filter support G_i , we can write the optical flow equation as:

$$u \left(\frac{\delta I}{\delta x} G_i \right) + v \left(\frac{\delta I}{\delta y} G_i \right) + \frac{\delta I}{\delta t} G_i = 0 \quad i = 1 \dots N. \quad (11)$$

The minimization is obtained iteratively by using the Weighted Least Squares algorithm (WLS):

$$V = \arg \min \sum_i w_i \rho(r_i). \quad (12)$$

With $w(x) = \frac{1}{x} \frac{\delta \rho(x)}{\delta x}$ is a weighting function and $\rho(x)$ is a Euclidean norm.

The function $\rho(x)$ can be chosen from a wide range of functions. It must be even defined, positive, and should not have a single minimum at zero. In addition, in order for the function $\rho(x)$ to minimize the influence of erroneous data, it has to be less than the quadratic function. In the case of a local estimate, it is preferable that the aberrant data be deleted. So we chose to use the M-estimator [5].

5 Application

The Lucas and Kanade together with the Gabor algorithm were tested on three different sequence images: a cube, taxi, and a rainy cell (Figs. 2a, 3a, and 4a, b).

- Figure 3a represents the image of an artificial sequence called “cube” consisting of a cube placed on a turntable plateau. The image size is 256×240 pixels.
- Figure 4a represents the image of an artificial sequence called “taxi” in a street intersection. It is composed of three cars and a pedestrian moving. The two bottom vehicles moving in the opposite direction. The taxi turned the corner in the middle of the street. The image size is 256×190 pixels.
- Figure 5a, b represent the sequences of two stratiform cells isolated from images of size 64×64 pixels. It represents one of the rain cells observed in the radar image taken in December 26th, 2004 at 17:30 (Fig. 4a) and 15 min later (Fig. 4b). The first cell has $2,685 \text{ km}^2$ of surface and 19.44 dBZ of reflectivity.

6 Results and Interpretations

6.1 Lucas and Kanade Method

The regularization coefficient α (α) is simply to avoid cancellation of the determinant in the least squares estimation. Its value should be close to zero, between 10^{-3} and 10^{-6} . The gradient that give the best results was calculated using the mask $[-1 \ 1, \ -1 \ 1]$ for a horizontal displacement x , and $[-1 \ -1, \ 1 \ 1]$ for

Fig. 3 Optical flow estimated for cube image by patch 7×7 . **(a)** Cube image; **(b)** field estimated by Lucas–Kanade; **(c)** field estimated by Gabor filters

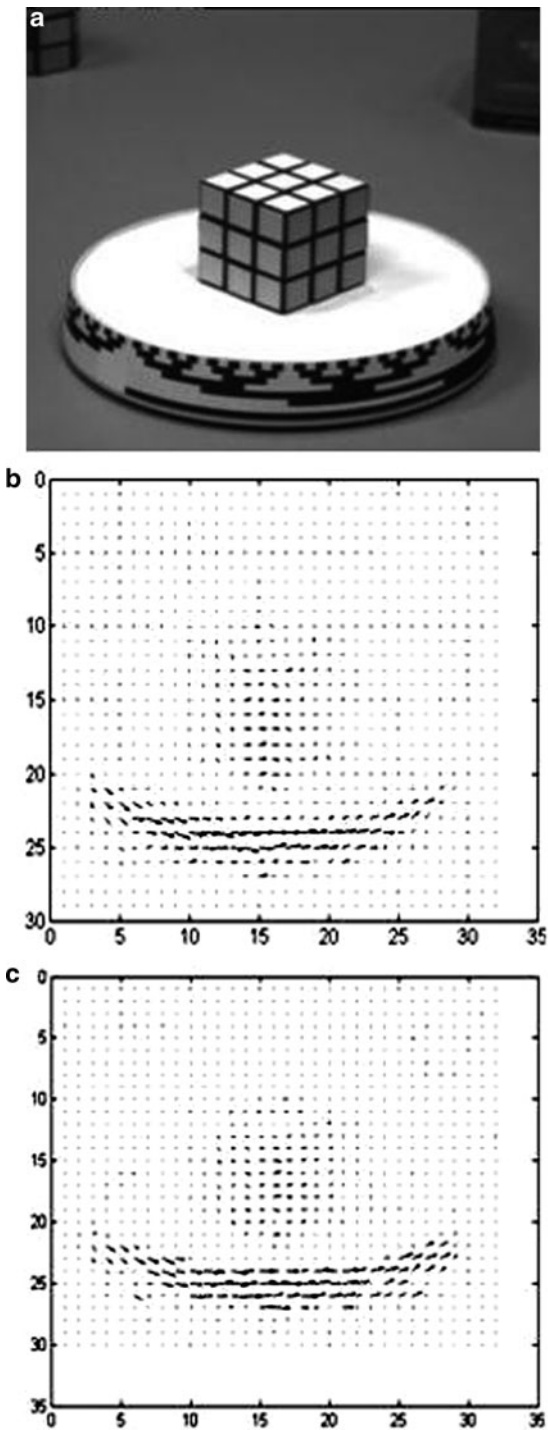
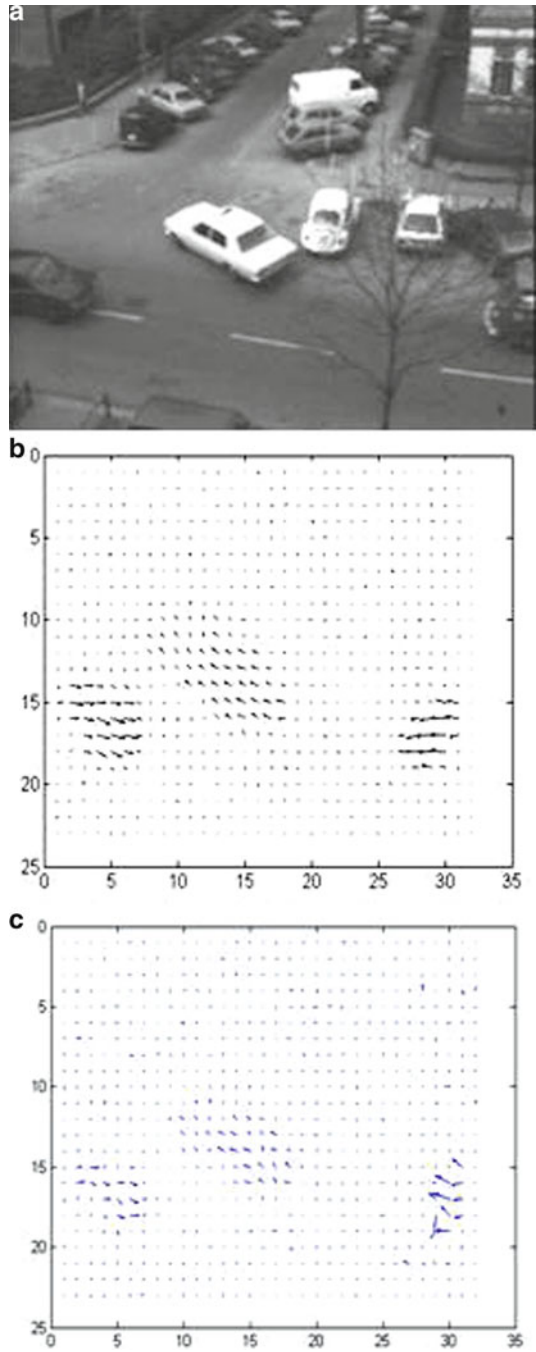


Fig. 4 Optical flow estimated for taxi image by patch 7×7 . **(a)** Taxi image; **(b)** field estimated by Lucas–Kanade; **(c)** field estimated by Gabor filters



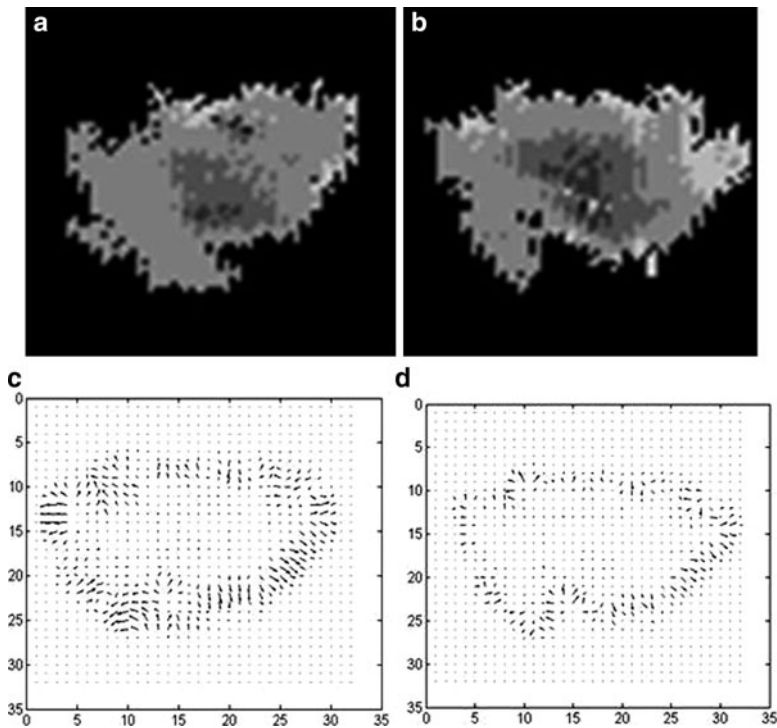


Fig. 5 Optical flow estimated for rainy cell. (a) First cell of rain cloud; (b) second cell of rain cloud; (c) field estimated by Lucas–Kanade method; (d) field estimated by Gabor filters

vertical displacement y . After several trial tests, the best results were obtained under the following conditions:

- The cube level of the pyramid 3, patch 7×7 pixels, $\alpha = 10^{-3}$: the velocity field in Fig. 2b shows the rotation of the plateau.
- Taxi: level of the pyramid 3, patch 7×7 pixels, $\alpha = 10^{-3}$: Fig. 3b shows three cars moving and that both gray cars are moving in opposite directions and the white taxi in the center turns on the corner. The movement of the pedestrian is not detected.
- Rainy cell: pyramid level 1 patch of 5×5 pixels, $\alpha = 10^{-3}$: In Fig. 4c appears vectors almost directed to the right indicating the direction of development of the rainy cloud. We also note that the estimated motion is almost zero at the center of the cell and increases gradually as we come near the edges, which is logical because the cloud changes little in 15 min. Outside the rain cell the field is considered zero.

The size of the images used in the sequences “Cube” and “taxi” is important, which justifies the use of third-level pyramid. Concerning the cell extracted from the radar image, first-level pyramid is used directly given the size of the image (5×5 pixels).

6.2 Gabor Filter Method

The estimated flow shows all the objects in motion in the three image sequences. The tuning parameters of this algorithm do not differ greatly in all three cases, thus:

- Cube (Fig. 2c): $K = 3$, $\alpha = 10^{-3}$, $\sigma = 4$, $f_0 = 0.14$.
- Taxi (Fig. 3c): $K = 3$, $\alpha = 10^{-3}$, $\sigma = 4$, $f_0 = 0.18$.
- Rain cell (Fig. 4d): $K = 1$, $\alpha = 10^{-3}$, $\sigma = 4$, $f_0 = 0.14$.

We also note that the density increases with velocity vectors N to reach a high where it does not change, for $N = 6$. However, the increment of N greatly increases the computing time. The density also increases with σ up to a maximum value beyond which we obtain a smooth velocity field, since the assumption of locally constant optical flow is no longer respected.

We also note that variations of the frequency centre are very low. This is due to that the global variations of the considered photometric variable (brightness or reflectivity) are low-frequency type.

On setting the parameters for obtaining good results, there is a trade-off between α and the number of iterations since the higher number of iteration gives a better motion boundaries area.

In these three examples, we have considered a single iteration loop for refinement: i.e., $L = 1$. Increasing L will have little influence on the velocity fields of the moving objects. Tests on the size of the patch showed that whenever the window size is increased, the fields motion appear, which is more and more dense until reaching an optimum value. Thus, if we continue to increase the window size, the optical flow fields became smooth, i.e. the contour areas will begin to disappear and the movement becomes less clear.

7 Conclusions

In this chapter, we have presented two techniques for estimating local motion of rainy cells in which the quality of estimation depends mainly on the right choice of the parameters in each method. In both cases and when applied on three different sequence images, the estimated optical flow detected all the moving objects and their direction. We also found that the method based on Gabor filters gives a less dense optical flow with a lack of precision in the estimated field near the border. This was in contrast with the method of Lucas and Kanade. Moreover, it was found that the computing time required to run the algorithm of the latter is much smaller than that of the Gabor filters. Based on this, we can say that optical flow can be used successfully to determine rainy clouds motion in weather radar images.

References

1. Baron J.L., Fleet DJ. and Beauchemin SS: Performance of optical flow techniques. *In International Journal on Computer Vision*, vol. 12, pp. 43–77 (1994).
2. Bouguet J.: Pyramidal Implementation of the Lucas Kanade Feature tracker, Intel Corporation, Microprocessor Research Labs (2000).
3. Bruno E: De l'estimation locale à l'estimation globale de mouvement dans les séquences d'images. Thèse, Université Joseph Fourier, Grenoble, France (2001).
4. Darricaud, J.(Ed). Physique et theorie du radar, 3^{ème} ed., t.2: principes et performances de base (Physics and theory of the radar, 3rd ed., t.2: Basic principles and performance). Paris: Sodipe (1993).
5. Huber P.: Robust statistics, Wiley(Ed), New York (1981).
6. CRANE, R.K. (Ed): Electromagnetic wave propagation through rain, 269 p. New York: J. Wiley and sons (1996).
7. Keeler, R.J. and Passarelli, R.E.: Signal processing for atmospheric radars. In *Radar probing and measurement of the planetary boundary layer*; D.H. Lenschow (Ed.), pp.199–229.Boston: American Meteorological Society (1986).
8. Mrazat J.: Estimation temps réel du flot optique; Rapport de stage ingénieur; INRIA, Yvelines (2008).
9. Meischner, A.: Weather radar: principles and advanced applications. Springer (2005).
10. Mémin E.: Estimation du flot optique-contribution et panorama des différents approches; Habilitation à diriger des recherches. Université de Rennes 1, Rennes (2003).
11. Raaf, O. and Adane, A.: Image-Filtering Techniques for meteorological radar. *In International Symposium on Industrial Electronics (IEEE ISIE08)*, 30 June-02 July 2008, Cambridge, pp. 2561–2565.Sauvageot, H.(1992). Radar Meteorology. Artech House(Ed), Boston (2008).
12. Renaut, D.: Les satellites meteorologiques (Meteorological satellites). *La meteorologie*, **45**, pp. 33–37 (2004).
13. Sauvageot, H.: Radar Meteorology, 366p.Boston: Artech House (1992).

Hydrodynamic Modeling of Port Foster, Deception Island (Antarctica)

Juan Vidal, Manuel Berrocoso, and Bismarck Jigena

1 Introduction

Deception Island is part of the South Shetland Island chain with Smith, Snow and Livingston Islands to the north and west, and King George Island to the north and east. These islands form the northern boundary of the Bransfield Strait and the Antarctic Peninsula forms its southern boundary. Deception Island is situated between latitudes $62.89\text{--}63.02^\circ$ South and longitudes $60.49\text{--}60.75^\circ$ West. Deception Island is an active volcano that has a central flooded caldera, which is open onto Bransfield Strait through a shallow and narrow sill, at Neptune's Bellows [17].

Bransfield Strait tides are a mixture of diurnal and semidiurnal frequencies. The M2 component is the most important with an amplitude around 0.40 m while O1, K1 and S2 tidal components, all with amplitudes around 0.28 m, are also significant [13]. Amplitudes of the main tidal constituents are higher in the northwestern Weddell Sea than at the northwestern side of the peninsula [7]. The local circulation in the Bransfield Strait is strongly influenced by tides. Both at the surface and at deeper layer, the flow patterns exhibit a strong diurnal and semidiurnal periodicity in agreement with the barotropic character of tides [12]. The Bransfield Current is a northeastward geostrophic current. Based on geostrophic studies calculated to a reference level, various authors proved that the currents are oriented parallel to the front with speeds from 0.05 to 0.30 m s^{-1} , [14] and [8].

J. Vidal (✉)

Centro Andaluz de Ciencia y Tecnologías Marina, Campus Universitario de Puerto Real,
Puerto Real (Cadiz), 11510 Spain
e-mail: juan.vidal@uca.es

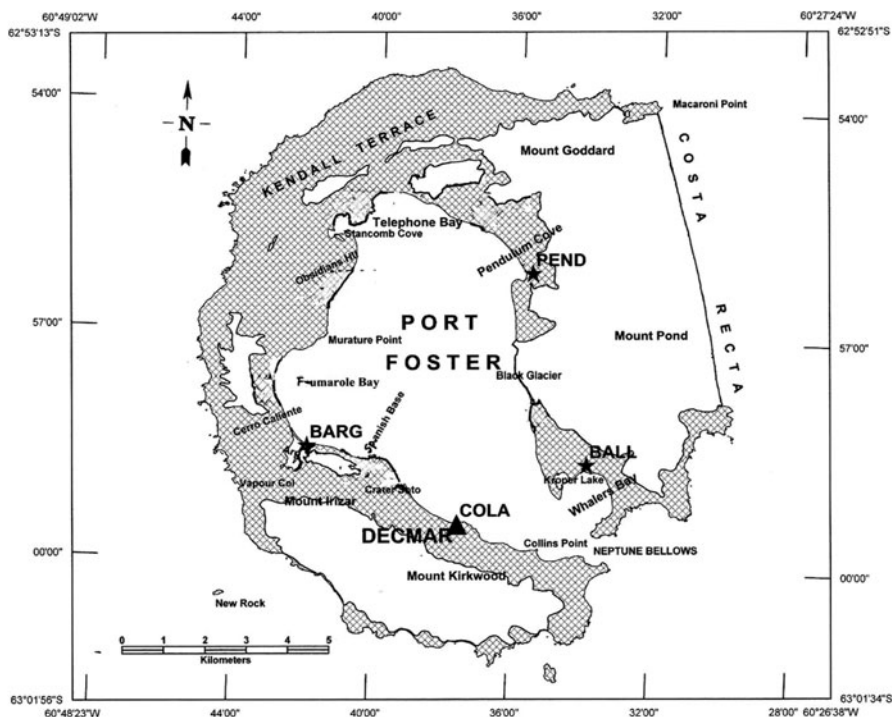


Fig. 1 Tidal Stations in Deception Island: (BALL) Balleneros Station, (COLA) Colatinas Station and (PEND) Pendulo Station

However, Port Foster water mass characteristics differ significantly from Bransfield Strait water mass characteristics, indicating that the sill at Neptune's Bellows limits cross-sill exchange below 30 m. The small tidal currents in Port Foster, less than 0.1 m s^{-1} , also suggest that the larger Bransfield Strait tidal currents do not direct water into Deception Island [11]. The maximum tidal currents are obtained at Neptune's Bellows, with 0.64 m s^{-1} during spring tides [4].

Several hydrodynamical numeric models, including the Bransfield Strait have been recently implemented by several authors, [7] and [16]. However, these models are larger in scale. The main objectives of this paper are to present a detailed study of the tidal characteristics in Port Foster. The amplitudes and phases obtained through the analysis of the time series generated by the model allowed the attainment of co-tidal maps and ellipses of currents maps; these results are compared with the harmonic constants of several coastal stations, which positions are given in Fig. 1. The tidal simulations are a first step toward the modeling of the hydrodynamic flushing.

2 Model Approach

2.1 Governing Equations

The hydrodynamic model UCA2D here applied has been developed at University of Cadiz. A completed description of the model can be found in [2]. It has already been applied successfully to several coastal environments. See [1, 3, 18].

The model resolves the equations in their formulations with water levels and transports:

$$\frac{\partial U}{\partial t} - fV + gH \frac{\partial \zeta}{\partial x} + RU + X = 0 \quad (1)$$

$$\frac{\partial V}{\partial t} + fU + gH \frac{\partial \zeta}{\partial y} + RV + Y = 0 \quad (2)$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} = 0, \quad (3)$$

where f is the Coriolis parameter, ζ is the water level, g is the gravitational acceleration, $H = h + \zeta$ is the total water depth, and U and V are the vertically integrated velocities. R is the friction coefficient which is expressed as:

$$R = C_b \frac{\sqrt{u^2 + v^2}}{H}, \quad (4)$$

where C_b is the bottom drag coefficient. The terms X and Y of equations contain the nonlinear terms:

$$X = u \frac{\partial U}{\partial x} + v \frac{\partial U}{\partial y} - A_H \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} \right) \quad (5)$$

$$Y = u \frac{\partial V}{\partial x} + v \frac{\partial V}{\partial y} - A_H \left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \right), \quad (6)$$

where the last terms in equations represent the horizontal turbulent diffusion and A_H is the horizontal eddy viscosity.

At the open boundaries, the water levels are prescribed in accordance with the Dirichlet condition, while at the closed boundaries the normal velocity is set to zero and the tangential velocity is a free parameter.

2.2 Residence Times and Water Exchange

The residence time may be defined as the time it takes for a particular water parcel to leave a water body through its inlet. However, if the residence time is interpreted as an average, and does not refer to individual water parcels, it may be expressed as

$T_r = V/Q$, where V is the volume of the water body and Q is the water exchange (sum of the net freshwater supply and exchange with the open sea), [15] and [20].

Hydrodynamic flushing feature in an estuary is a measure of its self-purification capability and it determines the renewal of ocean water. The flushing time is equivalent to the reposition time [5], also known as replacement time. The definition is strictly oriented at the time needed for the volume introduced by the ocean flux to be equal to the fresh water volume in the bay, but says nothing about the processes of replacement.

Conventionally, estuarine flushing time (conversely flushing rate) is computed by simplified approaches, such as the tidal prism or salt balance method [10]. We have carried out a study of its flushing characteristics with a numerical tracer experiment by solving the vertically integrated mass transport equation.

3 Application to Port Foster

3.1 The Model Setup

The numerical computation has been carried out on a spatial domain that represents Port Foster through a finite element grid, which consists of 7,980 rectangular elements with a resolution of 100×100 m (Fig. 2). The bottom drag coefficient has been set to $cd = 3.0 \times 10^{-3}$. In barotropics models, the bottom drag coefficient is usually determined by fitting the modeled M2 tidal elevation and the observations as tide gauge stations [6]. All simulations presented in this work have been carried out using a maximum time step of 5 s.

The open boundary oscillations of the model consider components M2, S2, O1, and K1 that represent more than 90% of the tidal energy in Port Foster [13]. We use the amplitude and phase of the major tidal constituents from Antarctic Tide Gauge Database (<http://www.esr.org>) at the oceanic border.

For simulations of tidal velocities, all variables, including sea levels and velocities, were initially set to zero and the model was forced at the open boundary by specifying the sea level as above.

Although the seasonal cycle is the dominant signal in the temperature, characterized by the onset of stratification and deepening of the thermocline, using a two-dimensional model, we assumed that the water column is well mixed. This situation occurs during the austral winter [14]. Corresponding wind stress and spatial salinity variation were omitted. For the horizontal turbulent diffusivity, a low value of $3 \text{ m}^2 \text{ s}^{-1}$ has been chosen.

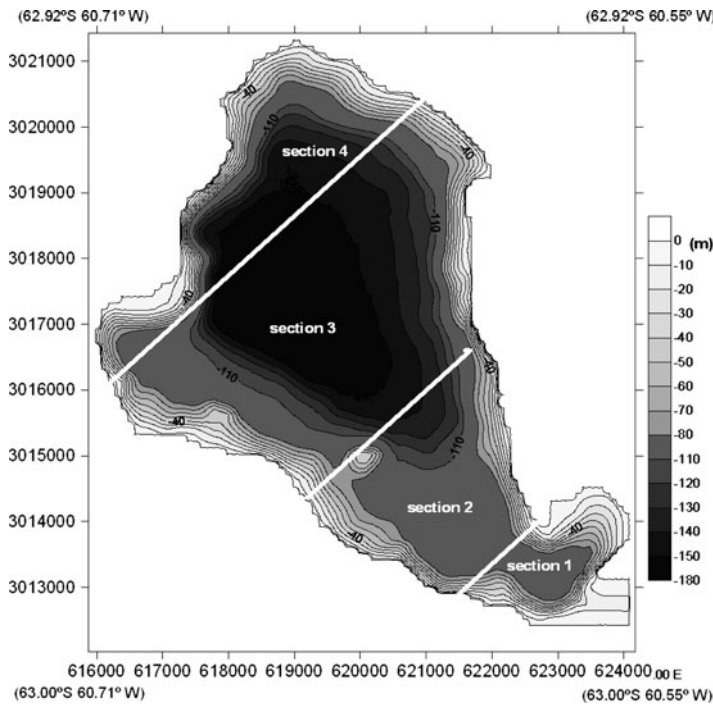


Fig. 2 Grid of Port Foster

3.2 Tidal and Current Measurements for Model Calibration

The model has been validated comparing the amplitudes and phases of the major components generated by the model with those derived from harmonic analysis of the tide gauge data within the island. We used the results of the harmonic analysis of series of direct sea levels measurements at Pendulo and Balleneros stations studied by Dragani et al. [7]. Additionally, water measurements were undertaken from December to March 2008, using a recording tide gauges. The tide gauges was deployed at Colatina (DECMAR station) [19]. The accuracy given by the pressure sensor is 1 cm. Harmonic analysis applied to tide gauge data is also shown in Table 1. Current measurements were carried out during four days at Colatinas station. However, for security reasons of the instrument, it was deployed too close to the shore about mid-depth at a total depth of approximately 8 m. Harmonic analysis was carried out also on the current measurements from Colatina.

Table 1 Elevations; harmonic analysis results

Station	Ball			Pend			Cola		
	A(m)		θ (degrees)	A (m)		θ (degrees)	A(m)		θ (degrees)
	O	M		O	M		O	M	
M2	0.46	0.39	280	0.44	0.39	281	0.40	0.40	281
S2	0.28	0.26	x	0.29	0.26	x	0.26	0.26	351
O1	0.29	0.27	48	0.29	0.27	55	0.27	0.27	53
K1	0.26	0.30	66	0.26	0.30	73	0.30	0.30	74
Error		0.03			0.03			0.00	

Observed (O) and calculated (M) amplitudes and phase lags at Balleneros (Ball), Pendulo (Pen) and Colatinas (Cola) stations.

For each constituent of each simulation, the error in the amplitudes (e_A) and the error in the phases (e_F) have been calculated by using:

$$e_A = \sqrt{\frac{1}{N} \sum_i (A_{oi} - A_{mi})^2 / 2} \quad (7)$$

$$e_F = \frac{1}{N} \sum_i A_{mi}^2 (\theta_{oi} - \theta_{mi}) / (\sum_i A_{mi}^2). \quad (8)$$

The error in the amplitudes was calculated from the difference between the observations (A_o) and the results (A_m) of the model, and the error in the phases was calculated from their difference ($\theta_{oi} - \theta_{mi}$), but weighted with the values of the observed amplitudes (A_m).

3.3 Modeling of Tidal Flushing

Based on the validated hydrodynamic model, a numerical dye tracer experiment is performed. The entire bay is initially labeled (at high water) at a uniform 100% concentration with a passive tracer at time $t = 0$. The subsequent change in the tracer mass in different parts of the bay is then tracked by solving the 2D mass transport equation. At the open boundary, the inflow tracer dye concentration is 0. The computed flushing time is taken as the time when mass removal falls below 1/e of the original vertically integrated mass.

4 Results

Table 1 presents the comparison of the analysis of model results (M) with harmonic constants (O) of coastal stations (Balleneros, Pendulo, and Colatinas). The tidal amplitudes are in meters and the phases angle are in degrees relative to local time.

Model results show that the amplitudes of the components decrease as they spread into the caldera from the ocean through Neptune's Bellows. The phase lags are also important in this area due to the increased friction. Inside Port Foster, where depths are around 100 m deep, the amplitudes and phase lags do not vary significantly.

One of the most important results obtained by modeling is that the amplitudes of the velocities are very low (less than 5 cm s^{-1}) within Port Foster, excluding the area of Neptune's Bellows. Amplitudes and phase lags for the M_2 constituent are shown in Fig. 3. The velocities near the narrow channel of Neptune's Bellows show that the east–west direction is more important than the north–south, coinciding with the orientation of the narrow passage. For this area, through which it produces all the water exchange between the caldera and the open ocean, the results for simulations

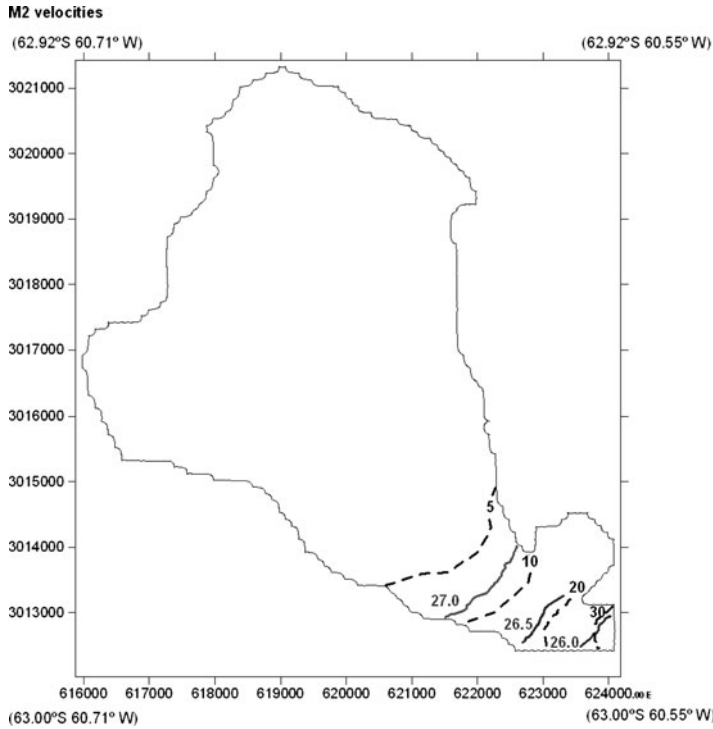


Fig. 3 Tidal velocities: amplitudes (*black dashed line*, in cm s^{-2}) and phases lags (*gray solid line*, in degrees) for M_2 constituent

of tidal velocities show maximum speeds of 0.76 m s^{-1} during spring tides. Table 2 shows the amplitudes of the velocities for the east–west components of the main harmonic constituents in a center point of the narrow and near Cola Station.

The volume of the basin is approximately $3,500 \text{ Hm}^3$ and the cross-section at Neptune’s Bellows is about $5,500 \text{ m}^2$. Figure 4 shows tidal elevation data obtained by the model at Neptune’s Bellows. A mean tidal level variation is about 1.20 m . Tidal elevation data obtained by the model are shown in Fig. 4. A height of 1.20 m implies that about 1.15% of the total volume of Port Foster is exchanged with Bransfield Strait during the flood and the ebb tide. This implies a exchange of 42 Hm^3 during filling or emptying of the Bay (in this case the predominant semi-diurnal tide cycle of 12.42 h), obtaining a calculated residence time of 82 days. Additionally, the total sea water volume V_d within the bay can be estimated as [9]:

$$V_d = \left(1 - \frac{S_m}{S_o}\right) V_l \quad (9)$$

where S_m is the mean salinity of the bay, S_o is the sea water salinity entering through the bay and V_l is the bay volume at high water. According to Lenn et al. [11], the

Table 2 Velocities: harmonic analysis results

Station	Neptuno’s Bellows		Cola			
	<u>A (m s⁻¹)</u>		<u>A (m s⁻¹)</u>			
	<u>θ (degrees)</u>		<u>θ (degrees)</u>			
	M	M	O	M	O	M
M2	0.40	27	0.13	0.10	33	36
S2	0.25	90	x	0.07	x	98
O1	0.13	148	x	0.05	x	156
K1	0.15	166	0.09	0.06	173	179
Error				0.02		2

Observed (O) and calculated (M) amplitudes and phase lags at Neptuno’s Bellows and Pendulo Colatinas (Cola) stations.

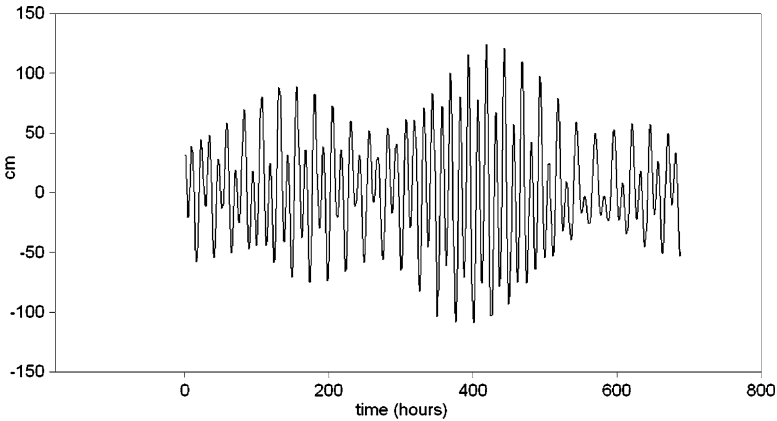


Fig. 4 Tidal elevation data obtained by the model

mean salinity in Port Foster is about 34.04 psu (practical salinity units). The mean salinity in the wester and central basin of the Bransfield strait is 34.4 psu [12]. For salinity to remain in a steady state, the volume of sea water required to this is 40.7 Hm³, similar to the previous result. The maximum range, defined as the largest difference (at spring) between a maximum and the next minimum, is around 2.20 m. For the case of spring tides, this amount is approximately 2.2% of the total volume. The residence time is 45 days.

To compute the different flushing feature in Deception Bay, the entire bay water was grouped into four segments as it is shown in Fig. 2. Using the typical tidal and associated salinity conditions, the flushing time was calculated as shown in Table 3. A quick renewal occurs in the lower bay, because the water flushed out in a few days. The upper bay is characterized by large flushing time, because the velocities are weaker-less than 0.1 m s⁻¹.

Table 3 Flushing time in deception bay

Section	Days
1	2.7
2	16.1
3	51.2
4	78.0

5 Conclusion

The tide in Port Foster was satisfactorily reproduced for all cases studied in this work. Currents obtained with the model are very small inside, with maximum values of 0.76 m s^{-1} in the area of the Neptune's Bellows. Neptune's Bellows is the region where most of the tidal energy is dissipated. Approximately 90% of the total energy dissipates in this area.

The residual circulation due to tides is very small in the center of the caldera where the residence times are more than 35 days. Port Foster in Deception Island has a rate of poor water exchange (between 1.1% and 2.2% volume exchange over each tidal cycle). The two-dimensional model results presented here are in general consistent with previous three-dimensional model experiments on Deception. However, in a two-dimensional model, the absence of the vertical dispersion of the tracer may lessen the spread of the tracer patch. To implement a realistic simulation of the dye trace experiment, a three-dimensional model must be employed with the real hydrography and topography.

Acknowledgements This work was possible thanks to the project CGL2005-0789-C03-01/ANT INVESTIGACIONES GEODESICAS, GEOFISICAS Y DE TELEDETECCION EN LA ISLA DECEPCION Y SU ENTRONO (PENINSULA ANTARTICA, ISLAS SHETLAND DEL SUR), financed by Spanish Ministry of Science and Technology through the National Program of Antarctic Research of Natural Resources. We also thank the Las Palmas Crew and the members of the Spanish Antarctic Station Gabriel de Castilla for their collaboration during the surveying campaigns.

References

1. Alvarez, O., Tejedor, B., Tejedor, L.: Simulación hidrodinámica en el área de la Bahía de Cádiz: Análisis de las constituyentes principales. IV Jornadas Españolas de Puertos y Costas, pp. 125–136. Servicio de Publicaciones de la Universidad Politécnica de Valencia, Valencia (1997)
2. Alvarez, O.: Simulación numérica de la dinámica de marea en la Bahía de Cádiz: análisis de las constituyentes principales, interacción marea-brisa e influencia del sedimento en suspensión. Ph. D. Thesis, Universidad de Cádiz, Cádiz (1999)
3. Alvarez, O., Izquierdo, A., Tejedor, B., Maanes, R., Tejedor, L., Kagan, B. A.: The Influence of Sediment Load on Tidal Dynamics, a Case Study: Cadiz Bay. *Estuarine, Coastal and Shelf Science*. **48**, 439–450 (1999)
4. Antarctic Pilot: Comprising the coasts of Antarctica and all Islands Southward of the Usual Route of Vessels, 4th Edition. Hydrographer of the Navy. Taunton, Somerset (1994)

5. Bolin, B., Rodhe, H.: A note on the concepts of age distribution and transit time in natural reservoirs. *Tellus* **25**, 58–62 (1973)
6. Crean, P.B., Murty, T. S., Stronach, J. A.: Mathematical Modelling of tides and Estuarine Circulation: the Coastal Seas of Southern British Columbia and Washington State. *Coastal Estuarine Stud.* **30**, pp.1-47, Washington (1991)
7. Dragani, W. C., D'onofrio, E. E., Fiore, M. E., Speroni, J. O.: Propagación y amplificación de la marea en el sector norte de la Península Antártica. V Simposio Argentino y Latinoamericano sobre Investigaciones Antárticas, Buenos Aires (2004)
8. Garcia, M. A., Castro, C. G., Rios, A. F., Doval, M. D., Roson, G., Gomis, D., Lopez, O.: Water masses and distribution of physico-chemical properties in the western Bransfield Strait and Gerlache Strait during austral summer 1995/96. *Deep-Sea Research II.* **49**, 585–602 (2002)
9. Ketchum, B.H.: The exchanges of fresh and salt waters in tidal estuaries. *L. Mar. Res.* **10**, 18–37 (1951)
10. Lee, J. H., Qian, A.: Three-dimensional modeling of hydrodynamic and flushing in deep bay. International Conference on Estuaries and Coasts November 9-11, Hangzhou (2003)
11. Lenn, Y. D., Chereskin, T. K., Glatts, R. C.: Seasonal to tidal variability in currents, stratification and acoustic backscatter in a Antarctic ecosystem at Deception Island. *Deep-Sea Research II.* **50**, 1665–1683 (2003)
12. Lopez, O., Garcia, M. A., Sanchez-Arcilla, A. S.: Tidal and residual currents in the Bransfield Strait, Antarctica. *Ann Geophysicae.* **12**, 887–902 (1994)
13. Lopez, O., Garcia, M. A., Gomis, D., Rojas, P., Sospedra, J., Sanchez-Arcilla, A. S.: Hydrographic and hydrodynamica. characteristics of the eastern basin of the Bransfield Strait (Antartica). *Deep-Sea Research I.* **46**, 1755–1778 (1999)
14. Niiler, P. P., Amos, A., Hu, J. H.: Water masses and 200m relative geostrophic circulation in the western Bransfield Strait region. *Deep-See Research A.* **38**, 943–959 (1991)
15. Officer, C.B., Kester, D.R.: On estimating the non-advective tidal exchanges and advective gravitational circulation exchanges in an estuary. *Estuarine, Coastal and shelf Science.* **32**, 99–103 (1991)
16. Padman, L., Fricker, H. A., Coleman, R., Howard, S., Erofeeva, L.: A new Tide Model for the Antarctic Ice and Seas. *Annals of Glaciology.* **3**, 1–14 (2002)
17. Smith, K. L. Jr., Baldwin, R. J., Kaufmann, R. S., Sturz, A.: Ecosystem studies at Deception Island, Antartica: an overview. *Deep-See Research II.* **50**, 1595– 1609 (2003)
18. Vidal, J.: Caracterización dinámica de la marea y del sedimento en el canal de Sancti Petri. Ph. D. Thesis, Universidad de Cádiz, Cádiz (2002)
19. Vidal, J., Berrocoso, M., Fernandez, A.: Study of the tide in Deception and Livingston Islands (Antarctica). *Antarctic Science.* **in press**, (2010)
20. Wijeratne, E.M.S., Rydberg, L.: Modelling and observations of tidal wave propagation, circulation and residence times in Puttalam Lagoon, Sri Lanka. *Estuarine Coastal and Shelf Science.* **74**, 697–708 (2007)

Part II
Nonlinear and Complex Dynamics:
Applications in Biological Systems

Localized Activity States for Neuronal Field Equations of Feature Selectivity in a Stimulus Space with Toroidal Topology

Evan C. Haskell and Vehbi E. Paksoy

1 Introduction

The formation of spatially localized activity profiles or “bumps” have long been proposed as a mechanism for feature selectivity in models of cortex. Conditions for the existence and stability of persistent localized activity were first obtained by Amari [1]. Amari’s original results have been extended to networks that are not of lateral inhibition type [9, 10], spiking models [8], planar networks [10, 13, 14] and to sphere topology [4].

We seek to extend the mean field approach of Amari to study the formation of localized activity states in the topology of a torus. This is motivated by recent evidence that the topology of population activity is correlated to the topology of the stimulus space [12]. When the population is tuned to respond preferentially to two circular variables such as orientation and color hue, the underlying topology could be represented by a torus. Indeed, areas of primary visual cortex do demonstrate spatial maps that are clustered for orientation [6] and those clustered for color [3] implying the appropriateness of considering a torus model for areas with an ordered set of two-dimensional feature maps with circular variables. To what extent such a topology is applicable to other cortical areas is unclear. However, throughout cortex cells sharing similar receptive field properties tend to be arranged in a columnar structure [5]. It is hypothesized that such a clustering of cells by receptive field properties in the visual system facilitates the computations required to generate linking features used to associate related parts of an image [2, 7]. Here we explore a model of the computations performed for cells with receptive-fields tuned for two circular stimulus variables.

E.C. Haskell (✉)

Division of Math, Science, and Technology, Farquhar College of Arts and Sciences,
Nova Southeastern University, Ft. Lauderdale, FL 33314, USA
e-mail: haskell@nova.edu

2 Model

To model feature selectivity with two circular variables, we consider a continuum of neurons distributed on a Torus. The mean membrane potential of neurons located at a position (θ, ϕ) in the stimulus space with $\theta \in [-\pi, \pi)$ and $\phi \in [-\pi, \pi)$ is designated by $u(\theta, \phi, t)$. Following Amari, the field equation for the evolution of $u(\theta, \phi, t)$ for neurons receiving a homogeneous external input h , and an inhomogeneous external input $s(\theta, \phi, t)$ is taken to be of the form:

$$\tau \frac{\partial u(\theta, \phi, t)}{\partial t} = -u(\theta, \phi, t) + \int_T w(\theta, \phi | \theta', \phi') f[u(\theta', \phi', t)] D(\theta', \phi') + h + s(\theta, \phi, t) \quad (1)$$

where T designates the torus, $w(\theta, \phi | \theta', \phi')$ is the mean efficacy of synapses from neurons at (θ', ϕ') to (θ, ϕ) , and $f[u]$ is the mean activity level of the neurons. The integration measure, $D(\theta, \phi) = \frac{(a+b\cos(\phi))}{4\pi^2 a} d\theta d\phi$, is defined for the torus given in parametric form as:

$$\begin{aligned} x &= (a + b\cos(\phi))\cos(\theta) \\ y &= (a + b\cos(\phi))\sin(\theta) \\ z &= b\sin(\phi) \end{aligned}$$

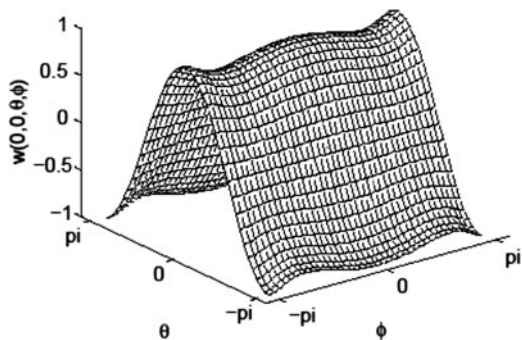
with meridian radius a and vertical radius b . For this analysis we will use an indicator function for $f[u]$, which is 1 if $u > 0$ and 0 otherwise.

Given the torus topology it is natural to construct a weight function, $w(\theta, \phi | \theta', \phi')$, that is invariant to the orientation preserving isometries of the torus. Although there are two such orientation preserving isometries corresponding to individual angular coordinate rotations and reflections, nontrivial dynamics should occur with a weighting function that is invariant to one of these isometries. We construct a weight function that is invariant to rotations about the z-axis by any angle. This is represented by the rotation group $O(2)$. Such a weighting function can be constructed from the angular separation, σ , between two points (θ, ϕ) and (θ', ϕ') on the torus. The angular separation between these points is given by the relation $\cos(\sigma) = \frac{(a+b\cos(\phi))(a+b\cos(\phi'))\cos(\theta-\theta') + b^2\sin(\phi)\sin(\phi')}{\sqrt{(a^2+2ab\cos(\phi)+b^2)}\sqrt{(a^2+2ab\cos(\phi')+b^2)}}$. Thus, we suggest a simple nontrivial form for a local weight distribution that will lead to nontrivial dynamics given by:

$$w(\theta, \phi | \theta', \phi') = W_0 + W_1 \frac{(a+b\cos(\phi))(a+b\cos(\phi'))\cos(\theta-\theta') + b^2\sin(\phi)\sin(\phi')}{\sqrt{(a^2+2ab\cos(\phi)+b^2)}\sqrt{(a^2+2ab\cos(\phi')+b^2)}}, \quad (2)$$

where W_0 and W_1 are parameters of the weighting function. The use of rotational symmetry to construct an $O(2)$ invariant weight function for the ring or an $O(3)$ invariant weight function for the sphere implies that the pattern of connections

Fig. 1 Surface plot of $w(0,0|\theta,\varphi)$ in the (θ,φ) plane for the ring torus with $a = 2$, $b = 1$, $W_0 = 0$ and $W_1 = 1$



within the network depends only upon the relative distance between the cells [4]. For a spherical topology, the angular separation coincides with the metric distance with respect to the intrinsic geometry of the sphere. The same cannot be said for the torus. Nevertheless, the metric preserving automorphisms and symmetries of the torus can be extended to that of ambient Euclidean space and symmetries preserve the angles. Therefore, one can consider angular separation for weight function.

It should be pointed out that for a horn torus where $a = b$, angular separation is not defined everywhere as the inclusion of the origin as a point on the torus allows for vectors of zero length. However, angular separation can be used to define a weight function for the ring torus where $a > b$ and the spindle torus where $a < b$. Using angular separation to define a weighting function for the sphere, we can construct a network of lateral inhibition type where neurons that are similar in feature preference are mutually excitatory and those that are dissimilar are mutually inhibitory. This is not the case for the torus. Nevertheless, the $O(2)$ invariant weight function for the torus utilized here provides similar results as that found with the $O(3)$ invariant weighting function for a sphere.

Figure 1 gives an example of the $O(2)$ invariant weight function $w(\theta, \varphi|\theta', \varphi')$ defined by (2) for the ring torus. As we can see this defines a network of lateral inhibition type in the θ variable only. The φ variable provides only a small modulation to the weighting level set by the θ variable. This would be indicative of a network that receives input representing two independent circular variables but is primarily selective to one of the variables. However, as we will see in the analysis the φ variable will have an impact on the radius of a localized activity state.

Figure 2 gives an example of the $O(2)$ invariant weight function $w(\theta, \varphi|\theta', \varphi')$ defined by (2) for the spindle torus. This weighting function is not properly of lateral inhibition type. While the strength of interaction does initially decay from the point of identical feature selectivity $(0, 0)$, at the points of self-intersection of the torus with itself the weighting function becomes positive and returns to a maximum value when both variables are maximally dissimilar, $|\theta - \theta'| = \pi$ and $|\varphi - \varphi'| = \pi$. The spindle torus intersects itself in the two circles $\varphi = \cos^{-1}(\frac{-a}{b})$ and $\varphi = \cos^{-1}(\frac{a}{b})$,

Fig. 2 Surface plot of $w(0,0|\theta,\phi)$ in the (θ,ϕ) plane for the spindle torus with $a = 2$, $b = 4$, $W_0 = 0$ and $W_1 = 1$

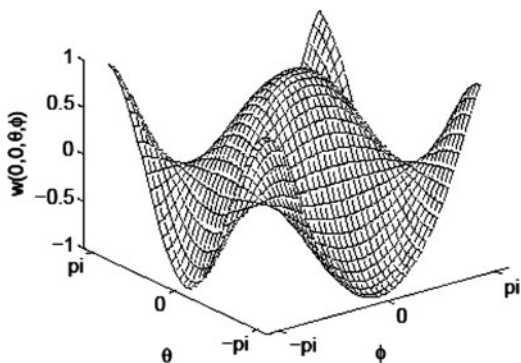
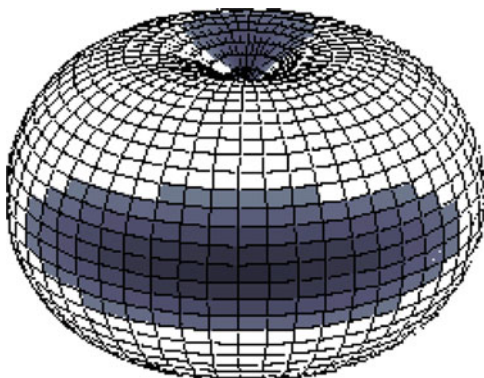


Fig. 3 Example of a mirror bump for the spindle torus with $a = 2$, $b = 3$, $W_0 = -1$, $W_1 = 4$, and $h = -0.05$



thus, the region where cells with dissimilar preferences for both features excite each other can be restricted by keeping the ratio of a to b close to 1. The spindle torus has the added complication that the integration measure defined above vanishes on these circles. However, by having these regions where cells with dissimilar preferences in each preference excite each other allows for the potential formation of two mirror bumps to the primary localized activity state generated by a transient initial stimulus $s(\theta, \phi, 0)$. An example of such a localized activity state and a corresponding mirror bump is shown in Fig. 3.

It is unclear if there are cortical areas with cells that are mutually excitatory when their feature preferences are both similar and very dissimilar while being mutually inhibitory when one feature preference is similar and the other dissimilar. Also, although multiple localized activity states can arise through long-range interactions between cells with similar feature preferences in different columns, it is unclear whether such phenomena would occur in the scenario described here. As such, we will focus our analysis for conditions on the existence of localized activity states to the ring torus.

3 Equilibrium Solutions for Ring Torus

We follow the approach of Amari [1] to classify and derive conditions for equilibrium solutions of the neural field equation with homogeneous inputs (i.e. $s(\theta, \varphi, t) = 0$) since this is applicable to arbitrary choices of $O(2)$ invariant weight distributions. equilibria solutions, $U(\theta, \varphi)$, of (1) in the absence of an inhomogeneous input satisfy

$$U(\theta, \varphi) = \int_T w(\theta, \phi | \theta', \phi') f[U(\theta', \phi')] D(\theta', \phi') + h. \quad (3)$$

Letting $R[U] = \{(\theta, \varphi) | U(\theta, \varphi) > 0\}$ denote the region over which the field is excited, (3) can be rewritten as

$$U(\theta, \varphi) = \int_{R[U]} w(\theta, \phi | \theta', \phi') D(\theta', \phi') + h. \quad (4)$$

Following the previous analysis for the sphere topology [4] we distinguish three types of equilibrium solution: \emptyset -, π -, and α -solutions. A \emptyset -solution satisfies $U(\theta, \varphi) \leq 0$ for $(\theta, \varphi) \in T$ so that the excited region is empty, $R[U] = \emptyset$. This is complementary to the π -solution where $R[U] = T$ in which the entire space is excited. α -solutions are localized activity states where a simply-connected proper subset of the torus is excited. Given the homogeneity and $O(2)$ invariance of the weighting function (2), we can choose the center of the α -solution to be at any arbitrary point (Θ, Φ) on the torus. The boundary of the excited region over which $U(\theta, \varphi) = 0$ is then given by $\cos(\alpha) = \frac{(a+b\cos(\phi))(a+b\cos(\Phi))\cos(\theta-\Theta)+b^2\sin(\phi)\sin(\Phi)}{\sqrt{(a^2+2ab\cos(\phi)+b^2)}\sqrt{(a^2+2ab\cos(\Phi)+b^2)}}$ with the angular radius $\alpha \in (0, \pi)$ determined self consistently from the equilibrium equation (4). Furthermore, without loss of generality we consider α -solutions centered at $\theta = 0$. Given symmetry considerations we have that for α -solutions, $R[U] = \{(\theta, \varphi) | -\alpha < \theta < \alpha, \varphi \in [-\pi, \pi], \alpha \in (0, \pi)\}$. Solutions of (4) then reduce to the form $U(\theta, \varphi) = G(\theta, \varphi, \alpha) + h$ where,

$$\begin{aligned} G(\theta, \phi, \alpha) &= \int_{-\pi-\alpha}^{\pi} \int_{-\alpha}^{\alpha} w(\theta, \phi | \theta', \phi') D(\theta', \phi') = \frac{w_0}{\pi} \alpha + \frac{w_1}{2\pi^2 a} \frac{a + b\cos(\phi)}{\sqrt{a^2 + 2ab\cos(\phi) + b^2}} \\ &\quad \times \left(\int_{-\pi}^{\pi} \frac{(a + b\cos(\phi'))^2}{\sqrt{a^2 + 2ab\cos(\phi') + b^2}} d\phi' \right) \cos(\theta) \sin(\alpha). \end{aligned} \quad (5)$$

It can be seen for any fixed value of φ that for $W_1 > 0$, this function decreases as we move away from the center of the bump at $\theta = 0$. Hence as in the ring and sphere models, a necessary condition for the existence of an α -solution in which $U > 0$ for $-\alpha < \theta < \alpha$ and $U < 0$ for $|\theta| > \alpha$ is $W_1 > 0$. Further note, that for the ring torus the terms containing φ in (5) are strictly positive contributions to the

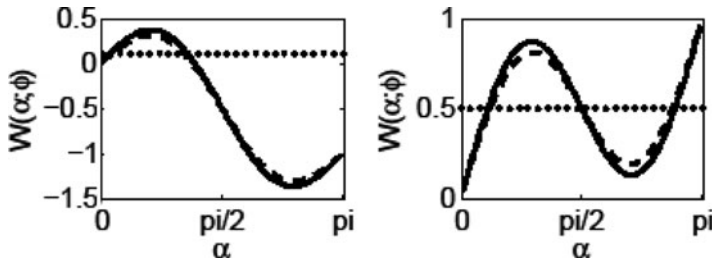


Fig. 4 Plot of $W(\alpha; \phi)$ as a function of α for $\phi = 0$ (solid line) and $\phi = \pi/2$ (dashed-line). In the left panel $W_0 = -1$ and $W_1 = 4$. In the right panel $W_0 = 1$ and $W_1 = 4$. The dotted line represents $-h$. The existence of α -solutions can be found by solving $W(\alpha; \phi) = -h$ graphically for various values of h . For example, the left panel has two solutions for $h = -0.1$ and the right panel has three solutions for $h = -0.5$. In both panels $a = 2$ and $b = 1$

equation indicating that the radius α of the solution is dependent upon ϕ . If we define $W(\alpha; \phi) = G(\alpha, \phi, \alpha)$, then finding the radius of the α -system becomes a problem of solving $W(\alpha; \phi) + h = 0$ for α as a function of ϕ . This is in contrast to the previously studied cases for the ring and sphere the width of the localized activity state did not have dependence upon a feature variable [4]. Examples of $W(\alpha, \phi)$ are shown in Fig. 4 which correspond to $W_0 < 0$ and $W_0 > 0$, respectively. Notice that ϕ has impacts on both the minimum and maximum values of $W(\alpha; \phi)$ and the location of these critical values. It is important to note that $W(0; \phi) = 0$, $W(\pi; \phi) = W_0$, and $W(\alpha; \phi) = W_0 - W(\pi - \alpha)$.

Theorem 1. *In the absence of an inhomogeneous input:*

- (a) *There exists a ϕ -solution if and only if $h < 0$*
- (b) *There exists a π -solution if and only if $W_0 > -h$*
- (c) *There exists an α -solution if and only if $W_1 > 0$ and $W(\alpha; \phi) + h = 0$ with $\pi > \alpha > 0$*

Proof. (a) If there is a ϕ -solution then $U(\alpha) = h$ requiring $h < 0$. If $h < 0$ then $U(\alpha) = h$ is a ϕ -solution.

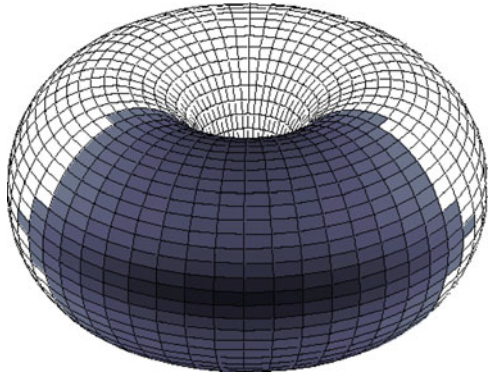
(b) For a π -solution, $u = W(\pi; \phi) + h = W_0 + h > 0$ which means that $W_0 > -h$. If $W_0 > -h$ then $u = W_0 + h$ is a π -solution.

(c) The conditions for the existence of an α -solution follows from (5) as discussed above. An α -solution is zero on the boundary of $R[U]$ if and only if $W(\alpha, \phi) = 0$ and is negative outside the domain of $R[U]$ if and only if $W_1 > 0$.

An example of an α -solution for the ring torus is shown in Fig. 5. Notice that the solution is a band around the torus centered at the preferred value of the θ variable that is widest at the preferred value of the ϕ variable.

We can further partition the parameter space (W_0, W_1) according to the sign of W_0 and the nature of the stationary points $W_{\max} = \max_{\alpha} W(\alpha; \phi)$ and $W_{\min} = \min_{\alpha} W(\alpha; \phi)$. In the ring and sphere topologies, five cases were identified

Fig. 5 Numerical example of an α -solution for the ring torus. The solution has been half-wave rectified for ease of presentation. Light and dark regions correspond to active and inactive regions respectively. Here $W_0 = -1$, $W_1 = 5$, $h = -0.1$, $a = 5$, and $b = 4$



in the parameter space that gave a classification of the equilibrium solutions for various values of h in the different portions of the parameter space [4]. We point out that the width of the localized activity state will depend on both the value of φ and the partitioning of the parameter space will depend on the torus radii a and b . The five cases are:

Case Ia: $W_1 > 0$, $W_0 < 0$ and $W_1 + W_0 < 0$: $W(\alpha; \varphi)$ is a monotonically decreasing negative function (no stationary points).

Case Ib: $W_1 > 0$, $W_0 < 0$ and $W_1 + W_0 > 0$: $W(\alpha; \varphi)$ has two stationary points with $W_{\max} > 0 > W_{\min}$ (see left panel of Fig. 4).

Case IIa: $W_0 > 0$, $0 < W_1 < \gamma_1 W_0$: $W(\alpha; \varphi)$ is a monotonically increasing positive function (no stationary points).

Case IIb: $W_0 > 0$, $\gamma_1 W_0 < W_1 < \gamma_2 W_0$: $W(\alpha; \varphi)$ has two stationary points with $W_0 > W_{\max} > W_{\min} > 0$ (see right panel of Fig. 4).

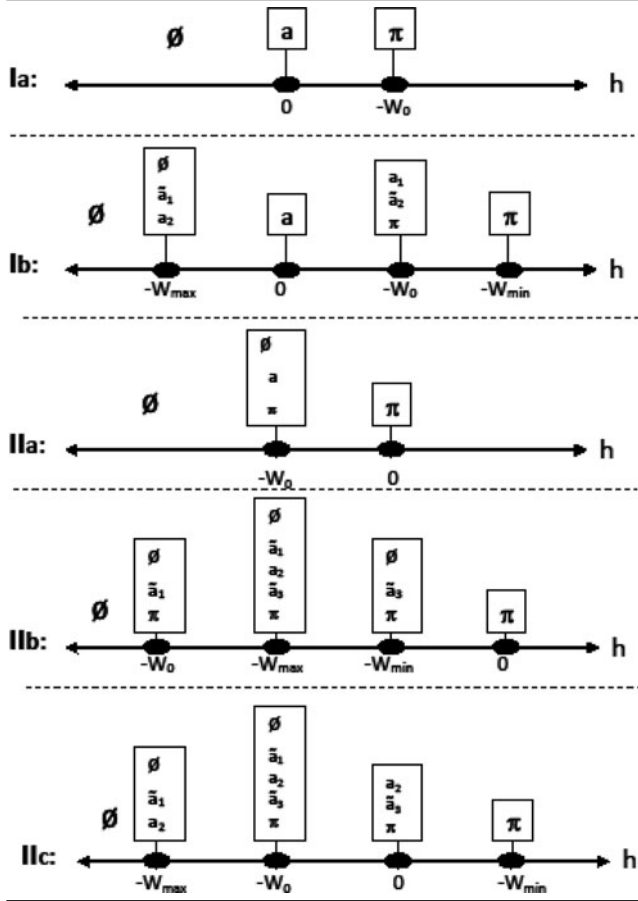
Case IIc: $W_0 > 0$, $W_1 > \gamma_2 W_0$: $W(\alpha; \varphi)$ has two stationary points with $W_{\max} > W_0 > 0 > W_{\min}$.

Case 1 follows from recognizing that $\frac{a+b\cos(\phi)}{\sqrt{a^2+2ab\cos(\phi)+b^2}} \leq 1$ and $\int_{-\pi}^{\pi} \frac{(a+b\cos(\phi))^2}{\sqrt{a^2+2ab\cos(\phi)+b^2}} d\phi \leq \int_{-\pi}^{\pi} (a+b\cos(\phi)) d\phi = 2\pi a$. The parameters γ_1 and γ_2 satisfying $0 < \gamma_1 < \gamma_2$ depend upon the torus radii a and b and φ . With this notation, the taxonomy of the sets of equilibrium solutions for the various values of h is identical to that of the sphere model presented by Haskell and Bressloff [4] and is reproduced here in Table 1.

Theorem 2. *The possible sets of equilibrium solutions for various values of h is given in Table 1.*

Proof. The existence of ϕ -, α -, and π -solutions follows from the qualitative properties of the function $W(\alpha; \varphi)$ in each of the five cases. Solving for $W(\alpha; \varphi) + h = 0$ will result in the various types of α -solution.

Table 1 Equilibria solutions for various values of h with ϕ -, π -, and a -denoting the existence of ϕ -, π -, and a -solutions, respectively. Also, $a_1 < a_2 < a_3$ and \tilde{a} indicating unstable equilibria (analysis not shown here). The bottom row of each table specifies the minimum value of h within a given interval with h increasing from left to right



What remains to be determined are the coefficients γ_1 and γ_2 . To find γ_1 we note that the condition $W_0 > 0$ and $W_1 = \gamma_1 W_0$ is the critical point when the two stationary points W_{\max} and W_{\min} coincide. By symmetry arguments this occurs when $\alpha = \pi/2$. We have that

$$W'(\alpha; \phi) = \frac{W_0}{\pi} + \frac{W_1}{2\pi^2 a} K(a, b; \phi) \cos(2\alpha), \quad (6)$$

where,

$$K(a, b; \phi) = \frac{a + b \cos(\phi)}{\sqrt{a^2 + 2ab \cos(\phi) + b^2}} \int_{-\pi}^{\pi} \frac{(a + b \cos(\phi'))^2}{\sqrt{a^2 + 2ab \cos(\phi') + b^2}} d\phi'.$$

Table 2 Values of the parameters γ_1 and γ_2 . The value of γ_2 given for the ring model is approximate

	Ring	Sphere	Ring torus
γ_1	1	2	$\frac{2\pi a}{K(a,b;\phi)}$
γ_2	4.6	8	$\frac{1}{\gamma_2} \frac{4\pi a}{K(a,b;\phi)} \alpha + \sin(2\alpha) = 0$

From (6) see that $W'(\pi/2) = 0$ when $W_I = \gamma_1 W_0$ where $\gamma_1 = \frac{2\pi a}{K(a,b;\phi)}$. The condition $W_0 > 0$ and $W_I = \gamma_2 W_0$ corresponds to the case where a pair of zeros of $W(\alpha; \phi)$ coincide such that $W_{\min} = 0$. To find the value of γ_2 requires solving the transcendental equation $\frac{W_0}{W_1} \frac{4\pi a}{K(a,b;\phi)} \alpha + \sin(2\alpha) = 0$ to find the ratio of W_0 to W_1 that provides a solution for a torus with given radii a and b .

4 Tristability

Tristable states in networks of lateral inhibition type in which a stable α -solution coexists with stable \emptyset - and π -solutions for noncompact topologies are not found in one-dimension [1] and were first reported for two-dimensional planar networks [14]. For compact topologies, tristable states have been identified for the one-dimensional ring and two-dimensional sphere [4]. However, in contrast to the planar and line networks, the existence of a localized activity state is not dependent upon a presence of global inhibition [4]. In the earlier analysis of the ring model, the value of γ_2 required to have a stable α -solution in the presence of a predominantly excitatory network was misidentified. The correct value comes from finding the value of γ_2 where two roots of the transcendental equation $W(a) = \frac{W_0}{\pi}(a + \gamma_2 \sin(2a))$ coalesce so that the minimum value is zero. Numerically this is found to occur approximately when $\gamma_2 = 4.6$.

Table 2 provides a summary of the parameter values γ_1 and γ_2 that partition the (W_0, W_1) parameter space by the different possible behaviors of the network. For the ring torus, these parameters will depend on the radii of the torus and are given in functional form. As with the ring model, finding the values of γ_2 for the ring torus involves finding the solution of a transcendental equation.

5 Discussion

We have analyzed the existence of localized activity states in neuronal network models of feature selectivity with a torus topology. Our worked is based upon a foundation presented by Amari for studying localized activity states for one-dimensional infinite neural fields of lateral inhibition type [1]. Amari showed that such fields are either monostable or bistable. When extended to a two-dimensional planar network, a regime of tristability appears in which a localized activity state coexists with both a quiescent and fully excited homogeneous state [14]. Extension

to compact topologies of the ring and sphere showed that these topologies supported the monostable, bistable, and tristable states as well; however, in compact topologies a global inhibition is no longer required for the presence of a localized activity state [4]. In the compact topology of the torus, we are able to repeat these results as we have done in this paper. Experimental evidence continues to grow between the relationship between the structure of stimulus space and neuronal activity [11, 12] increasing the need to further understand the relationship between network topology and the computations permitted by cortical networks.

The feature selectivity properties of neurons is generally not limited to two features and the interactions of those features may be quite complex. As part of a first step toward understanding the computational behavior of neuronal networks with higher dimensional feature spaces, we have examined here the common properties found in networks of lateral inhibition type for the compact network topologies of the ring, sphere, and ring torus.

References

1. Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol Cybernetics*, **27**, 77–87.
2. Barlowe, H. B. (1986). Why have multiple cortical areas? *Vision Res*, **26**, 81–90.
3. Conway, B. R., & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *PNAS*, **106**, 18034–18039.
4. Haskell, E. C., & Bressloff, P. C. (2003). On the Formation of Persistent States in Neuronal Network Models of Feature Selectivity. *J Integ Neurosci*, **2**, 103–123.
5. Horton, J. C., & Adams, D. L. (2005). The cortical column: A structure without a function. *Philos Trans R Soc Lond B Biol Sci*, **360**, 837–862.
6. Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional echitecture of monkey striate cortex. *J Physiol*, **195**, 215–243.
7. Hubel, D. H., & Wiesel, T. N. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *J Comp Neurol*, **158**, 267–293.
8. Laing, C. R., & Chow, C. C. (2001). Stationary bumps in networks of spiking neurons. *Neural Comp*, **13**, 1473–1494.
9. Laing, C. R., Troy, W. C., Gutkin, B., & Ermentrout, G. B. (2002). Multiple bumps in a neuronal model of working memory. *SIAM J on Appl Math*, **63**, 62–97.
10. Potthast, R., & Graben, P. B. (2010). Existence and Properties of solutions for neural field equations. *Math Meth Appl Sci*, **33**, 935–949.
11. Simoncelli, E.P., & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Ann Rev Neurosci*, **24**, 1193–1216.
12. Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., & Ringach, D. L. (2008). Topological analysis of population activity in visual cortex. *J Vis*, **8**, 1–18.
13. Taylor, J. G. (1999). Neural ‘bubble’ dynamics in two dimensions: foundations. *Biol Cybernetics*, **80**, 393–409.
14. Werner, H., & T. Richter. (2001). Circular stationary solutions in two-dimensional neural fields. *Biol Cybernetics*, **85**, 211–217.

Intrinsic Fractal Dynamics in the Respiratory System by Means of Pressure–Volume Loops

Clara M. Ionescu and J. Tenreiro Machado

1 Introduction

The study of fractional order systems received considerable attention recently [20], due to the fact that many physical systems are well characterized by fractional models [18]. With the success in the synthesis of real noninteger differentiators, the emergence of new electrical elements and the design of fractional controllers [15, 19], fractional calculus (FC) has been applied in a variety of dynamical processes [21]. The importance of fractional order mathematical models is that it can be used to make a more accurate description and to give a deeper insight into the physical processes underlying long range memory behaviors. In previous works, it was demonstrated that the respiratory system has fractal dynamics and constitutes a good test-bed for their study [9, 10].

The pseudo phase space (PPS) is used to analyze signals with nonlinear behavior. For the two-dimensional case, it is called pseudo phase plane (PPP) [14]. In this paper, we propose the use of pressure–volume loops to analyze data from lung function tests in healthy subjects and in patients with respiratory disorders. The novelty of the proposed methodology is to combine the information from pressure–volume loops with the corresponding fractal dimension. In this way, the fractal dynamics of the respiratory system can be assessed and further analyzed [21]. One of the most common features extracted from lung function tests is the air-pressure and the air-flow variation during forced breathing or during breathing at rest. The data used in this application has been acquired in lung function tests during breathing at rest, namely the forced oscillation technique [17]. We aim to analyze three sets of patient data: healthy children, children with asthma and children with cystic fibrosis.

C.M. Ionescu (✉)

Ghent University, Technologiepark 913, B9052 Gent-Zwijnaarde, Belgium

e-mail: ClaraMihaela.Ionescu@UGent.be

This paper is organized as follows. Section two introduces the materials and methods employed in the paper. The breathing recordings from patients in time domain is presented for one patient in each group, along with information upon the various sets of data, biometric details of the population sets and measurement protocol. A short description of the pressure–volume (PV) loop is given in the same section. Next, the processing of the information extracted from the PV loops and the calculus of the fractal dimension follows, along with a power law model structure. Section three presents the results of the PV loop analysis and the feature study and section four discusses the geometrical and clinical interpretation of the emerged results. Finally, a conclusion section summarizes the main outcome of this paper.

2 Materials and Methods

2.1 Patient Database

There are three sets of patients available for analysis, whose biometric details are given in Table 1. This data has been previously published in [12].

The measurements on the 33 *healthy children* were performed at the St. Vincentius Basis School in Zwijnaarde, Belgium; the biometric details are given in Table 1. The children had no history of pulmonary disease, and were selected using a specific questionnaire. The questionnaire verified the absence of dyspnea, chronic cough, wheeze in the chest, etc. The measurements performed on healthy children have been verified with predicted values [7], based on their height. All measured data was validated within the 95% confidence interval values.

Asthma denotes a pulmonary disease in which there is obstruction to the flow of air out of the lungs, but the obstruction is usually reversible and, between attacks of asthma, the flow of air through the airways is usually good [5]. Asthma is caused by chronic (ongoing, long-term) inflammation of the airways, making them highly sensitive to various triggers. Asthma can be controlled using specific medication (inhaled steroids). The data for this study was recorded at the University Hospital Antwerp-Belgium from 44 asthmatic children data sets whose corresponding biometric and spirometric values are given in Table 1.

Table 1 Biometric parameters of the investigated children subjects

	Healthy <i>n</i> = 33	Asthma <i>n</i> = 44	CF <i>n</i> = 38
Children			
Age (years)	9 ± 1	11 ± 4	14 ± 6
Height (m)	1.35 ± 0.05	1.40 ± 0.2	1.49 ± 0.15
Weight (kg)	32 ± 6	36 ± 15	40 ± 11

Values are presented as mean ± standard deviation
CF cystic fibrosis, *n* the number of tested volunteers

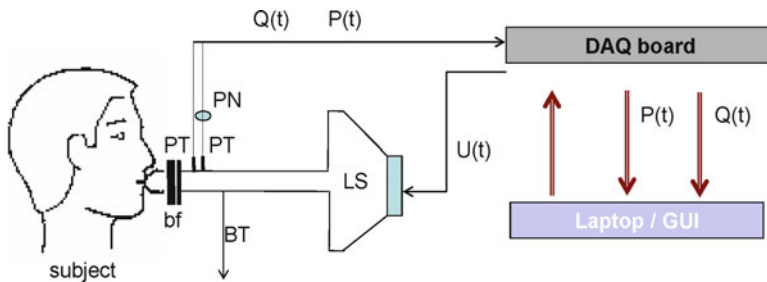


Fig. 1 A schematic overview of the device with patient performing the FOT lung function test. *LS* loudspeaker, *BT* bias-tube, *PN* pneumotachograph, *PT* pressure transducer, *bf* biological filter with mouthpiece, *Q* flow, *P* pressure, *U(t)* driving signal (test input)

Cystic fibrosis (CF) is one of the most common severe (genetic) diseases, characterized by the production of abnormal secretions, leading to mucus build-up and persistent inflammation in a variety of organs [4, 8]. Inflammation and infection also cause injury and structural changes to the lungs, originating a variety of symptoms and eventually respiratory failure. The data used in this study was recorded at the University Hospital Antwerp from 38 data sets from children diagnosed with cystic fibrosis whose biometric and spirometric values are given in Table 1.

2.2 Lung Function Testing: The Forced Oscillation Technique

The impedance was measured using the Forced Oscillation Technique (FOT) standard setup, commercially available, assessing respiratory mechanics from 4 to 48 Hz in steps of 2 Hz. The subject is connected to the typical setup from Fig. 1 via a mouthpiece, suitably designed to avoid flow leakage at the mouth and dental resistance artifact. The oscillation pressure in most recent FOT devices is generated by a loudspeaker (LS) connected to a chamber, namely the SPH-165KEP MONACOR, with a range from 3 to 1,000 Hz. The LS is driven by a power amplifier fed with the oscillating signal generated by a computer, namely a HP Pavilion dv1000 with a Pentium M processor, 1.5 GHz, 512 MB with 266 MHz SDRAM. The movement of the LS cone generates a pressure oscillation inside the chamber, which is applied to the patient's respiratory system by means of a flexible respiratory tube of 1 m length and 2 cm diameter, connecting the LS chamber and the bacterial filter (bf). A side opening (BT) of the main tubing allows the patient to decrease total dead space re-breathing (i.e., 40 ml). This bias tube exhibits high impedance at the excitation frequencies to avoid the loss of power from the LS pressure chamber [2]. During the measurements, the patient wears a nose clip and keeps the cheeks firmly supported to reduce the artifact of upper airway shunt.

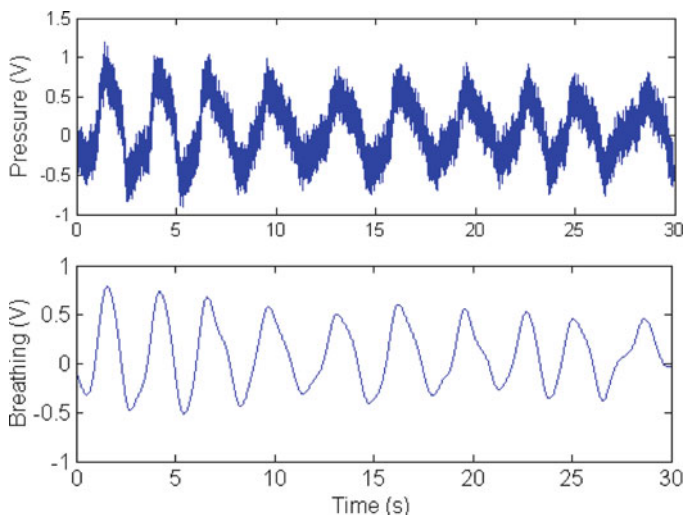


Fig. 2 Typical recorded airway pressure signal from the patient and filtered breathing signal

Pressure and flow are measured at the mouthpiece, respectively, by means of (a) a pressure transducer (PT) and (b) a pneumotachograph (PN) plus a differential pressure transducer (PT). The high precision pressure transducers are BSDX0050D4D, with a bipolar pressure range from 0 to 1 kPa, accuracy of 0.004 kPa and a common mode rejection ratio of 80 dB over the frequency range of interest. The working range is a peak-to-peak size between 0.1 and 0.3 kPa, in order to ensure optimality, patient comfort and stay within a narrow range in order to assume linearity [17]. The flow is measured using a Hans Rudolph pneumotachograph, covering a range from 0 to 400 l/min, (6.6 l/s), 4830B series, with a dead space volume between 0 and 6.66 ml.

Averaged measurements from 3 to 5 technically acceptable tests were taken into consideration for further signal processing (i.e., coherence function above 0.8). The signals were acquired using a PCMCIA DAQ card 4026E series, 12-bit from National Instruments, sampled every 1 ms. Typical time-records are depicted in Fig. 2. All patients were tested in the sitting position, with cheeks firmly supported and elbows resting on the table.

The output of the system is the airway pressure, where both the low-frequency component (i.e., breathing) and the excitation signal (i.e., multisine) are superimposed. Since the breathing frequency is one decade below the first excited harmonic, a low-pass filter can be applied and the breathing signal $b(t)$ can be extracted, as given in Fig. 2 (right). For the purpose of this investigation, we applied a Butterworth filter of order four and cut-off frequency at 1 Hz. From the flow measurement, volume has been obtained by integration.

2.3 The Pressure–Volume Curve

In clinical terms, the air-pressure and air-volume variations in one breathing cycle plotted against each other form a closed loop known as the PV loop [16]. The area inside this loop, and the slope of the axis of the minimal-to-maximal points in the PV loop are used to evaluate the respiratory mechanics of the patient. The interpretation of the PV loop is then made with respect to inspiratory and expiratory parameters, such as air-flow resistance and work of breathing.

The PV loops are defined by

$$\text{Area} = \int_0^T V(t) dP(t) = \int_0^T P(t) dV(t) \quad (1)$$

with $P(t)$ the pressure and $V(t)$ the volume at time instants t . The air-flow is related to the air-volume by $Q(t) = dV(t)/dt$, and using this in (1) we obtain that the area is the integral of the power:

$$\text{Area} = \int_0^T P(t) Q(t) dt, \quad (2)$$

which is by definition the work (energy) of breathing to perform the cycle over the period T . With disease, the work of breathing is increased and the ratio between peak pressure and peak volume is altered. However, for the purpose of this study, the area within the PV loop is of interest to us within a geometrical context.

2.4 Fractal Dimension and Power Law Model

The fractal dimension F_d is a quantity that gives an indication of how completely a spatial representation appears to fill the space. There are many specific methods to compute the fractal dimension. The most popular and simple methods are the Hausdorff dimension and box-counting dimension [1]. In this paper, the box-counting dimension method is used due to its simplicity of implementation and is defined as:

$$F_d = \frac{\ln[N(\varepsilon)] - \ln(C)}{\ln(1/\varepsilon)}, \quad (3)$$

where C is a constant related to the total area, $N(\varepsilon)$ represents the minimal number of covering cells (e.g., boxes) of size ε required to cover the PV graph. For each box size value ε , follows a corresponding number of boxes $N(\varepsilon)$. The slope for various values of box-sizes provides an estimate of the fractal dimension F_d :

$$N(\varepsilon) = C(1/\varepsilon)^{F_d} \quad (4)$$

Further on, given for each patient we have a PV plot, we shall obtain for each patient a value for C and a value for F_d . If one represents these values in a log–log plot, a power law model can be fitted to each group of patients, allowing deriving a model for each pathology:

$$C = A \times F_d^B \quad (5)$$

with A and B identified constants.

3 Results

For each measured set of signals, we have extracted the breathing signal and the pressure–volume curves. An illustrative example is given in Fig. 3.

Based on the PV plots, the box-counting method can be applied to obtain the fractal dimension F_d and the constant C for each set of data. The models are identified for each set of measurements from volunteer using the least-squares algorithm [13]. Consequently, the box-counting values $N(\varepsilon)$ and the box-sizes $1/\varepsilon$ values are obtained for each patient, in each data set. The result is depicted for each group in Figs. 4–6, respectively. From the F_d and C data obtained, one may fit a trend-line for each data set, following an identified model parameters from (5) valid for each respiratory disease. The values of the fitted trend-lines are given in Table 2, with the corresponding plots in Fig. 7.

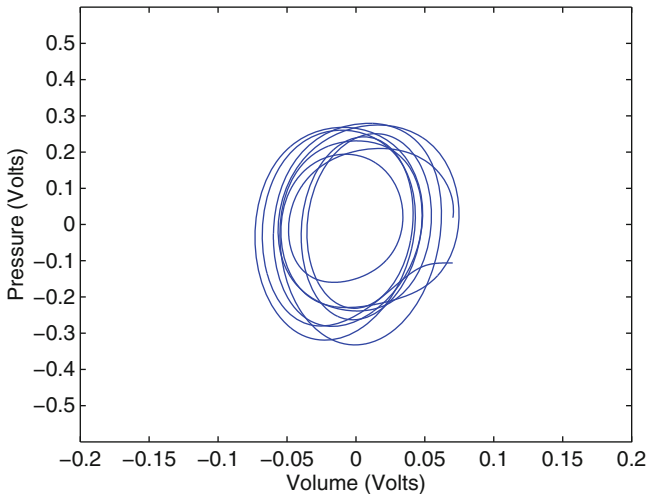


Fig. 3 Illustrative example of the PV loop for a healthy child

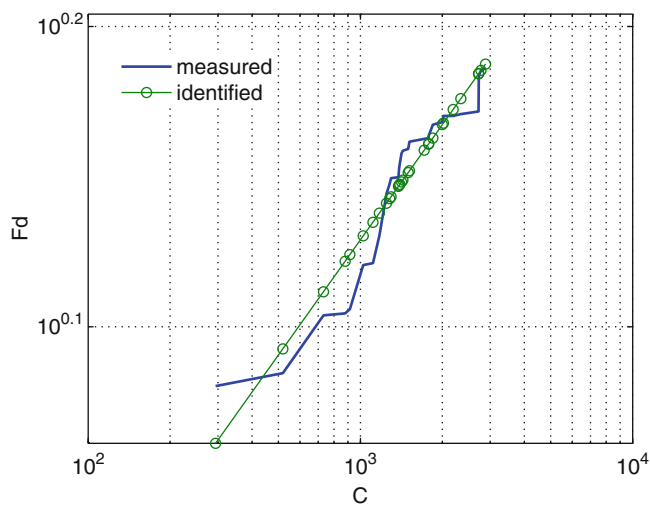


Fig. 4 Healthy children: the information extracted for each patient in terms of C and F_d , and the result of the identification

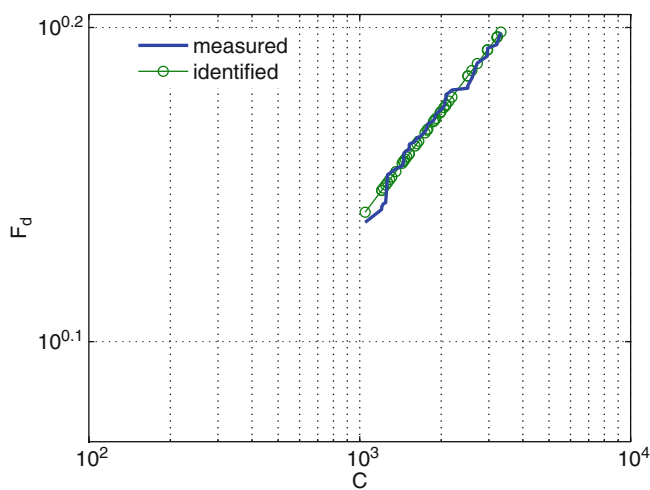


Fig. 5 Asthma children: the information extracted for each patient in terms of C and F_d , and the result of the identification

4 Discussion

In its most simple representation, the respiratory system can be represented as a series connection of a resistance R_e and a compliance C_e . It assumes patient's respiratory muscles inactive and the external equipment is driving the flow into the

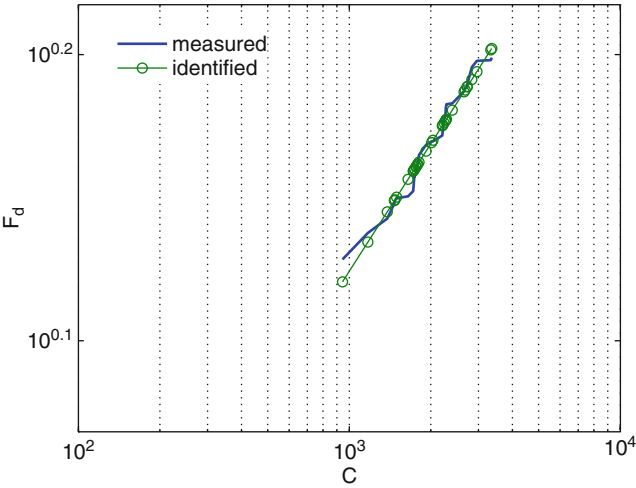


Fig. 6 Cystic fibrosis children: the information extracted for each patient in terms of C and F_d , and the result of the identification

Table 2 Fitted power-law models (5) for each group of data

Data	A	B
Healthy children	0.5579	0.1275
Asthma	0.6245	0.1145
CF	0.4786	0.1475
CF cystic fibrosis		

lungs [3]. The driving pressure $P(t)$ generates flow $Q(t)$ across the resistance and the volume $V(t)$ changes in the compliance. If $P_r(t)$ and $P_e(t)$ are the resistive and elastic pressure drops respectively, we have that:

$$R_r = \frac{P_r(t)}{Q(t)}; C_r = \frac{V(t)}{P_e(t)} \text{ and } P(t) = P_e(t) + P_r(t). \tag{6}$$

It results that:

$$P(t) = R_r \times Q(t) + \frac{V(t)}{C_r} \tag{7}$$

This represents the first-order equation in the motion-equation for a single compartment model of the respiratory system: a single balloon with compliance C_r on a pipeline with a resistance R_r . This system can be studied using the exponential decay of volume $V(t)$ as resulting from a step input V_0 :

$$V(t) = V_0 \times e^{-\frac{t}{\tau}}, \tag{8}$$

where t is time and τ is the time constant which characterizes the system, denoted by the product of $R_r C_r$ [9, 10, 17].

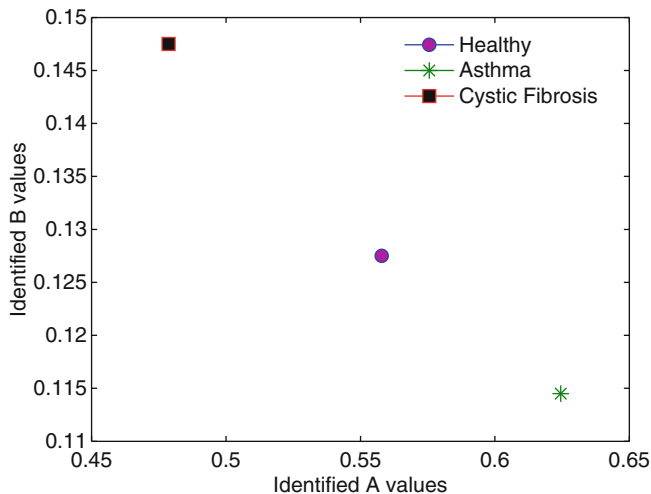


Fig. 7 The plot of the identified A and B values from model (5) for each set of data

In the representation of the PV plots, we have the breathing signal expressed as pressure and the volume. From (7) can be observed that there exists a relation between pressure and flow ($Q(t) = dV/dt$). In clinical terms, the pressure–volume loop during one breathing period is able to tell the clinician something about the dynamic compliance of the respiratory system and its work. The area enclosed by the PV loop is called the physiologic work of breathing, denoting the resistive work performed by the patient to overcome the resistance present in the airways.

During cycling loading, the stress that develops in the viscoelastic body (respiratory tissue) displays:

- A component in phase with strain, which is the elastic stress contributing to the storage modulus E_S (elastance).
- A component out of phase with strain, corresponding to the viscous dissipation and contributing to the loss of modulus E_D (damping).

In [6, 11] was shown that the respiratory system can be indeed modeled as a combination of series RC elements in a cascade arrangement of consecutive airways, by using their mechanical analogue representation, springs K_{rs} and dashpots B_{rs} . In this mechanical model, it follows that the pressure–volume relationship equivalent to the stress–strain relationship is given by:

$$P(t) = \frac{K_{rs}\ell_{rs}}{A_{rs}}V(t) + \frac{B_{rs}\ell_{rs}}{A_{rs}}\frac{dV(t)}{dt}, \quad (9)$$

with P the air-pressure, V the air-volume, ℓ_{rs} and A_{rs} the changes in length and area of airways during the breathing cycle, and K_{rs}, B_{rs} the constants of the spring and dashpot, respectively [6]. This relation suggests that the PV loop is indeed related to position and speed of air during breathing.

In summary, it can be concluded that the PV plots are able to provide information upon the intrinsic fractal dynamics of the breathing which changes with disease. Using the derived parameters extracted from the information delivered by the PV plots, i.e., models (4) and (5), we were able to obtain a classification between all three sets of measurements. It is interesting to notice that changes in the respiratory dynamics implies changes in the fractal dimension, as suggested by the values given in Table 2. This is indeed in agreement with the fact that airways and tissue structure are changing with disease, leading to changes in the fractal dimension of the respiratory tree [9, 10].

5 Conclusion

In this paper, a pressure–volume analysis has been performed on respiratory data from three groups of children: healthy, asthmatic, and cystic fibrosis. The results suggest the usefulness of using pressure–volume loops for mapping purposes, describing (dis)similarities within the breathing dynamics in groups of patients and between groups of patients.

Acknowledgments C. Ionescu gratefully acknowledges the children from primary school in Zwijnaarde who volunteered to perform lung function testing in our laboratory. We also acknowledge the technical assistance provided at University Hospital Antwerp, Belgium, to measure asthma and CF diagnosed children. J. Tenreiro Machado would like to acknowledge FCT, FEDER, POCTI, POSI, POCI, POSC, POTDC, and COMPETE for their support to R&D Projects and GECAD Unit.

References

1. Baker G.L., Gollub J.B., (1996) *Chaotic Dynamics: An Introduction*, 2nd ed. Cambridge, England: Cambridge University Press.
2. Birch M., MacLeod D., Levine M., (2001) An analogue instrument for the measurement of respiratory impedance using the forced oscillation technique, *Phys Meas* 22:323-339
3. Blom J., (2004) *Monitoring of respiration and circulation*, CRC Press.
4. Brennan S., Hall G., Horak F., Moeller A., Pitrez P., Franzmann A., Turner S., de Clerck N., Franklin P., Winfield K., Balding E., Stick S., Sly P., (2005) Correlation of forced oscillation technique in preschool children with cystic fibrosis with pulmonary inflammation, *Thorax* 60:159-163
5. Busse W., Lemanske R., (2001) Asthma, *New Engl J Med*, 344(5): 350-362
6. De Geeter N., Ionescu C., De Keyser R., (2009) A mechanical model of soft biological tissue - an application to lung parenchyma, In: *Proceedings of the 31st Annual Int Conf of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, USA, 2-6 September, ISBN 978-1-4244-3296-7, 2863-2866
7. Duiverman E., Clement J., Van de Woestijne K., Neijens H., van den Bergh A., Kerrebijn K., (1985) Forced oscillation technique: reference values for resistance and reactance over a frequency spectrum of 2-26 Hz in healthy children aged 2.3-12.5 years, *Clinical Resp Physiol* 21:171-178

8. Elizur A., Cannon C., Ferkol T., (2008) Airway inflammation in cystic fibrosis, *Chest* 133(2):489-495
9. Ionescu C., Segers P., De Keyser R., (2009) Mechanical properties of the respiratory system derived from morphologic insight, *IEEE Trans Biomed Eng* 56(4):949-959
10. Ionescu C., Muntean I., Machado J.T., De Keyser R., Abrudean M., (2010) A theoretical study on modelling the respiratory tract with ladder networks by means of intrinsic fractal geometry, *IEEE Trans Biomed Eng* 57(2):246-253
11. Ionescu C., Kosinsky W., De Keyser R., (2010) Viscoelasticity and fractal structure in a model of human lungs, *Archives of Mechanics* 62(1): 21-48
12. Ionescu C., (2009) Fractional-order models for the respiratory system, Doctoral Thesis, ISBN 978-90-8578-318-3
13. Ljung L., (1999) System identification: theory for the user, Prentice Hall
14. Moon F.C., (1987) Chaotic Vibration, New York: John Wiley,
15. Monje A., Chen Y., Vinagre B., Xue D., Feliu V., (2010) Fractional order systems and controls, Springer-Verlag
16. Northrop R., (2002) Non-invasive instrumentation and measurement in medical diagnosis, CRC Press
17. Oostveen E., Macleod D., Lorino H., Farré R., Hantos Z., Desager K., Marchal F., (2003) The forced oscillation technique in clinical practice: methodology, recommendations and future developments, *Eur Respir J* 22:1026-1041
18. Podlubny I., (2002) Geometrical and physical interpretation of fractional integration and fractional differentiation, *Journal of Fractional Calculus and Applied Analysis* 5(4):357-366
19. Tenreiro Machado J. A., (1997) Analysis and Design of Fractional-Order Digital Control Systems, *Journal Systems Analysis-Modelling-Simulation*, Gordon and Breach Science Publishers 27:107-122
20. Tenreiro Machado J. A., (2003) A Probabilistic Interpretation of the Fractional-Order Differentiation, *Journal of Fractional Calculus and Applied Analysis* 6(1):73-80
21. West B., (2010) Fractal physiology and the fractional calculus: a perspective, *Frontiers in Fractal Physiology* 1(12):1-17 (open source: www.frontiersin.org)

Part III
Nonlinear and Complex Dynamics:
Applications in Financial Systems

Forecasting Project Costs by Using Fuzzy Logic

M. Bouabaz, M. Belachia, M. Mordjaoui, and B. Boudjema

1 Introduction

The concept of fuzzy logic is derived from the theory of fuzzy sets. The theory of fuzzy sets was developed by Zadeh [1]. It is ranked among the methods of artificial intelligence. The method of fuzzy logic provides a way for coping with problems arising from unexpected situations. It is a means for solving hard problems, by determining a mathematical model that describes the system behavior, known as an unsupervised learning method.

The proposed model consists of two parts: First, optimization is the process to determine the number of clusters from data input–output, which respectively serves for subsequent use. The second part involves the extraction of fuzzy rules.

In this paper, we evaluate the use of the expectation maximization clustering algorithm in modeling and estimating projects costs.

1.1 The Fuzzy Clustering Algorithm

The Takagi–Sugeno [2] was the first model developed in 1985. This model can effectively represent complex nonlinear systems using fuzzy sets. The clustering technique is an essential method in data analysis and pattern recognition. Fuzzy clustering allows natural grouping of data in a large data set and provides a basis for constructing a rule-based fuzzy model. It is a partitioning method of data into subsets

M. Bouabaz (✉)

Civil Engineering Department, University of 20August, Skikda, Algeria

e-mail: mbouabaz@hotmail.fr

or groups based on similarities between the data [2]. The representation of fuzzy rules for the Takagi–Sugeno model takes the form:

$$\begin{aligned} R_i : & \text{ If } x_1 \text{ is } A_{i,1} \text{ and } \dots \text{ and } x_p \text{ is } A_{i,p} \\ & \text{ Then } y_i = a_{i0} + a_{i1}x_1 + \dots a_{ip} \end{aligned} \quad (1)$$

where, R_i is the rule number, x_j is the j -th input variable, A_{ij} is the fuzzy set of the j -th input variable in the i -th rule, y_i is the output of the i -th fuzzy rule.

1.2 The Fuzzy C-Means Algorithm

Fuzzy C-means algorithm (FCM) is a fuzzy clustering technique that is different from C-means that uses hard partitioning. The fuzzy c-means uses fuzzy partitioning in which a data point can belong to all clusters with different grades between 0 and 1. The FCM is an iterative algorithm that aims to find cluster centers that minimize the objective function.

$$J_{\text{FCM}}(Z; \Phi, V) = \sum_{i=1}^c \sum_{k=1}^N (\mu_{ik})^m D_{ikA}^2 \quad (2)$$

where $V = [v_1, v_2, \dots, v_c]$, v_i are the clusters centers to be determined. $\phi = \{\mu_{ik}\}$ is a fuzzy partition matrix; (μ_{ik}) is a membership degree between the i th cluster and k th data which is subject to conditions (3).

$$\mu_{ik} \in [0, 1], \quad 0 < \sum_{k=1}^N \mu_{ik} < N, \quad \sum_{i=1}^c \mu_{ik} = 1 \quad (3)$$

1.3 The Expectation Maximization Algorithm

The EM algorithm was proposed by Abonyi [3]. It is an extension of the algorithm of Gath and Geva [4], with a covariance matrix has nonzero diagonal elements, which creates an error in the projection of these elements. The methodology of the used algorithm is as follows:

The partition matrix is expressed by:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^c \left(D_{ikA}^2 / D_{jkA}^2 \right)^{1/(m-1)}} \quad 1 \leq i \leq c, \quad 1 \leq k \leq N \quad (4)$$

The prototypes (center) are:

$$V_i^x = \frac{\sum_{k=1}^N \mu_{i,k}^{(l-1)} X_k}{\sum_{k=1}^N \mu_{i,k}^{(l-1)}} \quad (5)$$

The standard deviation of the Gaussian membership functions is:

$$\sigma_{i,j}^2 = \frac{\sum_{k=1}^N \mu_{i,k}^{(l-1)} (x_{j,k} - v_{j,k})^2}{\sum_{k=1}^N \mu_{i,k}^{(l-1)}} \quad (6)$$

The local model parameters are extracted as follows:

$$\theta_i = (X_e^T \Phi_i X_e)^{-1} X_e^T \Phi_i y \quad (7)$$

where Φ is the weights matrix having the membership degrees defined by:

$$\Phi_i = \begin{bmatrix} \mu_{i,1} & 0 & \cdots & 0 \\ 0 & & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{i,N} \end{bmatrix} \quad X = \begin{bmatrix} X_1^T \\ \vdots \\ X_N^T \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad (8)$$

The extended matrix X_e is given by; $X_e = [X \ 1]$

And the prior probability is expressed by:

$$\alpha_i = \frac{1}{N} \sum_{k=1}^N \mu_{i,k} \quad (9)$$

The weights on rules are expressed as follows:

$$w_i = \prod_{j=1}^n \frac{\alpha_i}{\sqrt{2\pi\sigma_{i,j}^2}} \quad (10)$$

with a distance norm given by the following expression:

$$\frac{1}{D_{i,k}^2} = w_i \times \exp \left(-\frac{1}{2} \frac{(x_{j,k} - v_{i,j})^2}{\sigma_{i,j}^2} \right) \times \exp \left(\frac{(y_k - f_i(x_k, \theta_i))^T (y_k - f_i(x_k, \theta_i))}{2\sigma_i^2} \right) \quad (11)$$

where $f_i(x_k, \theta_i)$ is the model consequents.

The proposed algorithm is summarized as follows:

Let $Z = \{z_{k1}, z_{k2}, \dots, z_{kn}\}^T$

Select the number of clusters $c > 1$

Select the weithning exponent ($\mathbf{m} = 2$)

and the termination criterion ($\varepsilon > 0$)

Initialize the partition matrix such as (3)

Start

1: Compute the clusters centers using (5)

2: Compute fuzzy covariance matrix by (6)

3: Compute (7), (9), and (10)

4: Compute the distances using (11)

5: Update the partition matrix using (4)

If stopping criterion $|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}| \leq \varepsilon$ satisfied then stop

Otherwise $l \leftarrow l + 1$ and go to step 2

End

2 Validation

2.1 Indices

It is important to determine the number of clusters for use in simulation. For this, different indices for validation have been proposed by Bezdek [5] in data clustering. This can be done by the partition coefficient (PC) and the partition entropy (PE).

The PC measures the amount of overlapping between clusters.

$$PC(c) = \frac{1}{N} \sum_{i=1}^c \sum_{k=1}^N (\mu_{ik})^2 \quad (12)$$

The PE measures the fuzzyness of the cluster

$$PE(c) = -\frac{1}{N} \sum_{i=1}^c \sum_{k=1}^N \mu_{ik} \log(\mu_{ik}) \quad (13)$$

where $PC(c) \in [1/c, 1]$, $PE(c) \in [0, \log_a]$, with an increase of c , the values of PC and PE are decreased/increased, respectively. The above mentioned cluster validity indices are sensitive to fuzzy coefficient m . When $m \rightarrow 1$, the indices give the same value for all c . When $m \rightarrow \infty$, both PC and PE exhibit significant knee at $c = 2$. The number corresponding to a significant knee is selected as the optimal number of clusters.

2.2 The Performance Error

To evaluate the performance of the Fuzzy model, we use the following criteria proposed by Bezdek [6]. They are used for evaluating the output of the model.

The root mean squared error is expressed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^N (y_k - \hat{y}_k)^2} \quad (14)$$

where y is the output of the process and y' is the output of the model.

The variance accounted for is expressed by:

$$\text{VAF} = 100\% \left[1 - \frac{\text{var}(y - \hat{y})}{\text{var}(y)} \right] \quad (15)$$

Where n is the number of projects and i is the project number ($i = 1 \dots n$), and y' is the actual output

3 Application of Fuzzy Logic to Cost Estimation

The production of an accurate estimate for estimating projects costs is a challenging task for the estimator at the early stage of a project.

In an attempt to overcome the problem, a soft computing method has been used to solve the problem of uncertainties in estimating and constructing an accurate model for forecasting the final cost of a project.

3.1 The Model Development

The present model has been developed in three phases. First, it consists of the determination of the number of cluster. Second, the learning phase, and finally the testing phase.

3.2 The Data

A wide range of project contracts made on cost–significance work packages were used for modelling purposes [7].

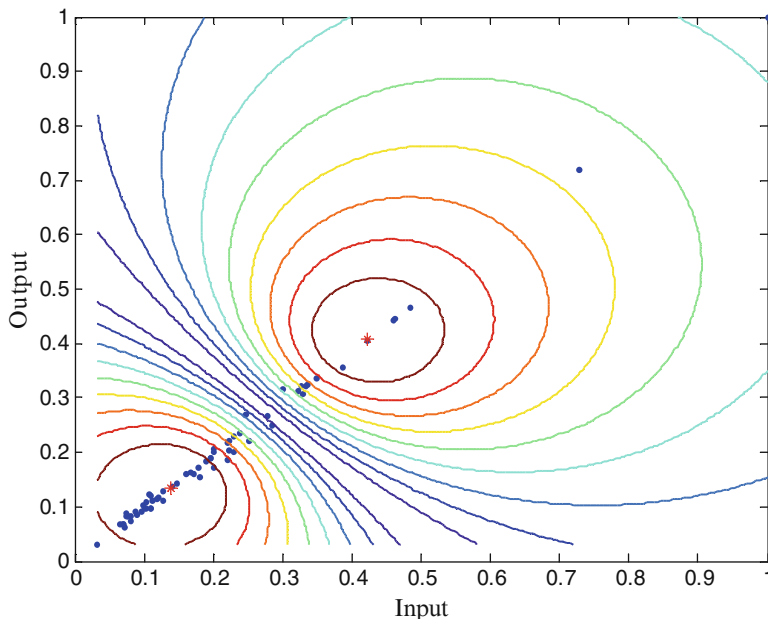


Fig. 1 The clustering map using expectation maximisation algorithm

3.3 The Clustering

The clustering consists of the selection of the number of clusters, which depends on the PC and the classification entropy.

The clustering map by the expectation maximisation algorithm is presented in Fig. 1.

We can deduce that the clustering number is equal to 2 clusters according to PC and CE from Fig. 2.

The values for PC and PE at optimization stage are given in Tables 1 and 2.

3.4 The Model Simulation

The proposed model was developed utilizing a set of projects. The developed model was generated from 68 data samples using Matlab toolbox [3,8] in a microcomputer. It has 1 input and 1 output. The training was stopped when the variance accounted-for (VAF) reached the maximum percentage value of 99.5304 in an elapsed time of 1.335000s. The termination tolerance of the clustering algorithm was 0.01. The training error (RMSE) is 0.0107. The simulation model at learning phase is shown graphically on Fig. 3.

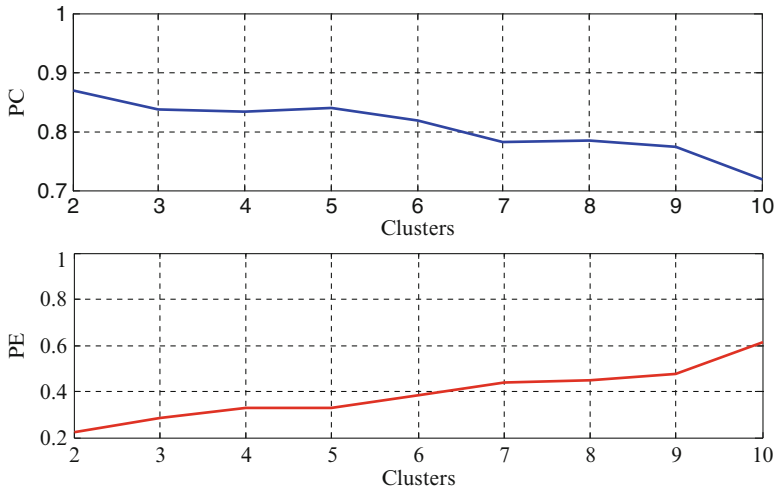


Fig. 2 Clusters validity indices for PC and PE

Table 1 Summarize the clustering results

Indices	Clusters								
	2	3	4	5	6	7	8	9	10
PC =	0.869	0.837	0.832	0.839	0.818	0.781	0.783	0.774	0.719
CE =	0.224	0.284	0.328	0.327	0.385	0.436	0.449	0.475	0.611

Table 2 The values for the clusters centres

Rules	$y(k-1)$	U
R1	1.61×10^{-1}	1.69×10^{-1}
R2	4.20×10^{-1}	4.28×10^{-1}

In order to illustrate sense of membership functions of the actual data versus simulated data obtained by the projection of the clusters by the expectation maximization algorithm. Fig. 4 shows fuzzy sets of the model with their rules presented in Table 3.

3.5 The Testing Model

A testing phase was investigated on the adopted rules in order to determine the accuracy of the model. Some flattering results are shown in Table 4.

Figure 5 shows the plot of the results given in Table 4 from the obtained rules at testing phase.

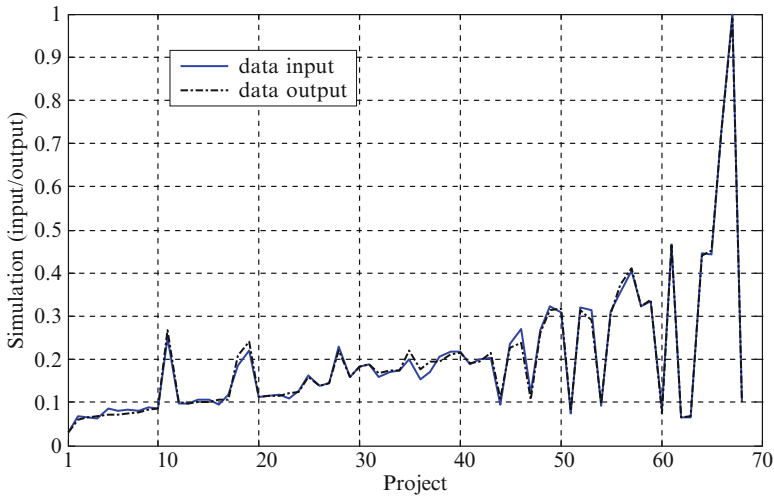


Fig. 3 Predicted model at learning phase

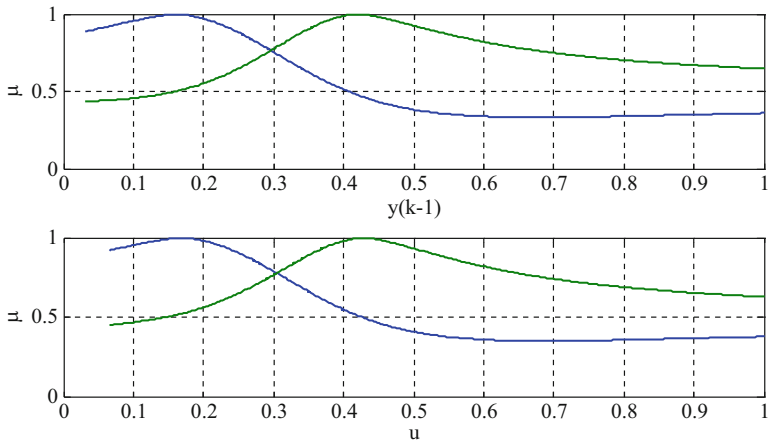


Fig. 4 The membership functions of the model

Table 3 Fuzzy rules obtained by expectation maximization algorithm

Rule 1
If $y(k-1)$ is A_{11} and u is A_{12} Then
$y(k) = 4.50 \times 10^{-2}y(k-1) + 9.23 \times 10^{-1}u + 1.04 \times 10^{-3}$
Rule 2
If $y(k-1)$ is A_{21} and u is A_{22} Then
$y(k) = 5.27 \times 10^{-3}y(k-1) + 1.04 \times 10^0u + 270 \times 10^{-2}$

Table 4 Results of testing rules

Project	Value of cswp's	Simulated fuzzy model	Actual bill value	Cpe
N ^o	(£)	(£)	(£)	(%)
1	21,402	26,662	26,753	0.000
2	47,158	57,931	57,775	−0.267
3	37,520	47,145	47,451	0.650
4	29,668	37,086	37,794	1.909
5	44,500	54,884	55,152	0.489
6	39,400	50,407	49,663	−1.475
7	57,898	72,373	71,702	−0.925
8	91,023	112,374	115,301	2.605
9	62,890	80,481	78,833	−2.046
10	110,264	137,241	140,700	2.520
Mean error				0.345
Standard deviation				1.617

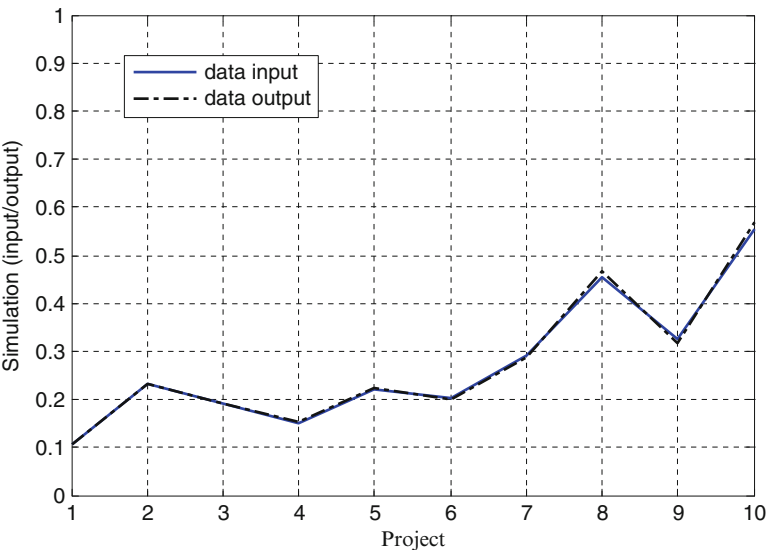


Fig. 5 Plot of simulated values on actual at testing phase

4 Conclusion

A modified fuzzy clustering algorithm based on the expectation maximization algorithm was used to construct a fuzzy model. Based on the results of the study, it can be concluded that the objectives of the research outlined has been achieved and the accuracy is significantly improved by means of fuzzy logic. It is also encouraged that fuzzy logic could be used in estimating and financing projects, were bills of

quantities made on cost-significance work packages for civil engineering projects contracts models are ready for use. As a conclusion fuzzy clustering approach based on the expectation–maximization algorithm seems to reveal promising results in modeling and forecasting projects costs in terms of accuracy.

References

1. Zadeh, L. A. Fuzzy sets. *Information and Control*. **8**, pp 338–353 (1965)
2. Takagi T. and M. Sugeno Fuzzy Identification of Systems and its Applications to Modeling and Control, *IEEE Transactions on Systems, Man, and Cybernetics* **15**(1), pp 116–132 (1985)
3. Abonyi J (2003). *Fuzzy Modeling Identification for Control*. Birkhäuser, Boston (2003)
4. Gath I and Geva, A.B (1989). Unsupervised Optimal Fuzzy Clustering. *IEEE Trans Pattern and Machine Intell*, vol. **11** (7), pp 773–781 (1989)
5. Bezdeck J C.; Ehrlich, R.; Full, W. FCM: Fuzzy C-means Algorithm. *Computers and Geoscience*, 10(2–3);pp 191–203 (1984)
6. Bezdeck, J C. Cluster Validity With Fuzzy Sets, *Cybernetics* **3**(3), pp 58–73 (1975)
7. Bouabaz M. and Horner RMW. Modeling and predicting bridge repair and maintenance costs, *Proc. Int. Conf. on Bridge Management*, London pp 187–197 (1990)
8. Abonyi J Balasko B, and Balazs L *Fuzzy Clustering and Data Analysis Toolbox for Use With Matlab* (2005)

Why You Should Consider Nature-Inspired Optimization Methods in Financial Mathematics

A. Egemen Yilmaz and Gerhard-Wilhelm Weber

1 Introduction

Financial Mathematics is a flourishing area of modern science. The subject has developed rapidly into a substantial body of knowledge since the days of pioneering people of this discipline such as Black, Scholes, and Merton. As of today, numerous applications of financial mathematics have become vital for the financial institutions, especially as regards to trading, asset management, and risk control of complicated financial positions.

The basic mathematics that underlies the subject is probability theory, with its strong connections to partial differential equations and numerical analysis. On the finance side, the main topics of importance are the pricing of derivatives, the evaluation of risk, and the management of portfolios. In fact, in today's world, many aspects of capital markets management are becoming more quantitatively and computationally sophisticated; however, it is still a valid argument to say that everything began with derivatives.

As complicated as the problems in consideration get, or as complicated as the approaches/models for handling of these problems get, conventional tools or methodologies become insufficient. In this paper, we will try to focus on the portfolio optimization problem, which is one of the main topics in financial mathematics. We will try to identify why and when conventional methods become insufficient, and metaheuristics (especially nature-inspired optimization methods) might constitute a remedy.

The organization of this chapter is as follows. After this introductory section, we will try to revisit the definition of the portfolio optimization problem with existing models in the literature. In Sect. 3, we will give brief descriptions of

A.E. Yilmaz (✉)

Ankara University Department of Electronics Engineering, 06100 Tandoğan, Ankara/Turkey
e-mail: aeYilmaz@eng.ankara.edu.tr

some popular nature-inspired optimization algorithms. Section 4 is nothing but a condensed literature review about the application of the nature-inspired methods for the solution of the portfolio optimization problem. In Sect. 5, we will try to give our concluding remarks.

2 Portfolio Optimization

Portfolio is nothing but the allocation of wealth (or resources in hand) among several assets. Portfolio optimization, which addresses the ideal assignment of resources to existing assets, have been one of the important research fields in modern risk management, or more generally financial management.

A fundamental answer to this problem was given by Markowitz [53, 54], who proposed the mean-variance model, which is now considered as the basis of modern portfolio theory. In Markowitz's approach, the problem was formulated as an optimization problem with two criteria:

- The profit (sometimes also referred to as reward or return) of a portfolio (measured by the mean) that should be maximized.
- The risk of the portfolio (measured by the variance of return) that should be minimized.

In the presence of two criteria, there is not a single optimal solution to the problem (i.e., a single optimal portfolio), but a set of optimal portfolios. Certainly, there is a trade-off between risk and return.

Since the mean-variance theory of Markowitz, research has been performed about extending or modifying the basic model in three directions [1]:

1. The simplification of the type and amount of input data.
2. The introduction of alternative measures of risk.
3. The incorporation of additional criteria and/or constraints.

In the following sections, we will try to summarize the basic models extending that of Markowitz while trying to identify the differences.

2.1 Mean–Variance Model

Markowitz's mean–variance model, in which the variance or the standard deviation is considered as a measure of risk, has been regarded as a quadratic programming problem. In spite of its popularity during the past, the mean–variance model is based on the assumptions that an investor is risk averse and that either (1) the distribution of the rate of return is multivariate normal or (2) the utility of the investor is a quadratic function of the rate of return [6].

However, neither (1) nor (2) holds in practice unfortunately. It is now widely recognized that the real world portfolios do not follow a multivariate normal distribution. Many researchers suggested that one cannot blindly depend on mean–variance model. That is why various risk measures such as semi-variance model, mean absolute deviation model, and variance with skewness model have been proposed.

2.2 *Semi–Variance Model*

Standard mean–variance model is based on the following assumptions:

- An investor is risk averse.
- The distribution of the rate of return is multivariate normal.

With this model, eventually the variance component of the Markowitz’s quadratic objective function can be replaced by other risk functions such as semi-variance. With an asymmetric return distribution, the mean–variance approach leads to an unsatisfactory prediction of portfolio behavior. Indeed, Markowitz himself suggested that a model based on semi-variance would be preferable.

2.3 *Mean Absolute Deviation Model*

Konno and Yamazaki [44] were the ones who first proposed a mean absolute deviation portfolio optimization model as an alternative to the Markowitz mean–variance portfolio selection model, with the advantage of the portfolio selection problem to be formulated and solved via linear programming.

It has been shown that this model yields similar results to the mean–variance model. Moreover, due to its simplicity, computational it outperforms to the mean–variance model [43,45].

2.4 *Variance with Skewness*

Samuelson [68] was the one who first noticed the importance of the third order moment in portfolio optimization. A portfolio return may not be a symmetric distribution. The distribution of individual asset returns tends to exhibit a higher probability of extreme values than is consistent with normality.

In order to capture the characteristics of the return distribution and to provide further decision-making information to investors, this model includes skewness into the mean–variance model. Although the existence of skewness in portfolios has been demonstrated many times, only a few studies to date have proposed incorporating

skewness into the portfolio optimization problem [6]. Konno and Yamamoto [46] showed that a mean–variance skewness portfolio optimization model can be solved exactly in a fast manner by using the integer programming approach.

3 Nature-Inspired Optimization Methods

Nature-inspired optimization methods fall into the class of metaheuristics. These are nothing but some methods influenced by the existing behaviors/phenomena for the solution of an optimization-like problem in nature. A very simple example of inspiration is the behavior of a colony or a swarm while searching for the best food source.

In computer science, the term metaheuristic is used to describe a computational method which optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. In other words, such methods are nothing but systematical trial-and-error approaches. Metaheuristics (sometimes also referred to as derivative free, direct search, black box, or indeed just heuristic methods) make few or no assumptions about the problem (such as modality or dimension) being optimized and can search very large spaces of candidate solutions. Moreover, most of these algorithms by definition are easily adaptable to parallel computing, which makes them applicable in very large scale problems.

However, it should be noted that metaheuristics do not guarantee that an optimal solution is ever found. On the other hand, for each algorithm, numerous studies (most of which are empirical) have been carried out in order to understand how the algorithm parameters should be adjusted for increasing the success probability.

Originally, metaheuristics were proposed for combinatorial optimization in which the optimal solution is sought over a discrete search space. An example is the traveling salesman problem, where the search space of candidate solutions grows exponentially as the size of the problem increases, which makes an exhaustive search for the optimal solution infeasible. Popular metaheuristics for combinatorial problems include simulated annealing [42], tabu search [32, 33], genetic algorithms [35], and ant colony optimization [21–24].

Later, metaheuristics for problems over real-valued search-spaces were also proposed. In such problems, the conventional approach was to derive the gradient of the function to be optimized, and then to employ gradient descent or a quasi-Newton method. Metaheuristics do not use the gradient or Hessian matrix; hence their advantage is that the function to be optimized need not be continuous or differentiable; moreover, it can also have constraints. Popular metaheuristic optimizers for real-valued search spaces include particle swarm optimization [41], differential evolution [75] and evolution strategies [66, 72].

All algorithms of this sort were initially proposed for single-objective problems. However, throughout the years, multiobjective extensions of these algorithms have been proposed. One of the early attempts was the extension of simulated annealing to multiobjective problems by Czyzak and Jaskiewicz [16]. Another one was that of

Hansen [34] for extension of tabu search. In a review article, Ehrgott and Gandibleux [26] listed a bibliography of multiobjective optimization approaches for combinatorial (not only considering metaheuristics, but also the conventional approaches). In another review article, Coello [12] identified the historical development of the research studies about extending these algorithms to multiobjective problems. This review was not limited to the combinatorial problems; but only the studies regarding population-based metaheuristics were considered.

In this paper, we will focus on the nature-inspired metaheuristics and their applications to the portfolio optimization problem. In the upcoming sections, we will briefly summarize the most popular and well-known ones.

3.1 Genetic Algorithm

Influenced from the “survival of the fittest” principle in the evolution theory, Holland [35] proposed genetic algorithms for the solution of combinatorial optimisation problems. The method simply relies on representation of the solution candidates by means of chromosomes, via which the relevant objective function is evaluated. The solution candidates constitute a population, which will be evolved throughout the generations (with the programmer’s perspective, the generations correspond to the populations in succeeding iterations). Performing these evaluations and considering the fitness of each solution candidate (i.e., the value of the objective function corresponding to that candidate), it is decided which candidates deserve to survive and to be transferred to the next generation. Certainly, as in the evolution process, diversity is added by means of some operators such as crossover (yielding the hybridization of high-quality solution candidates) and mutation. The main flow and the basic idea of the method are illustrated in Fig. 1.

Even though genetic algorithm was originally proposed by Holland for single-objective combinatorial problems, later it has been extended to real-valued optimization problems; even to multiobjective optimization problems (such as Schaffer [70], Corne et al. [13], Zitzler and Thiele [84], Zitzler et al. [85], Deb [17], Deb et al. [18]). A review by Coello [11] lists and identifies the genetic algorithm based multiobjective optimization techniques.

3.2 Genetic Programming

Inspired from the genetic algorithm, Koza [47] proposed the genetic programming approach in order to achieve a software program with a desired capability defined in terms of numerous input–output pairs. The approach is quite similar to the genetic algorithms; but this time, the genes inside the chromosomes are the building blocks (i.e., the functions, components, or modules) of the sought computer program [73].

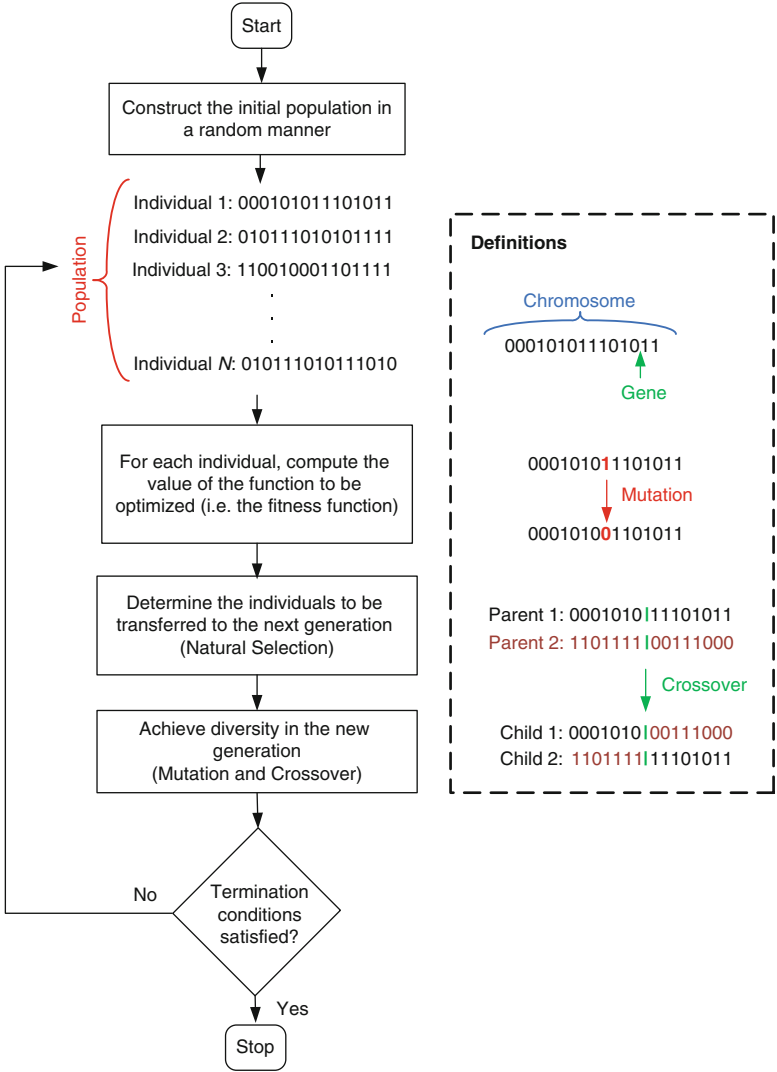


Fig. 1 The terminology and the main flow of the genetic algorithm

3.3 Differential Evolution

Extended from the genetic algorithm, the differential evolution is a recent meta-heuristic originally proposed by Storn and Price [75] for single-objective continuous problems. Again, the method relies on evolutionary operators' crossover and mutation, in addition to the concept of so-called differential weight. Despite its simplicity, the method has so far proven itself in numerous occasions with benchmark problems

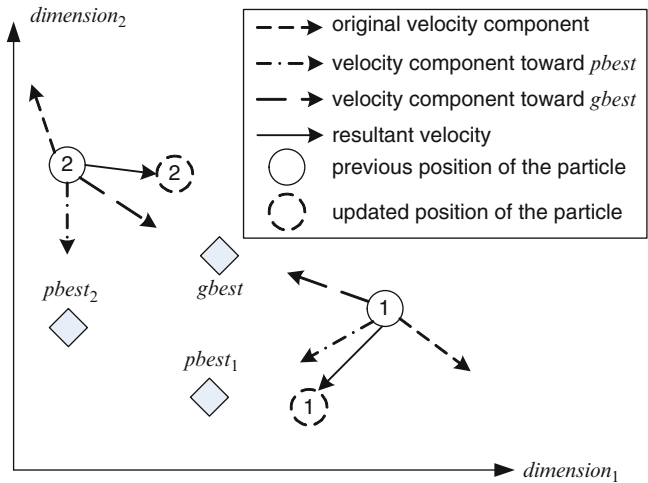


Fig. 2 The pictorial description of the particle tendencies in particle swarm optimization for a 2-D optimization problem

and outperformed many other optimization algorithms. Later, many differential evolution variants for other purposes (i.e., for the solution of the continuous and multiobjective problems, etc.) have been proposed. Unlike other nature-inspired optimization methods, it is possible to guarantee the trial-error procedure imposed by the algorithm to converge.

3.4 Particle Swarm Optimization

Particle swarm optimization is a method proposed by Kennedy and Eberhart [41] after getting inspired by the behaviors of the animal colonies/swarms. Similar to such swarms searching for the best place for nutrition in 3-dimensional space, this method relies on the motions of the swarm members (so-called particles) searching for the global best in an N -dimensional continuous space. This time, the position of each particle is a candidate solution of the problem in hand. As seen in Fig. 2, each member of the swarm has:

- A cognitive behavior (i.e., having tendency to return positions related with good memories); as well as
- A social behavior (i.e., having tendency to go where the majority of the swarm members are located); in addition to
- An exploration capability (i.e., the tendency for random search throughout the domain).

The balance among these three tendencies is the key to the success and the power of the method. So far, the method has been successfully applied to various multidimensional continuous and discontinuous problems. In fact, the results of a similar analysis were recently reported by Poli [63] in a review article (More detailed version of this review is also available on the web [62]). The power of the method is its simplicity allowing implementation in almost every platform and every programming language as well as its ease of parallelization.

Even though particle swarm optimization was originally proposed for single-objective continuous problems, later its discrete variants have also been published. Also, so far more than 30 versions of multiobjective particle swarm optimization extensions have been proposed, most of which have been reviewed by Reyes-Sierra and Coello [67].

3.5 Ant Colony Optimization

Ant Colony Optimization is another algorithm originally proposed by Dorigo [21] and later by Dorigo et al. [22–24] for the solution of combinatorial problems such as the traveling salesman or the shortest path.

Dorigo was inspired from the behaviors of ants while transporting food to their nests. The algorithm depends on the following principles: Initially, ants have random movements; but upon finding food they lay down pheromone trails returning home. Other ants have a tendency to follow these pheromones instead of keeping their random behavior. By this time, all pheromone trails start to evaporate and reduce their attractiveness. However, since pheromones over shorter paths are traced faster, and new pheromones are laid over the same path; new pheromone laying-out rate overcomes the evaporation rate. Due to this positive feedback mechanism, the popularity of shorter paths (i.e., pheromone density) increases in an accelerated manner as seen in Fig. 3. This is the key to the success of the ant colony optimization for the solution of relevant problems.

Similarly, continuous and multiobjective variants of ant colony optimization have later been published.

3.6 Other Nature-Inspired Optimization Methods

There are some other recent nature-inspired optimization algorithms with self-descriptive names such as:

- Bees algorithm [61]
- Invasive weed optimization [55]
- Artificial bee colony algorithm [37]
- Saplings growing-up algorithm [38]

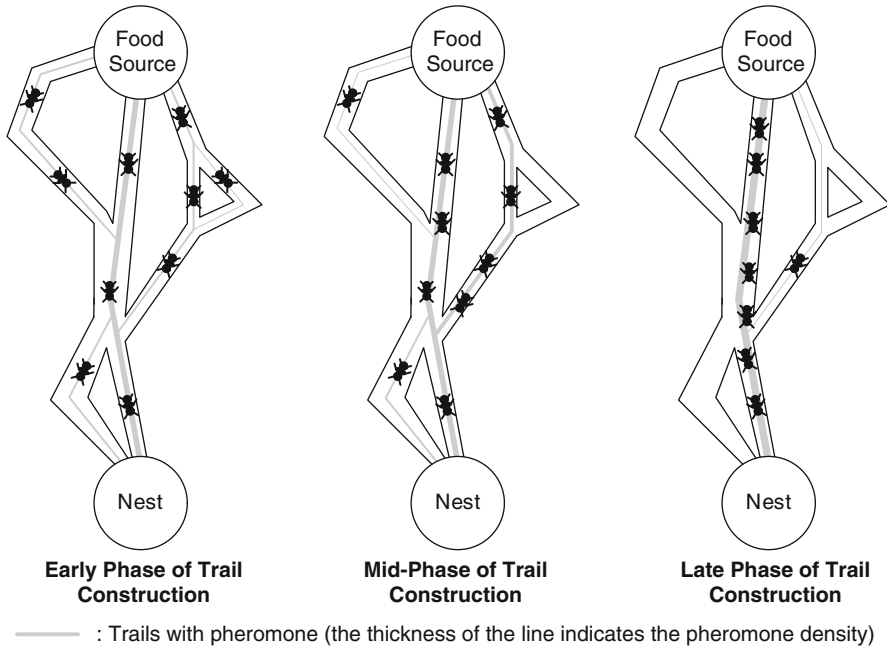


Fig. 3 Pictorial description of the basic idea underlying the ant colony optimization

- Monkey search [56]
- Central force optimization [29–31]
- Viral systems [14]
- Intelligent water drops algorithm [74]
- Gravitational search algorithm [65]
- Glowworm swarm optimization [49]
- Cuckoo search [82, 83]
- Firefly algorithm [80, 81]

which are still waiting for further research and promotion.

4 Application of Metaheuristics to the Portfolio Optimization Problem: A Literature Review

As stated before, portfolio optimization problem is actually a constrained multiobjective optimization problem. But in the early decades (between 1950s and early 1990s), due to lack of powerful methods and computational power, the problem used to be handled with oversimplifying assumptions. With the diffusion of metaheuristics to all disciplines, researchers started to apply them to problems of

their own branches. Eventually, portfolio optimization took its share. As of 2001, as pointed by Chen and Kuo [7], there have been about 400 publications regarding the application of metaheuristics to problems in economy and finance. Certainly, publications related to portfolio optimization constituted the majority.

Early attempts were applications of metaheuristics to the single-objective portfolio optimization problem with no constraints. Dueck and Winker [25] used a local-search based heuristic for the solution. Arnone et al. [4] were the ones who first applied genetic algorithm to the portfolio selection problem.

Later research can be considered in two main directions:

1. Incorporation of constraints in the problem model
2. Handling the problem as a multiobjective one

Regarding the studies about incorporation of the constraints in the problem: Jobst et al. [36] and Fieldsend et al. [28] discussed the computational aspects in the presence of the discrete asset choice constraints. Chang et al. [5] applied tabu search, simulated annealing and genetic algorithm to the portfolio optimization problem considering the cardinality constraint; afterwards Schaerf [69] and Kellerer et al. [39] applied local search algorithms; Streichert et al. [76] and Diosan [19] applied various evolutionary algorithms for the same purpose.

One of the very early multiobjective solution approaches in portfolio optimization problem was by Lin et al. [52], who applied genetic algorithm for this purpose [51]. Crama and Schyns [15] discussed how to apply simulated annealing in complex portfolio optimization problems. Ong et al. [59] applied a multiobjective evolutionary algorithm, whereas Armananzas and Lozano [2] applied multiobjective greedy-search, simulated annealing and ant colony optimization by considering the portfolio optimization problem as a triobjective one. Doerner et al. [20] applied ant colony optimization; Subbu et al. [77], Diosan [19], and Chiam et al. [10] applied various evolutionary algorithms; Kendall and Su [40] applied particle swarm optimization; Yang [79] applied the genetic algorithm. Application of differential evolution to the field is brand new. As of today, Ardia et al. [3] and Krink et al. [48] are the ones who have so far applied differential evolution to the portfolio optimization problem.

For deeper and broader surveys of the literature, interested readers can take a look at Schlottmann and Seese [71] or Tapia and Coello [78] the on application of multiobjective evolutionary algorithms in economics and finance in general.

Application of metaheuristics, or more specifically nature-inspired methods to similar problems such as index fund management, credit portfolio construction, etc. is also possible. Orito et al. [60], Kyong et al. [50], Oh et al. [57, 58] used genetic algorithms for the index fund management problem.

Another issue in portfolio management is the cost of transactions, which is usually neglected during the modelling of the problems. It is possible to incorporate this factor while applying metaheuristics. For example, Chen and Zhang [8] applied a particle swarm optimization variant to the portfolio optimization problem considering the transaction costs.

Meanwhile, genetic programming also found application in finance. Potvin et al. [64] applied genetic programming for generating trading rules in stock markets; more recently, Etemadi et al. [27] used it for bankruptcy prediction, meanwhile Chen et al. [9] used a time-adaptive version of the technique to portfolio optimization.

5 Conclusions

Metaheuristics, more specifically nature-inspired optimization algorithms, constitute powerful means for the solution of existing problems in economy and finance. The main factors promoting the usage of such algorithms can be summarized as follows:

- The algorithms make no assumptions (or require no a priori information) about the objective function.
- They do not require the objective function to be continuous or differentiable.
- They can handle complicated models with constraints.
- Almost all of them have variants for handling continuous and combinatorial problems.
- Almost all of them have extensions to multiobjective problems.
- Almost all of them support parallelization, which yields the solution of very large-scale problems.

Eventually, the literature is full of a plethora of publications about successful applications of such algorithms to problems, which could not have been considered and handled previously with conventional approaches. By observing the rate of increase in such publications, it can be easily forecasted that more and more applications will occur in the near future.

References

1. Anagnostopoulos, K.P. and G. Mamanis (2010). A portfolio optimization model with three objectives and discrete variables. *Computers & Operations Research*, **37**, 1285–1297.
2. Armananzas, R. and J.A. Lozano (2005). A multiobjective approach to the portfolio optimization problem. In: *Proceedings of IEEE Congress on Evolutionary Computation*, vol. 2, p. 1388–1395.
3. Ardia, D., K. Boudt, P. Carl, K.M. Mullen and B.G. Peterson (2010). Differential evolution (DEoptim) for non-convex portfolio optimization. *Munich Personal Repec Archive*, MPRA Paper No. 22135, [online], April 2010, Available at: http://mpra.ub.uni-muenchen.de/22135/1/MPRA_paper_22135.pdf.
4. Arnone, S., A. Loraschi and A. Tettamanzi (1993). A genetic approach to portfolio selection. *Neural Network World*, **6**, 597–604.
5. Chang, T.J., N. Meade, J.E. Beasley and Y.M. Sharaiha (2000). Heuristics for cardinality constrained portfolio optimization. *Computers & Operations Research*, **27**, 1271–1302.

6. Chang, T.-J., S.-C. Yang and K.-J. Chang (2009). Portfolio optimization problems in different risk measures using genetic algorithm. *Expert Systems with Applications*, **36**, 10529–10537.
7. Chen, S.-H. and T.-W. Kuo (2002). Evolutionary computation in economics and finance: A bibliography. In *Evolutionary Computation in Economics and Finance* (S.-H. Chen, (Ed.)), 419–455, Physica-Verlag, Heidelberg.
8. Chen, W. and W.-G. Zhang (2010). The admissible portfolio selection problem with transaction costs and an improved PSO algorithm. *Physica A*, **389**, 2070–2076.
9. Chen, Y., S. Mabub and K. Hirasawa (2010). A model of portfolio optimization using time adapting genetic network programming. *Computers & Operations Research*, **37**, 1697–1707.
10. Chiam, S.C., K.C. Tan and A.A.L. Mamum (2008). Evolutionary multi-objective portfolio optimization in practical context. *International Journal of Automation and Computing*, **5**(1), 67–80.
11. Coello, C.A.C. (2002). An updated survey of GA-based multiobjective optimization techniques. *ACM Computing Surveys*, **32**(2), 109–143.
12. Coello, C.A.C. (2006). Evolutionary multi-objective optimization: a historical view of the field. *IEEE Computational Intelligence Magazine*, **1**(1), 28–36.
13. Corne, D.W., J.D. Knowles and M.J. Oates (2000). The Pareto envelope-based selection algorithm for multiobjective optimization. In: *Proceedings of the Parallel Problem Solving from Nature VI Conference - Lecture Notes in Computer Science*, vol. 1917, p. 839–848.
14. Cortés, P., J.M. García, J. Muñizuri, and L. Onieva (2008). Viral systems: a new bio-inspired optimisation approach. *Computers & Operations Research*, **35**(9), 2840–2860.
15. Crama, Y. and M. Schyns (2003). Simulated annealing for complex portfolio selection problems. *European Journal of Operational Research*, **150**, 546–571.
16. Czyzak, P. and A. Jaskiewicz (1998). Pareto simulated annealing: a metaheuristic technique for multiple-objective combinatorial optimization. *Journal of Multi-Criteria Decision Analysis*, **7**, 34–47.
17. Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Chichester.
18. Deb, K., A. Pratap, S. Agarwal and T. Meyarivan (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, **6**(2), 182–197.
19. Diosan, L. (2005). A multi-objective evolutionary approach to the portfolio optimization problem. In: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation*, p. 183–188.
20. Doerner, K., W.J. Gutjahr, R.F. Hartl, C. Strauss and C. Stummer (2001). Ant colony optimization in multiobjective portfolio selection. In: *Proceedings of the 4th Metaheuristics International Conference*, p. 243–248.
21. Dorigo, M. (1992). *Optimisation, Learning, and Natural Algorithms*, Ph.D. Thesis, Politecnico di Milano, Italy.
22. Dorigo, M., V. Maniezzo, and A. Colomi. (1996). The ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man and Cybernetics*, **26**, 29–41.
23. Dorigo, M. and L. Gambardella (1997). Ant colony system: a cooperative learning approach to the travelling salesman problem. *IEEE Transactions on Evolutionary Computation*, **1**, 53–66.
24. Dorigo, M. and G. Di Caro (1999). The ant colony optimization meta-heuristic. In *New Ideas in Optimization* (D. Corne, M. Dorigo and F. Glover (Eds.)), p. 11–32, McGraw-Hill, London.
25. Dueck, G. and P. Winker (1992). New concepts and algorithms for portfolio choice. *Applied Stochastic Models and Data Analysis*, **8**, 159–178.
26. Ehrgott, M. and X. Gandibleux (2000). A survey and annotated bibliography of multiobjective combinatorial optimization. *OR Spektrum*, **22**, 425–460.
27. Etemadi, H., A.A.A. Rostamy and H.F. Dehkordi (2009). A genetic programming model for bankruptcy prediction: empirical evidence from Iran. *Expert Systems with Applications*, **36**(2), 3199–3207.
28. Fieldsend, J.E., J. Matatko and M. Peng (2004). Cardinality constrained portfolio optimisation. In: *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3177, p. 788–793.

29. Formato, R.A. (2007). Central force optimization: a new nature inspired computational framework for multidimensional search and optimization. In: *Proceedings of Nature Inspired Cooperative Strategies for Optimization (NICSO 2007) – Studies in Computational Intelligence*, vol. 129, p. 221–238.
30. Formato, R.A. (2009a). Central force optimization: a new deterministic gradient-like optimization metaheuristic. *OPSEARCH*, **46**(1), 25–51.
31. Formato, R.A. (2009b). Central force optimisation: a new gradient-like metaheuristic for multidimensional search and optimisation. *International Journal of Bio-Inspired Computation*, **1**(4), 217–238.
32. Glover, F. (1989). Tabu search - Part I. *ORSA Journal on Computing*, **1**(3), 190–206.
33. Glover, F. (1990). Tabu search - Part II. *ORSA Journal on Computing*, **2**(1), 4–32.
34. Hansen, M. (2000). Tabu search for multiobjective combinatorial optimization: TAMOCO. *Control and Cybernetics*, **29**, 799–818.
35. Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. University of Michigan Press, Michigan.
36. Jobst, N.J., M.D. Horniman, C.A. Lucas and G. Mitra (2001). Computational aspects of alternative portfolio selection models in the presence of discrete asset choice constraints. *Quantitative Finance*, **1**, 1–13.
37. Karaboga, D. and B. Basturk (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, *Journal of Global Optimization*, **39**(3), 459–471.
38. Karci, A. (2007). Theory of saplings growing up algorithm. In: *Proceedings of the International Conference on Adaptive and Natural Computing Algorithms*, vol. 4431, p. 450–460.
39. Kellerer, H. and D.G. Maringer (2003). Optimization of cardinality constrained portfolios with a hybrid local search algorithm. *OR Spectrum*, **25**(4), 481–495.
40. Kendall, G. and Y. Su (2005). A particle swarm optimisation approach in the construction of optimal risky portfolios. In: *Proceedings of Artificial Intelligence and Applications*, vol. 453, p. 140–145.
41. Kennedy, J. and R.C. Eberhart (1995). Particle swarm optimization. In: *Proceedings of IEEE International Conference on Neural Networks*, vol. 4, p. 1942–1948.
42. Kirkpatrick, S., C.D. Gelatt and M.P. Vecchi (1983). Optimization by simulated annealing. *Science - New Series*, **220**(4598), 671–680.
43. Konno, H. (2003). Portfolio optimization of small fund using mean-absolute deviation model. *International Journal of Theoretical and Applied Finance*, **6**(4), 403–418.
44. Konno, H. and H. Yamazaki (1991). Mean-absolute deviation portfolio in optimization model and its application to Tokyo stock market. *Management Science*, **37**, 519–531.
45. Konno, H., and T. Koshizuka (2005). Mean-absolute deviation model. *IEEE Transactions*, **37**, 893–900.
46. Konno, H. and R. Yamamoto (2005). A mean–variance–skewness model: Algorithm and applications. *International Journal of Theoretical and Applied Finance*, **8**(4), 409–423.
47. Koza, J.R. (1992). *Genetic Programming, on the Programming of Computers by means of Natural Selection*. MIT Press, Cambridge (MA).
48. Krink, T. and S. Paterlini (2010). Multiobjective optimization using differential evolution for real-world portfolio optimization. *Computational Management Science*, doi: 10.1007/s10287-009-0107-6 (in press).
49. Krishnanand, K.N. and D. Ghose 2009. Glowworm swarm optimisation: a new method for optimising multi-modal functions. *International Journal of Computational Intelligence Studies*, **1**(1), 93–119.
50. Kyong, J.O., Y.K. Tae and M. Sungky (2005). Using genetic algorithm to support portfolio optimization for index fund management. *Expert Systems with Applications*, **28**, 371–379.
51. Lin, C.C. and Y.T. Liu (2008). Genetic algorithms for portfolio selection problems with minimum transaction lots. *European Journal of Operational Research*, **185**(1), 393–404.

52. Lin, D., S. Wang and H. Yan (2001). A multiobjective genetic algorithm for portfolio selection problem. In: *Proceedings of the ICOTA 2001*, Hong Kong.
53. Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, **7**, 77–91.
54. Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*. John Wiley & Sons, New York.
55. Mehrabian, A. R. and C. Lucas (2006). A novel numerical optimization algorithm inspired from weed colonization. *Ecological Informatics*, **1**, 355–366.
56. Mucherino, A and O. Seref (2007). Monkey search: a novel meta-heuristic search for global optimization. In: *Proceedings of the Conference Data Mining, System Analysis and Optimization in Biomedicine (AIP Conference Proceedings 953)*, p. 162–173.
57. Oh, K.J., T.Y. Kim, S.H. Min (2005). Using genetic algorithm to support portfolio optimization for index fund management. *Expert Systems with Applications*, **28**, 371–379.
58. Oh, K.J., T.Y. Kim, S.H. Min and H.Y. Lee (2006). Portfolio algorithm based on portfolio beta using genetic algorithm. *Expert Systems with Applications*, **30**(3), 527–534.
59. Ong, C.S., J.J. Huang and G.H. Tzeng (2005). A novel hybrid model for portfolio selection. *Applied Mathematics and Computation*, **169**, 1195–1210.
60. Orito, Y., H. Yamamoto and G. Yamazaki (2003). Index fund selections with genetic algorithms and heuristic classifications. *Computers and Industrial Engineering*, **45**, 97–109.
61. Pham D.T., A. Ghanbarzadeh, E. Koç, S. Otri, S. Rahim and M. Zaidi (2006). The bees algorithm – a novel tool for complex optimisation problems, In: *Proceedings of IPROMS 2006 Conference*, p. 454–461.
62. Poli, R. (2007). An analysis of publications on particle swarm optimisation applications, *Technical Report- University of Essex Computer Science and Electrical Engineering CSM-469*, [online], May 2007 (rev. Nov. 2007). Available at: <http://www.essex.ac.uk/dces/research/publications/technicalreports/2007/tr-csm469-revised.pdf>.
63. Poli, R. (2008). Analysis of the publications on the applications of particle swarm optimisation. *Journal of Artificial Evolution and Applications*, Article ID: 685175, doi:10.1155/2008/685175.
64. Potvin, J., P. Soriano and M. Vallee (2004). Generating trading rules on the stock markets with genetic programming. *Computers & Operations Research*, **31**(7), 1033–1047.
65. Rashedi, E., H. Nezamabadi-pour, and S. Saryazdi (2009). GSA: A gravitational search algorithm. *Information Sciences*, **179**(13), 2232–2248.
66. Rechenberg, I. (1971). *Evolutionstrategie – Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*, Ph.D. Thesis, Germany.
67. Reyes-Sierra, M. and C.A.C. Coello (2006). Multi-objective particle swarm optimizers: a survey of the state-of-the-art. *International Journal of Computational Intelligence Research*, **2**(3), 287–308.
68. Samuelson, P. (1958). The fundamental approximation theorem of portfolio analysis in terms of means variances and higher moments. *Review of Economic Studies*, **25**, 65–86.
69. Schaerf, A (2002). Local search techniques for constrained portfolio selection problems. *Computational Economics*, **20**(3), 177–190.
70. Schaffer, J.D. (1985). Multiple objective optimization with vector evaluated genetic algorithm. In: *Proceedings of the First International Conference on Genetic Algorithms and their Applications*, p. 93–100.
71. Schlottmann, F. and D. Seese (2004). Modern heuristics for finance problems: A survey of selected methods and applications. In *Handbook of Computational and Numerical Methods in Finance* (S. Rachev (Ed.)), p. 331–360, Birkhauser, Berlin.
72. Schwefel, H.-P. (1974). *Numerische Optimierung von Computer-Modellen*, Ph.D. Thesis, Germany.
73. Sette, S. and L. Boullart (2001). Genetic programming: principles and applications. *Engineering Applications of Artificial Intelligence*, **14**(6), 727–736.
74. Shah-Hosseini, H. (2009). The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm. *International Journal of Bio-Inspired Computation*, **1**(1/2), 71–79.

75. Storn, R. and K. Price (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, **11**, 341–359.
76. Streichert, F., H. Ulmer and A. Zell (2003). Evolutionary algorithms and the cardinality constrained portfolio optimization problem. In: *Selected Papers of the International Conference on Operations Research*, p. 253–260.
77. Subbu, R., B. Bonissone, N. Eklund, S. Bollapragada and K. Chalermkraivuth (2005). Multiobjective financial portfolio design: a hybrid evolutionary approach. In: *Proceedings of IEEE Congress on Evolutionary Computation*, vol. 2, p. 1722–1729.
78. Tapia, M.G.C. and C.A.C. Coello (2007). Applications of multi-objective evolutionary algorithms in economics and finance. In: *Proceedings of IEEE Congress on Evolutionary Computation*, p. 532–539.
79. Yang, X. (2006). Improving portfolio efficiency: A genetic algorithm approach. *Computational Economics*, **28**, 1–14.
80. Yang X.-S. (2009). Firefly algorithms for multimodal optimization, In: *Stochastic Algorithms: Foundations and Applications (Lecture Notes in Computer Sciences 5792)*, p. 169–178.
81. Yang, X. S., (2010). Firefly algorithm, stochastic test functions and design optimisation. *International Journal of Bio-Inspired Computation*, **2**(2), 78–84.
82. Yang, X.-S. and S. Deb (2009). Cuckoo search via Lévy flights. In: *Proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)*, p. 210–214.
83. Yang, X.-S. and S. Deb (2010). Engineering optimisation by cuckoo search. *International Journal of Mathematical Modelling and Numerical Optimisation*, **1**(4), 330–343.
84. Zitzler, E. and L. Thiele (1999). Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, **3**(4), 257–271.
85. Zitzler, E., M. Laumanns and L. Thiele (2001). SPEA2: improving the strength of Pareto evolutionary algorithm. *TIK-103*, Department of Electrical Engineering, Swiss Federal Institute of Technology, Zurich, Switzerland.

Desirable Role in a Revenue-Maximizing Tariff Model with Uncertainty

Fernanda A. Ferreira and Flávio Ferreira

1 Introduction

It is well known that the outcome of a duopoly market varies when competition takes different forms such as Cournot or Stackelberg. Most of the economic literature assumes Cournot competition with simultaneous play as the natural order of moves in a quantity-setting game. However, recent advances in game theory argue that the assumed order of play should also be consistent with the players preferences over the time of actions. There are several comparative studies of these market structures. Notice that, in the typical models, compared to Cournot competition the Stackelberg leader can never be worse off: The leader can choose the Cournot quantity and then for the follower it is optimal to produce the same amount and we have Cournot outcome; so, leader either chooses Cournot quantity or if it chooses a different level of production it must be better off otherwise it would not do that. Here, we examine optimal trade in a market with demand uncertainty, in a duopoly in which a home firm competes with a foreign firm. Furthermore, we also assume that in that international market the home government imposes a tariff on the imported goods. In this case, we will see that the results can be different than the ones refereed above.

Tariff revenue may be an important source of government revenue for developing countries that do not have an efficient tax system. So, the government may use the maximum-revenue tariff. Brander and Spencer [1] have shown that a tariff has a profit-shifting effect in addition to its effect on tariff revenue. Larue and Gervais [8] studied the effect of maximum-revenue tariff in a Cournot duopoly. Ferreira and Ferreira [4] examined the maximum-revenue tariff under international Bertrand competition with differentiated products when rivals' production costs are unknown.

F.A. Ferreira (✉)

ESEIG - Polytechnic Institute of Porto and CMUP, Rua D. Sancho I, 981,
4480-876 Vila do Conde, Portugal
e-mail: fernandaamelia@eu.ipp.pt

Clarke and Collie [2] studied a similar question, when there is no uncertainty on the production costs. The propose of this paper is to study the maximum-revenue tariff under international quantity competition with demand uncertainty, with different possible timings of decisions.

We consider a two-country, two-good model where a domestic and a foreign good are produced by a home and a foreign monopolist, respectively. Since we assume that the two countries are perfectly symmetric, it is sufficient to describe only the domestic economy. We should mention that issues related to those of this paper have been studied by Ferreira et al. [5, 6], Ferreira and Pinto [7] and Spulber [9].

2 The Benchmark Model

There are two countries, home and foreign. Each country has one firm, firm F_1 (home firm) and firm F_2 (foreign firm) that produces homogeneous goods. Consider the home market, where the two firms compete in quantities (see [10]). We consider that the domestic government imposes an import tariff t per unit of imports from the foreign firm.

The inverse demand function is given by $p = A - q_i - q_j$, where $i, j \in \{1, 2\}$ with $i \neq j$ and q_i stands for quantity. In this section, we assume that the intercept A is commonly known since the begin of the game.

The model consists in the following two-stage game:

- In the first stage, the domestic government chooses the import tariff t per unit of imports from the foreign firm.
- In the second stage, both firms choose output levels.

Firms' profits, π_1 and π_2 , are, respectively, given by

$$\pi_1 = (A - q_1 - q_2)q_1 \quad \text{and} \quad \pi_2 = (A - q_1 - q_2 - t)q_2.$$

2.1 Simultaneous Decision

In this section, we suppose that, in the second stage of the game, both home and foreign firms play a Cournot-type game, i.e., each firm F_i independently chooses q_i . Let the superscript C denote the equilibrium outcome of the Cournot-type game.

We determine the subgame perfect Nash equilibrium by backwards induction. Suppose that the domestic government has chosen the import tariff t per unit of imports in the first stage.

Maximizing simultaneously both firms' profits, we get the following output levels:

$$q_1 = \frac{A+t}{3} \quad \text{and} \quad q_2 = \frac{A-2t}{3}.$$

Now, we can use the above results to derive the maximum-revenue tariff. The maximum-revenue tariff is the tariff rate that maximizes the tariff revenue collected by the government in the home country:

$$R = t \frac{A - 2t}{3}.$$

Theorem 1. *In the Cournot-type game, the maximum-revenue tariff is given by*

$$t^C = \frac{A}{4}.$$

So, we get the following result.

Theorem 2. *In the case of Cournot competition, the output levels at equilibrium are given by*

$$q_1^C = \frac{5A}{12} \quad \text{and} \quad q_2^C = \frac{A}{6}.$$

Thus, the aggregate quantity in the market is given by $Q^C = 7A/12$ and the price is given by $p^C = 5A/12$. The following results are also obtained straightforwardly.

Theorem 3. *In the case of Cournot competition, home firm's profit π_1^C and foreign firm's profit π_2^C are, respectively, given by*

$$\pi_1^C = \frac{25A^2}{144} \quad \text{and} \quad \pi_2^C = \frac{A^2}{36}.$$

Corollary 1. *In the case of Cournot competition, home firm profits more than foreign firm.*

2.2 Home Firm Is the Leader

In this section, we suppose that, in the second stage of the game, the home firm is the leader. Home firm F_1 chooses q_1 , and foreign firm F_2 chooses q_2 after observing q_1 . Let the superscript L denote the equilibrium outcome of the game where the home firm F_1 is the leader.

We determine the subgame perfect Nash equilibrium by backwards induction. Suppose that the domestic government has chosen the import tariff t per unit of imports in the first stage. Also, suppose that, the leader home firm F_1 produces q_1 . Then, maximizing foreign firm profit's $\pi_2 = (A - q_1 - q_2 - t)q_2$, we get

$$q_2 = \frac{A - t - q_1}{2}. \tag{1}$$

Now, maximizing home firm profit's $\pi_1 = (A - q_1 - q_2)q_1$, knowing the above quantity q_2 , we get

$$q_1 = \frac{A+t}{2}. \quad (2)$$

Putting (2) into (1), we get

$$q_2 = \frac{A-3t}{4}.$$

Now, we can use the above results to derive the maximum-revenue tariff. The maximum-revenue tariff is the tariff rate that maximizes the tariff revenue collected by the government in the home country. From the above results, the tariff revenue is

$$R = t \frac{A-3t}{4},$$

which leads us to the following result.

Theorem 4. *In the case of Stackelberg competition, with the home firm as the leader, the maximum-revenue tariff is given by*

$$t^L = \frac{A}{6}.$$

So, the output levels at equilibrium are as follows.

Theorem 5. *In the case of Stackelberg competition, with the home firm as the leader, home and foreign firms produce, respectively,*

$$q_1^L = \frac{7A}{12} \quad \text{and} \quad q_2^L = \frac{A}{8}.$$

Thus, the aggregate quantity in the market is given by $Q^L = 17A/24$ and the price is given by $p^L = 7A/24$. The following results are also obtained straightforwardly.

Theorem 6. *In the case of Stackelberg competition, with the home firm as the leader, home firm's profit π_1^L and foreign firm's profit π_2^L are, respectively, given by*

$$\pi_1^L = \frac{49A^2}{288} \quad \text{and} \quad \pi_2^L = \frac{A^2}{64}.$$

2.3 Home Firm Is the Follower

In this section, we suppose that, in the second stage of the game, the home firm is the follower. Foreign firm F_2 chooses q_2 , and home firm F_1 chooses q_1 after observing q_2 . Let the superscript F denote the equilibrium outcome of the game where the home firm F_1 is the follower.

We determine the subgame perfect Nash equilibrium by backwards induction. Suppose that the domestic government has chosen the import tariff t per unit of imports in the first stage. Also, suppose that, the leader foreign firm F_2 produces q_2 . Then, maximizing home firm profit's $\pi_1 = (A - q_1 - q_2)q_1$, we get

$$q_1 = \frac{A - q_2}{2}. \quad (3)$$

Now, maximizing foreign firm profit's $\pi_2 = (A - q_1 - q_2 - t)q_2$, knowing the above quantity q_1 , we get

$$q_2 = \frac{A - 2t}{2}. \quad (4)$$

Putting (4) into (3), we get

$$q_1 = \frac{A + 2t}{4}.$$

Now, we can use the above results to derive the maximum-revenue tariff. The maximum-revenue tariff is the tariff rate that maximizes the tariff revenue collected by the government in the home country. From the above results, the tariff revenue is

$$R = t \frac{A - 2t}{4},$$

which leads us to the following result.

Theorem 7. *In the case of Stackelberg competition, with the home firm as a follower, the maximum-revenue tariff is given by*

$$t^F = \frac{A}{4}.$$

So, the output levels at equilibrium are as follows.

Theorem 8. *In the case of Stackelberg competition, with the home firm as a follower, home and foreign firms produce, respectively,*

$$q_1^F = \frac{3A}{8} \quad \text{and} \quad q_2^F = \frac{A}{4}.$$

Thus, the aggregate quantity in the market is given by $Q^F = 5A/8$ and the price is given by $p^F = 3A/8$. The following result is also obtained straightforwardly.

Theorem 9. *In the case of Stackelberg competition, with the home firm as a follower, home firm's profit π_1^F and foreign firm's profit π_2^F are, respectively, given by*

$$\pi_1^F = \left(\frac{3A}{8}\right)^2 \quad \text{and} \quad \pi_2^F = \frac{A^2}{32}.$$

Corollary 2. *In the case of Stackelberg competition, with the home firm as a follower, the home firm profits more than the foreign leader firm.*

2.4 Comparisons

In this section, we are going to compare the results obtained in each model. First, we observe that, independently of the role, the sales of the home firm are increasing in the tariff, and the sales of the foreign firm are decreasing in the tariff. Furthermore, the total sales in the home country are decreasing in the tariff. Next corollary states that the domestic government imposes a lower tariff in the game where the home firm is the leader; and the tariffs are equal in the Cournot-type game and in the game where the home firm is the follower.

Corollary 3. *The tariffs in the different games are related as follows:*

$$t^L < t^F = t^C.$$

The total sales in the home market are higher in the game where the home firm is the leader; and they are lower in the Cournot-type game, as stated in the following corollary.

Corollary 4. *The total sales in the home market are related as follows:*

$$Q^C < Q^F < Q^L.$$

In the next corollary, we compare the profits of the firms obtained in each game. We note that the home firm profits more when it plays a Cournot-type game than in any other game. This result is in contrast to the well-known result that a Stackelberg leader firm profits more than a Cournot firm. Furthermore, the foreign firm prefers to be a Stackelberg leader firm, and the worse situation is to be a Stackelberg follower firm.

Corollary 5. *Home firm's profits are related as follows:*

$$\pi_1^F < \pi_1^L < \pi_1^C.$$

Foreign firm's profits are related as follows:

$$\pi_2^L < \pi_2^C < \pi_2^F.$$

3 The Model with Demand Uncertainty

In this section, we consider that the inverse demand function is given by

$$p = A - q_i - q_j,$$

where the intercept A is ex-ante unobservable, although its prior cumulative function $F(a)$ is commonly known to domestic government and both firms, with strictly positive finite mean $E(A)$ and variance $V(A)$. We assume that the exact realization of this intercept A becomes observable after the decision on the import tariff t fixed by the domestic government, and before both firms decide their output levels.

The model consists in the following two-stage game:

- In the first stage, the domestic government chooses the import tariff t per unit of imports from the foreign firm, without knowing the demand realization.
- In the second stage, both firms choose output levels, knowing the exact realization a of the demand,

Firms' profits, π_1 and π_2 , are given by

$$\pi_1 = (a - q_1 - q_2)q_1 \quad \text{and} \quad \pi_2 = (a - q_1 - q_2 - t)q_2.$$

3.1 Simultaneous Decision

In this section, we suppose that, in the second stage of the game, both home and foreign firms play a Cournot-type game, i.e., each firm F_i independently chooses q_i . Let the superscript C denote the equilibrium outcome of the Cournot-type game.

We determine the subgame perfect Nash equilibrium by backwards induction. Suppose that the domestic government has chosen the import tariff t per unit of imports in the first stage.

Maximizing simultaneously both firms profits, we get the following output levels:

$$q_1 = \frac{a+t}{3} \quad \text{and} \quad q_2 = \frac{a-2t}{3}.$$

Now, we can use the above results to derive the maximum-revenue tariff. The maximum-revenue tariff is the tariff rate that maximizes the tariff revenue collected by the government in the home country. Since the government does not know the exact demand, it will use the expected demand to compute that tariff. The tariff is t per unit of imports and the expected demand is $E(A)$; so, expected tariff revenue is

$$E(R) = t \frac{E(A) - 2t}{3}.$$

Theorem 10. *The maximum-revenue tariff is given by*

$$t^C = \frac{E(A)}{4}.$$

So, we get the following result.

Theorem 11. *In the case of Cournot competition, the output levels at equilibrium are given by*

$$q_1^C = \frac{4a + E(A)}{12} \quad \text{and} \quad q_2^C = \frac{2a - E(A)}{6}.$$

Thus, the aggregate quantity in the market is given by $Q^C = (8a - E(A))/12$ and the price is given by $p^C = (4a + E(A))/12$. The following result is also obtained straightforwardly.

Theorem 12. *Expected ex-ante profits of the two firms are given by*

$$E(\pi_1^C) = \left(\frac{5E(A)}{12} \right)^2 + \frac{V(A)}{9} \quad \text{and} \quad E(\pi_2^C) = \left(\frac{E(A)}{6} \right)^2 + \frac{V(A)}{9}.$$

3.2 Home Firm Is the Leader

In this section, we suppose that, in the second stage of the game, the home firm is the leader. Home firm F_1 chooses q_1 , and foreign firm F_2 chooses q_2 after observing q_1 . Let the superscript L denote the equilibrium outcome of the game where the home firm F_1 is the leader.

We determine the subgame perfect Nash equilibrium by backwards induction. Suppose that the domestic government has chosen the import tariff t per unit of imports in the first stage. Also, suppose that, the leader home firm F_1 produces q_1 . Then, maximizing foreign firm profit $\pi_2 = (a - q_1 - q_2 - t)q_2$, we get

$$q_2 = \frac{a - t - q_1}{2}. \quad (5)$$

Now, maximizing home firm profit $\pi_1 = (a - q_1 - q_2)q_1$, knowing the above quantity q_2 , we get

$$q_1 = \frac{a + t}{2}. \quad (6)$$

Putting (6) into (5), we get

$$q_2 = \frac{a - 3t}{4}.$$

Now, we can use the above results to derive the maximum-revenue tariff. The maximum-revenue tariff is the tariff rate that maximizes the expected tariff revenue

collected by the government in the home country. From the above results, the expected tariff revenue is

$$E(R) = t \frac{E(A) - 3t}{4},$$

which leads us to the following result.

Theorem 13. *In the case of Stackelberg competition, with the home firm as a leader, the maximum-revenue tariff is given by*

$$t^L = \frac{E(A)}{6}.$$

So, the output levels at equilibrium are as follows.

Theorem 14. *In the case of Stackelberg competition, with the home firm as a leader, home and foreign firms produce, respectively,*

$$q_1^L = \frac{6a + E(A)}{12} \quad \text{and} \quad q_2^L = \frac{2a - E(A)}{8}.$$

Thus, the aggregate quantity in the market is given by $Q^L = (18a - E(A))/24$ and the price is given by $p^L = (6a + E(A))/24$. The following result is also obtained straightforwardly.

Theorem 15. *In the case of Stackelberg competition, with the home firm as the leader, home and foreign firms' ex-ante expected profits are, respectively, given by*

$$E(\pi_1^L) = \frac{49(E(A))^2}{288} + \frac{36V(A)}{288} \quad \text{and} \quad E(\pi_2^L) = \frac{(E(A))^2}{64} + \frac{V(A)}{16}.$$

3.3 Home Firm Is the Follower

In this section, we suppose that, in the second stage of the game, the home firm is the follower. Foreign firm F_2 chooses q_2 , and home firm F_1 chooses q_1 after observing q_2 . Let the superscript F denote the equilibrium outcome of the game where the home firm F_1 is the follower.

We determine the subgame perfect Nash equilibrium by backwards induction. Suppose that the domestic government has chosen the import tariff t per unit of imports in the first stage. Also, suppose that, the leader foreign firm F_2 produces q_2 . Then, maximizing home firm profit

$$\pi_1 = (a - q_1 - q_2)q_1,$$

we get

$$q_1 = \frac{a - q_2}{2}. \tag{7}$$

Now, maximizing foreign firm profit $\pi_2 = (a - q_1 - q_2 - t)q_2$, knowing the above quantity q_1 , we get

$$q_2 = \frac{a - 2t}{2}. \quad (8)$$

Putting (8) into (7), we get

$$q_1 = \frac{a + 2t}{4}.$$

Now, we can use the above results to derive the maximum-revenue tariff. The maximum-revenue tariff is the tariff rate that maximizes the expected tariff revenue collected by the government in the home country. From the above results, the expected tariff revenue is

$$E(R) = t \frac{E(A) - 2t}{4},$$

which leads us to the following result.

Theorem 16. *In the case of Stackelberg competition, with the home firm as the follower, the maximum-revenue tariff is given by*

$$t^F = \frac{E(A)}{4}.$$

So, the output levels at equilibrium are as follows.

Theorem 17. *In the case of Stackelberg competition, with the home firm as the follower, home and foreign firms produce, respectively,*

$$q_1^F = \frac{2a + E(A)}{8} \quad \text{and} \quad q_2^F = \frac{2a - E(A)}{4}.$$

Thus, the aggregate quantity in the market is given by $Q^F = (6a - E(A))/8$ and the price is given by $p^F = (2a + E(A))/8$. The following result is also obtained straightforwardly.

Theorem 18. *In the case of Stackelberg competition, with the home firm as the follower, home and foreign firms' ex-ante expected profits are, respectively, given by*

$$E(\pi_1^F) = \left(\frac{3E(A)}{8} \right)^2 + \frac{V(A)}{16} \quad \text{and} \quad E(\pi_2^F) = \frac{(E(A))^2}{32} + \frac{V(A)}{8}.$$

3.4 Comparisons

In this section, we are going to compare the results obtained in each way of moving. First, we observe that, independently of the role, the sales of the home firm are

increasing in the tariff, and the sales of the foreign firm are decreasing in the tariff. Furthermore, the total sales in the home country are decreasing in the tariff. Next corollary states that the domestic government imposes a lower tariff in the game where the home firm is the leader; and the tariffs are equal in the Cournot-type game and in the game where the home firm is the follower.

Corollary 6. *The tariffs in the different games are related as follows:*

$$t^L < t^F = t^C.$$

The total sales in the home market are higher in the game where the home firm is the leader; but, they can be lower either in the Cournot-type game or in the Stackelberg-type game where the foreign firm is the leader, depending upon the value a of the realized demand, as stated in the corollary below. This result is in contrast to the benchmark model, in which the total sales are always lower in the Cournot-type game.

Corollary 7. *The total sales in the home market are related as follows:*

1. $Q^L > Q^C$ and $Q^L > Q^F$
2. $Q^C < Q^F$ if, and only if, $a > E(A)/2$

In the next corollary, we compare the ex-ante expected profits of the firms in each game. We note that, in the ex-ante analysis, the worse situation for the home firm is to play a Stackelberg-type game being a follower firm; and the best situation is to play a Cournot-type game, if the variance of the demand parameter is low, and to be leader in a Stackelberg-type game, if the variance is high. This result is in contrast to the benchmark model, in which the home firm always prefers to play a Cournot-type game (see [3]). Furthermore, in the ex-ante analysis, the foreign firm prefers to be a Stackelberg leader firm, and its worse situation is to be a Stackelberg follower firm.

Corollary 8. *Home firm's ex-ante expected profits are related as follows:*

1. $E(\pi_1^F) < E(\pi_1^L)$ and $E(\pi_1^F) < E(\pi_1^C)$
2. $E(\pi_1^C) > E(\pi_1^L)$ if, and only if, $V(A) < (E(A))^2/4$

Foreign firm's ex-ante expected profits are related as follows: $E(\pi_2^L) < E(\pi_2^C) < E(\pi_2^F)$.

4 Conclusions

In this paper, we studied an international quantity competition with demand uncertainty, where each firm is a Cournot competitor or a Stackelberg leader. We computed the maximum-revenue tariff, the quantities, the prices and the profits in each role of the model.

Acknowledgements We are grateful to Alberto A. Pinto for a number of very fruitful and useful discussions on this work and for his friendship and encouragement. We thank ESEIG – Instituto Politécnico do Porto, Centro de Matemática da Universidade do Porto and the Programs POCTI and POCI by FCT and Ministério da Ciência, Tecnologia e do Ensino Superior for their financial support.

References

1. Brander JA, Spencer BJ (1984) Tariff protection and imperfect competition. In: Kierzkowski H (ed) *Monopolistic Competition and International Trade*: 194–207, Oxford University Press, Oxford
2. Clarke R, Collie DR (2006) Optimum-welfare and maximum-revenue tariffs under Bertrand duopoly. *Scottish Journal of Political Economy* 53: 398–408
3. Ferreira FA, Ferreira F (2010) Simultaneous decisions or leadership in an international competition. In: Simos TE et al. (eds) *Numerical Analysis and Applied Mathematics, AIP Conference Proceedings* 1281: 804–807, American Institute of Physics, New York
4. Ferreira FA, Ferreira F (2009) Maximum-revenue tariff under Bertrand duopoly with unknown costs. *Communications in Nonlinear Science and Numerical Simulation* 14: 3498–3502
5. Ferreira FA, Ferreira F, Pinto AA (2007) Bayesian price leadership. In: Kenan Tas et al. (eds) *Mathematical Methods in Engineering*: 371–379, Springer, Dordrecht
6. Ferreira FA, Ferreira F, Pinto AA (2007) Unknown costs in a duopoly with differentiated products. In: Kenan Tas et al. (eds) *Mathematical Methods in Engineering*: 359–369, Springer, Dordrecht
7. Ferreira FA, Pinto AA (2011) Uncertainty on a Bertrand duopoly with product differentiation. In: Tenreiro Machado JA et al. (eds) *Nonlinear Science and Complexity*: 389–395 Springer, Dordrecht
8. B. Larue B, Gervais J-P (2002) Welfare-maximizing and revenue-maximizing tariffs with few domestic firms. *Canadian Journal of Economics* 35: 786–804
9. Spulber D (1995) Bertrand competition when rivals' costs are unknown. *Journal of Industrial Economics* 43: 1–11
10. Tirole J (1994) *The Theory of Industrial Organization*. MIT Press, Cambridge, Mass.

Can Term Structure of Interest Rate Predict Inflation and Real Economic Activity: Nonlinear Evidence from Turkey?

Tolga Omay

1 Introduction

Policymakers may use the information provided by financial indicators as a complementary to the information provided by monetary aggregates to make monetary policy decisions. Therefore, it is crucial for policymakers to know whether these financial indicators can predict future real economic activity and inflation (i.e., [1]). There are numerous studies, for instance, on whether the interest rate spread (i.e., the difference between the long-term interest rate and the short-term interest rate) can predict future economic activity [1–9]. However, most of the previous literature applied a linear approach to tackle the problem. More recently, Venetis et al. [10] has suggested that Smooth Transition Regression (STR) models are better suited to investigate spread-output relationships as they can contain regime switching-type behavior and time-varying parameters. On the other hand, there are numerous studies on whether the interest rate spread can predict future inflation [11–15]. More recently, Telatar et al. [16] has also suggested that nonlinear models (Markov Switching) are useful and better suited to investigate spread-inflation relationship as they can accommodate to regime switching-type behavior and time-varying parameters.

In addition to the bivariate single equation investigations of the predictability of term structure, there are also some “equation system” studies in order to identify the predictability relationship. In the equation-system approach class, we see that the most influential study is Estrella’s [1] 1997 paper. In this paper, he has derived reduced form relationships of output growth and spread, inflation and spread and finally, interest rate and spread from a simple structural model of the

T. Omay (✉)

Department of Economics, Çankaya University, Ogretmenler Cad. No:14,
YY., Balgat, Ankara, Turkey
e-mail: omayt@cankaya.edu.tr

economy. Estrella [1] has found that empirical relationships are not structural, and alternative monetary policy regimes could lead to very different outcomes. Estrella's [1] approach in this study is a unification of the theories which uses the term structure of interest rates for predicting different important economic variables. In the empirical part of his paper, he has employed a theoretical impulse response analysis to this unified model and found out that it can imitate a real economic system. Moreover, Jardet [17] has used this methodology in her study to analyze US and Canadian economies and employed a VAR-VECM model. She has found out that the predictability of the term structure of interest rate is reduced by the structural breaks, related to monetary policy changes. Furthermore, Estrella [1] has also found similar results for the US economy in his 2004 research.

In this paper, we have applied a nonlinear approach to an emerging market, Turkey, concerning a particular period, a period that was influenced by high inflation, high budget deficits, and political instability. Besides these handicaps, Turkey has been encompassed by other disadvantages of testing the predictability issue like not having wide and deep financial markets for the same periods.¹ Because of these features, all possible methods must be applied only very carefully to the Turkish economy. In particular, linear models are not sufficient to analyze these types of economies. For example, for predictability of inflation, Mishkin [11] has stated that "when the relationship between the term structure and inflation is unstable, it would not be possible to provide correct information about the inflationary pressures in the economy by using the estimated parameters of the linear model," and therefore the term structure of interest rates are no longer a guide for future monetary policies aiming at ensuring price stability. Following this advice and other reasons, we have employed a nonlinear model. Unfortunately, all papers mentioned above have investigated the stability issue with a structural breaks analysis, in linear bivariate single equation models except [10] to investigate spread-output relationships and [16] for spread-inflation relationship. In our analysis, we have employed STR-type nonlinearity to both the relationships like [10], but our analysis is the generalization of their analysis applied in multivariate case. By using this analysis, the stability and the causes of the predictability issue among these nexus will further be investigated. For this purpose, we are employing a Logistic Smooth Transition Vector Autoregressive (LSTVAR) model too obtain the Generalized Impulse Response Function (GIRF). GIRF analysis has several advantages over the traditional counterparts in analyzing these kinds of systems. Traditional Impulse Response Function (TIRF) analysis has its limitations and confines researchers to give one impulse response for one specific situation in the estimation period. For example, giving an impulse to interest rate variable has one response by inflation variable during the estimation period in a linear VAR model. But in a LSTVAR model, one can apply an impulse response to specific histories (events) in the estimation period. This feature enables researchers to employ impulse response analysis to specific variables both before and after an important event.

¹For the detailed information about the Turkish economy, [16] can be used.

Therefore, we have incorporated these advantages into our research to investigate the highlighted issues in our study. We are also investigating, the Expectation Theory, Interest Transmission Channel and Fisher Effect with assistance of the GIRF for Turkey through the estimation period. This analysis sheds light on the theoretical connections of predictability of term structure for Turkey. Moreover, we have mentioned above that the empirical relationships are not structural, and alternative monetary policy regimes could lead to varying outcomes. This result has the obvious but interesting empirical implication that shifts in the relationship between real output changes (or inflation) and the spread may provide evidence of shifts in the relative weights policy makers give to deviations of inflation or output from their targets (see [18]). In order to analyze the stability of these relationships, researchers can estimate the monetary policy reaction function and test whether the parameters of the monetary policy reaction function have changed or not. This feature can be cross-checked by a parameter constancy test which was established by Lin and Terasvirta [19]. In their specification test, they are testing linearity against multiple structural changes. On the other hand, there are other structural change tests like Bai-Perron, Chow, and Andrews. Therefore, after estimating a linear model of monetary policy reaction function, a researcher can apply these tests in addition to parameter constancy test. In our analysis, we are estimating a linear monetary policy reaction function in order to robustify the results of the stability of the relationship between real output changes (or inflation) and the spread. Therefore, this investigation may be seen as further proof of Estrella's [1] claim or not, in the case of Turkey.

The remainder of the paper is organized as follows: Sect. 2 reviews the previous literature and model. Section 3 discusses the data and methodology used. Section 4 presents the empirical results. Section 5 estimates the policy reaction function to establish the connection with monetary policy. Finally, Sect. 6 offers conclusions drawn from the study undertaken.

2 Previous Literature and Model

Harvey [2] is the pioneer study that investigates the relationship between real economic activity and spread. Harvey [2] shows that there is useful information in the real yield spread about future consumption growth. Estrella and Hardouvelis [4] argue that the nominal spread is a useful predictor of future growth in output, consumption, and investment. Plosser and Rounwenhorst [5] study shows that the term structure has a significant predictive power for long-term economic growth. Haubrich and Dombrosky [6] contend that the yield spread is an excellent predictor of four quarter economic growth but its predictive content has changed over time. Dotsey [7] has thoroughly investigated the forecasting properties of the yield spread for economic activity. To our knowledge, the only other study that has investigated the yield spread-future economic activity relationship in an emerging market is by

Kim and Limpaphayom [20]. Their results show that the term structure of interest rates has significant power in forecasting Indian real economic activity.

Mishkin [11] is the first study that has investigated the relationship between inflation and spread. In his study, he uses the difference between an n -months interest rate and an m -months interest rate to predict the difference between average inflation rates over n -months and m -months concerning the future, where $n, m \leq 12$. He has examined the spread for US Treasury bills. Although, he has found that for maturities of 6 months or less, the term structure of nominal interest rates have provided so little information about the future path of inflation, he has concluded, nevertheless, that for longer maturities of 9 and 12 months, the spread of nominal interest rates indeed provides information about the future path of inflation. Mishkin [12] obtains even better results with longer maturities, i.e., from 1 to 5 years. In both of these papers, the researchers have developed the theoretical background from the application of both the Fisher equation and the Expectations Hypothesis of the term structure of interest rates. On the other hand, Mishkin [11] has also examined data obtained from ten OECD countries and has found out that shorter maturities have provided almost no information about the future path of inflation. Estrella and Mishkin [15] study shows that the term structure has a significant predictive power for subsequent economic activities and the inflation rate by using Europe and the US data. Finally, Frankel [14] shows that a different measure of spread has a better predictive power by using US data.

All of these empirical studies above use linear models, which have limitations as they cannot control the possibility of asymmetric effects or structural shifts. Moreover, it is common knowledge that a number of economic variables exhibit a nonlinear behavior. Following this common knowledge, a limited number of studies have been motivated to employ nonlinear approaches like [10], and [16]. Venetis et al. [10] uses a TV-STAR model in order to investigate the relationship between real economic activity and the spread, for three developed countries. Telatar et al. [16] uses a time-varying parameter model with Markov-switching heteroskedastic disturbances in order to investigate the relationship between inflation and the spread for Turkey. Besides [1], there are two other studies that have investigated the Turkish economy, namely [21], and [22], which have employed linear models to test the predictability relationship between inflation and term structure of interest rate. Şahinbeyoğlu and Yalçın [21] have concluded that the term structure and inflation have a negative relationship between each other. They claim that the high volatility of real interest rates relative to expected inflation, as well as the negative correlation between these two variables, have produced significant and negative term structure coefficients as mentioned in [11] and [12].

Hence, we aim to contribute to the existing literature on this issue by employing a LSTVAR model to an emerging market and investigating the sources of the negative relationship between these nexus.

In this paper, we have used the model which is given below. This model was developed by Estrella [1]:

1. Phillips curve: $\pi_{t+1} = \pi_t + \alpha y_t + e_{t+1}$
2. IS curve: $y_{t+1} = b_1 y_t - b_2 \rho_t + \eta_{t+1}$

3. Monetary policy reaction function:

$$r_{t+1} = g_1 r_t + g_2 \pi_{t+1} + g_3 y_{t+1} + (1 - g_1 - g_2) z_{t+1}$$

4. Monetary shock: $z_{t+1} = z_t + v_{t+1}$

5. Fisher equation: $R_t = \rho + \frac{1}{2} (E_t \pi_{t+1} + E_t \pi_{t+2})$

6. Expectation hypothesis: $R_t = \frac{1}{2} (r_t + E_t r_{t+1})$

Parameter restrictions in the equation system:

$$0 < a < 1, 0 < b_1 < 1, 0 \leq b_2 \leq 1, 0 \leq g_1 \leq 1, g_2, g_3 \geq 0$$

where π_t is the inflation rate in period t , y_t is the output gap in period t , r_t is the short term (1 period) nominal interest rate, R_t is the long-term (2 period) nominal interest rate in period t , ρ_t is the long-term (2 period) real interest rate in period t , Z_t is the inflation target in period t , E_t is the rational expectations operator based on period t information and ε_t , $v_t \eta_t$: i.i.d. random variables.

Estrella [1] derives reduced form equation for the relationship between inflation, changes in real output, and the spread with a simple structural model of an economy given as above. Estrella's [1] model is based on the model used by Fuhrer and Moore [23] for empirical purposes. But, Estrella [1] has found a theoretical, implicit solution for this model. The advantage of Estrella's [1] model is the flexible choice of a monetary policy reaction function. The monetary policy reaction function can be simplified by the choice of parameters to either the [24] reaction function ($g_1 = 0$) or Fuhrer-Moore ($g_1 = 1$) version of the Taylor rule.

Estrella [1] and researches derived the reduced forms of this model by forward and backward methods and show that the coefficient linking expected future output, inflation, and short-term interest rates to spread, depending on the parameters in the monetary reaction function. For example, if the central bank only reacts to output deviations where $g_2 = 0$, the coefficient linking expected future output and spread in the reduced form is given as $(2/g_3)$. As a conclusion of Estrella's [1] work, monetary policy is a key determinant of the precise relationship between the term structure of interest rates and macroeconomic variables such as real output and inflation. With respect to Estrella's [1] rational expectation's model, he notes that with alternative monetary policy rules, one or more of the following can occur: the term structure spread is the optimal predictor of real output; the term structure is the optimal predictor of changes in inflation; the short-term nominal rate is the best predictor of real output; the short-term nominal rate is perfectly correlated with the long-term real rate; interest rates are not informative with regard to future output and inflation. However, Estrella [1] has found that the term structure of interest rates has at least some predictive power for both real output and inflation under almost all circumstances. From this discussion, it is clear that the empirical relationships are not structural, and alternative monetary policy regimes could lead to very different outcomes. This issue can be further analyzed by the method that we mentioned in the introduction.

3 Data and Methodology

3.1 Data

For the paper, we are using industrial production index and the consumer price index (CPI) downloaded from the website of the Republic of Turkey's Central Bank. In order to calculate the term structure of interest rate (spread), 1-month and 3-month government security interest rates are taken from the secondary market. These data are extracted from the Istanbul Stock Exchange (ISE) monthly journal.² Therefore, the spread variable sp is obtained by subtracting 1-month interest rate from 3-month interest rate.

3.2 Basic Regression of Predictability: Real Economic Activity and Inflation

In the paper, we consider the annual growth rate of the monthly industrial production index as “cumulative,” whereas some other papers use marginal growth rates:

$$\Delta^k y_t = \frac{12}{k} (y_t - y_{t-k}) \quad (1)$$

where y_t is the logarithm of the industrial production index at time t . We compute the slope of the nominal yield curve as the difference between the long-term bond yield $LR\ i_t$ and the short-term yield $SR\ i_t$ as $(LR\ i_t - SR\ i_t)$ at time t .

The following basic regression is a way to describe the relationship between the spread and future activity:

$$\Delta^3 y_t = \zeta + \beta sp_{t-1} + \varepsilon_t \quad (2)$$

where k is the forecast horizon and ε_t forecast error. Depending on the previous literature, the best forecasting horizon is a 3-month period for Turkey. Therefore, we only consider the forecast horizon of 3 months in this study. Moreover, Turkey is a high inflationary country thus, a long horizon can be at most 3 months with regard to long term economic decision. The fact that we are working with monthly data creates some temporal correlation between the successive error terms. In order

²The daily interest rates obtained from ISE are weighted with the transaction volumes to calculate the interest rates for 1-month and 3-month maturities. The 1-month interest rate is defined to have a maturity in the range of 20–40 days and the 3-month interest rate is defined to have a maturity in the range of 80–100 days.

to remedy this serial correlation problem, we have used [25] for standard error terms (see [17]). The estimated equation is given as below:

$$\Delta^3 y_t = 0.018 - 0.0004 s p_{t-1} + \varepsilon_t$$

(0.039) (0.0001)

$$r - \text{square} = 0.006, \hat{\sigma}_e = 0.337, \text{SK} = 0.313 (0.197), \text{EK} = -0.043 (0.930),$$

$$\text{JB} = 1.715 (0.424) \text{DW} = 1.254, \text{ARCH} (1) = 3.284 (0.069) \quad (3)$$

where heteroscedasticity consistent (hcc) standard errors are given in the parentheses below the parameter estimates, ε_t denotes the regression residual at time t , $\hat{\sigma}_e$ is the residual standard deviation, SK is skewness, EK excess kurtosis, JB the Jarque–Bera test of normality of the residuals, and ARCH is the LM test of no Autoregressive Conditional Heteroscedasticity (ARCH). Normality of residual is not rejected with Jarque–Bera test. SK and EK are also rejected. But, Durbin–Watson test shows that there is a significant autocorrelation problem. Moreover, the LM test for ARCH has significant value which can be the indication of neglected nonlinearity (see [26]). This final conjecture is investigated further by applying the LM-type linearity test of [27]. From (3), we can observe that the relationship between real economic activity and the spread is significant and negative like the previous findings. This linear model suggested that the spread has information content about real economic activity for Turkey; but this relationship is negative, whereas the relationship has been found to be positive in developed countries. This issue has one of the main questions posed in this paper and will be analyzed further in Sect. 4. On the other hand, the other main question is the stability of this relationship. For investigating the stability issue, we have employed two different approaches; one of them is a recursive Chow structural break test with which we can date the structural breaks; and the other one is parameter stability test, which is developed by Lin and Terasvirta [19]. A recursive Chow structural test for (3) shows three consecutive structural change points between 1999 and 2000. These structural points are $1999:3 = 2.901(0.038)$, $1999:4 = 2.885(0.039)$, and $1995:5 = 2.930(0.037)$. The first value is Chow test and the probabilities are given in parenthesis. The parameter stability test $\text{LM} = 2.283(0.081)$ suggests that the equation can be modeled by a TV-STAR model which shows that the parameter of the model is not constant. Hence, stability issue is not satisfied for the Turkish case by using linear model.

For the inflation case, we consider the marginal growth rate of monthly inflation rate as:

$$\Delta^k \pi_t = (\pi_t - \pi_{t-k}) \quad (4)$$

where π_t is the logarithm of the inflation rate, which is obtained by the CPI at time t .

The following basic regression shows the relationship between spread and future activity:

$$\Delta^k \pi_t = \zeta + \beta s p_{t-1} + \varepsilon_t \quad (5)$$

where k is the forecast horizon and ε_t forecast error. In this study we are only dealing with the forecast horizon of 3 months, because we are not dealing with the best predictability horizon. On the other hand, other papers, dealing with the predictability issue, have found that the best forecasting horizon is 3-month period for Turkey. This is theoretically sensible as well, because Turkey is a high inflationary country and a long horizon can be at most 3 months to decide to make sound long term investments.

$$\Delta^3 \pi_t = -40.063 - 6.870 sp_{t-1} + e_t$$

(31.173) (0.905)

$$r - \text{square} = 0.580 \hat{\sigma}_e = 0.054 \text{SK} = 0.001 (0.197) \text{EK} = -0.564 (0.930)$$

$$\text{JB} = 1.529 (0.424) \text{DW} = 0.707 \text{ARCH}(1) = 60.096 (0.000) \quad (6)$$

All the misspecification test results and their explanations are the same. However, this time linear model has better information content than the real economic activity case, which we can decide as looking at the r -square criteria.³ Again we have investigated the stability issue; we employed two different approaches, which are mentioned above. The recursive Chow structural test for (2) shows that there are many structural breaks at 1% significance level, starting from 96:3 = 5.815(0.000) until the end of the sample period. The most significant one is 99:8 = 29.771(0.000). The parameter stability test suggests that the equation can be modeled by a TV-STAR model, which shows that the parameter of the model is not constant and the test statistics here LM = 6.835(0.000). From these analyses, we have seen that the parameters of the linear models are not constant and the stability issues are not satisfied with the Turkish data.

3.3 *Specification of Smooth Transition Vector Autoregressive Model and Linearity*

3.3.1 Tests

The specification and estimation of multivariate STAR models are discussed in further details in [26] (Chap. 5). In the specification of a smooth transition vector auto-regression model (STVAR), we follow [26]. Let $x_t = (x_{1t}, \dots, x_{kt})'$ be a $(k \times 1)$ vector time series. We have $x_t = (y_t, p_t, i_t, sp_t)'$ where y_t is the log of industrial production index, i_t the log of nominal interest, p_t the log of the **CPI**, and sp_t the log of spread values. A k -dimensional STVAR then can be formulated as

³In both of the regression equations, we are using different dependent variables; hence, there is no direct comparison of R-squares.

$$\Delta x_t = \left\{ \phi_1 + \sum_{i=1}^{p-1} \phi_{1i} \Delta x_{t-i} \right\} \cdot (1 - F(s_t; \gamma, c)) + \left\{ \phi_2 + \sum_{i=1}^{p-1} \phi_{2i} \Delta x_{t-i} \right\} \cdot (F(s_t; \gamma, c)) + \varepsilon_t \quad (7)$$

where $\phi_j, j = 1, 2$ are $(k \times 1)$ vectors, $\phi_{j,i}, j = 1, 2, I = 1, \dots, p-1$ are $(k \times k)$ matrices, and $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{kt})$ is a k -dimensional vector white-noise process with mean zero and $(k \times k)$ covariance matrix Σ . The transition function $F(s_t; \gamma, c)$ is assumed to be a continuous function between zero and one, with parameters γ and c determining the smoothness and location of the change in the value of $F(s_t; \gamma, c)$, respectively. Here we have the following form logistic function

$$F(s_t; \gamma, c) = \frac{1}{1 + \exp\{-\gamma[s_t - c]\}}, \gamma > 0 \quad (8)$$

The specific-to-general approach for specifying univariate STAR models put forward by Terasvirta [28] can be adapted to the multivariate case. This procedure starts with specifying a vector autoregressive model for $x_t = (y_t, p_t, i_t, sp_t)'$, that is,

$$\Delta x_t = \phi_1 + \sum_{i=1}^{p-1} \phi_i \Delta x_{t-i} \cdot \varepsilon_t \quad (9)$$

where the order p should be such that residuals $\hat{\varepsilon}_t$ have zero autocorrelations at all lags. The choice of p is based on applying both the SIC and AIC in a linear VAR model for x_t with a deterministic linear trend. Both information criteria select $p = 1$ as the appropriate lag order.

The next step of the specification procedure is testing linearity against STAR-type nonlinearity as given in (1), with (2). However, the testing problem is complicated by the presence of unidentified nuisance parameters under the null hypothesis. This can be understood by noting that the null hypothesis can be expressed in multiple ways, either as $H_0 : \phi_1 = \phi_2$ and $\phi_{1,i} = \phi_{2,i}$ for $I = 1, \dots, p-1$ or as $H'_0 : \gamma = 0$. In order to overcome this problem, following the approach of [27], we replace the transition function $F(s_t; \gamma, c)$ with a suitable Taylor approximation. For example, a first-order Taylor approximation of the transition function results in the following auxiliary regression

$$\Delta x_t = A_0 + \sum_{i=1}^{p-1} B_i \Delta x_{t-i} + A_1 s_t + \sum_{i=1}^{p-1} B_{1,i} \Delta x_{t-i} s_t + e_t \quad (10)$$

where e_t comprises the original shocks ε_t as well as the error arising from the Taylor approximation. In (10) it is assumed that the transition variable s_t is not one of the elements of Δx_{t-i} , $i = 1, \dots, p-1$ or their linear combination. If this is not the case, the term $A_1 s_t$ must be dropped from the auxiliary regression. The parameters in A_j and $B_{j,i}, j = 0, 1, I = 1, \dots, p-1$, are functions of the parameters in the original STVAR. In this case, the null hypothesis of linearity can be expressed as $H''_0 : A_1 = B_{1,i} = 0, I = 1, \dots, p-1$, that is, the parameters associated with the auxiliary regressors are all zeros. This null hypothesis can be tested by a standard variable addition test in a straightforward manner. The test statistics, to be denoted

Table 1 LM-type test against linearity

Candidate transition variables	Equation				System-wide test
	Δp_t	Δi_t	Δy_t	sp_t	
Δp_{t-6}	3.880 (0.000)	1.135 (0.344)	1.809 (0.089)	2.415 (0.022)	65.664 (0.000)
$\Delta^3 p_{t-6}$	1.937 (0.067)	0.519 (0.818)	1.316 (0.246)	1.803 (0.091)	67.824 (0.000)
Δy_{t-1}	3.014 (0.005)	0.569 (0.779)	4.373 (0.000)	3.573 (0.001)	104.128 (0.000)
Δy_{t-6}	0.898 (0.509)	0.613 (0.744)	0.473 (0.852)	3.217 (0.003)	87.663 (0.000)
sp_{t-1}	6.205 (0.000)	4.385 (0.000)	4.585 (0.000)	23.774 (0.000)	110.094 (0.000)
Δsp_{t-1}	5.143 (0.000)	1.537 (0.159)	2.947 (0.006)	17.922 (0.000)	101.927 (0.000)
Δi_{t-2}	3.760 (0.000)	1.463 (0.184)	2.218 (0.036)	2.164 (0.040)	49.827 (0.006)
Δi_{t-2}	1.430 (0.197)	0.745 (0.634)	1.602 (0.139)	0.762 (0.619)	47.709 (0.011)
t	6.412 (0.000)	3.650 (0.001)	3.287 (0.002)	3.285 (0.002)	42.891 (0.035)

The most suitable candidate variables are given. Besides, the other test statistics are available upon request

as LM_1 , have an asymptotic χ^2 distribution with $k(q+1) + (p-1)k^2$ degrees of freedom under the null of linearity. Since the LM_1 statistic does not test the original null hypothesis $H'_0: \gamma = 0$ but rather the auxiliary null hypothesis $H''_0: A_1 = B_{1,i} = 0$, this statistic is usually referred to as an LM-type statistic.

As noted by Luukkonnien et al. [27], the LM_1 statistic has no power in situations where only the intercept in the VAR varies across regimes, that is, when $\phi_1 = \phi_2$ but $\phi_{1,i} = \phi_{2,i}$ for $I = 0, 1, \dots, p-1$, in STVAR given in (1). Luukkonnien et al. [27] suggest to remedy this problem by replacing the transition function $F(s_t; \gamma, c)$ by a third-order Taylor approximation, which results in the auxiliary regression (4) with, s_t^j and $\Delta x_{t-i} s_t^j$, $j = 2, 3$, $i = 1, \dots, p-1$ as additional auxiliary regressors. The test statistic computed from the augmented auxiliary regression is LM_2 statistic. Since only the parameters corresponding to s_t^2 and s_t^3 are functions of ϕ_1 and ϕ_2 , a parsimonious, or “economy” version of the LM_2 statistic can be obtained by augmenting (10) with additional regressors s_t^2 and s_t^3 . The resultant statistic is the LM_3 statistic.

To identify an appropriate transition variable s_t , the LM-type statistic can be computed for several candidates, and the one for which the associated p-value of the test statistic is smallest can be selected. In this paper, we consider 54 different candidate transition variables; lagged growth rates in output (Δy_{t-i}), lagged growth rates in interest (Δi_{t-i}), lagged inflation rates (Δp_{t-i}), and lagged rate of change of the spread rates (Δsp_{t-i}). By being a linear combination of real and monetary variables, such a transition variable can capture both nominal and real rigidities. Since monthly time series exhibit considerable short-run fluctuations which do not necessarily represent changes in regimes, we also consider quarterly changes in the above-mentioned variables (i.e., $\Delta_3 \theta_{t-i} = \theta_{t-i} - \theta_{t-i-3}$ where θ_t is one of the elements in the time series vector x_t). We set $i = 1, \dots, 12$ in all cases.

The results of the linearity tests are shown in Table 1, which reports only the candidate transition variables for which the null of linearity of the output equation is rejected.

The results of both system-wide linearity tests as well as a test of linearity of the output equation suggest Δp_{t-i} , Δy_{t-i} , Δi_{t-i} and the spread as appropriate transition

Table 2 Estimation results of LSTVAR

Parameters	DCPI	DGSMH	SP	DSRI
Constant	−0.059 (0.017)	0.194 (0.056)	−2.345 (32.931)	−2.352 (0.475)
$L - \Delta p_{t-1}$	0.983 (0.015)	0.072 (0.049)	−9.827 (28.813)	−1.519 (0.424)
$L - \Delta y_{t-1}$	−0.057 (0.024)	0.585 (0.077)	−60.994 (45.551)	−1.271 (0.651)
$L - sp_{t-1}$	−0.0002 (0.000)	0.0002 (0.000)	0.174 (0.068)	−0.003 (0.000)
$L - \Delta i_{t-1}$	0.015 (0.005)	−0.051 (0.016)	0.731 (9.430)	0.751 (0.138)
$U - \Delta p_{t-1}$	0.002 (0.001)	−0.015 (0.005)	−5.310 (3.070)	0.145 (0.044)
$U - \Delta y_{t-1}$	0.081 (0.042)	0.062 (0.136)	14.422 (79.991)	−1.701 (1.1436)
$U - sp_{t-1}$	0.0002 (0.000)	−0.0004 (0.000)	1.407 (0.139)	−0.006 (0.001)
$U - \Delta i_{t-1}$	−0.005 (0.004)	0.030 (0.013)	13.072 (7.713)	−0.236 (0.111)
R^2	0.985	0.521	0.634	0.487
γ	5.119 (8.550)			
C	10.736 (0.394)			

variables. Since the null of linearity is rejected more strongly for Δsp_{t-1} , in price, output, and short run interest rate equations, we use this variable as the system-wide transition variable in estimation of the STVAR.

Given the choice of the transition variable s_t , estimation of the parameters in the STVAR (1) is a relatively straightforward application of nonlinear least squares, which is equivalent to quasi-maximum likelihood based on a normal distribution. Under certain (weak) regularity conditions, which are discussed by White and Domowitz [29], and Pötscher and Prucha [30], among others, the NLS estimates are consistent and asymptotically normal.

The estimation can be performed using any conventional nonlinear optimization procedure. The burden on the optimization algorithm can be alleviated by using good starting values. For fixed values of the parameters in the transition function, γ and c , the STVAR model is linear in the parameters ϕ_j , $\varphi_{j,i}$, $j = 1, 2$, $I = 1, \dots, p - 1$, and therefore, can be estimated by OLS. Hence, a convenient way to obtain sensible starting values for the nonlinear optimization algorithm is to perform a two-dimensional grid search over γ and c . Furthermore, the objective function (the log of the determinant of the residual covariance matrix) can be concentrated with respect to ϕ_j , $\varphi_{j,i}$, $j = 1, 2$, $I = 1, \dots, p - 1$. This considerably reduces the dimensionality of the NLS estimation problem, as the objective function needs to be minimized with respect to the two parameters γ and c only. Parameter estimates of LSTVAR model are given in Table 2.

After estimating the LSTVAR model, we test whether the estimated model adequately captures the nonlinear features of the time series under examination. Specifically, we test the model for remaining nonlinearity and for parameter constancy. The testing procedure is outlined in Appendix D in [31], who generalize the procedure developed by Eitrheim and Terasvirta [32] to the multivariate context. In addition to the candidate transition variables used for testing linearity of the baseline model, we also use semiannual changes in the same variables (Table 3).

Table 3 Test against multiple regime LSTVAR

Candidate transition variables	Additive	Multiplicative	Additive and multiplicative together
Δp_{t-6}	57.106 (0.172)	56.313 (0.191)	34.420 (0.999)
$\Delta^3 p_{t-6}$	28.056 (0.990)	46.493 (0.534)	30.028 (0.999)
Δy_{t-1}	29.912 (0.981)	40.417 (0.773)	26.116 (1.000)
Δy_{t-6}	27.175 (0.993)	36.434 (0.889)	19.379 (1.000)
sp_{t-1}	49.731 (0.404)	51.182 (0.349)	39.359 (0.999)
Δsp_{t-1}	30.595 (0.976)	40.728 (0.762)	37.753 (0.999)
Δi_{t-2}	38.831 (0.824)	51.920 (0.323)	36.046 (0.999)
Δi_{t-4}	40.525 (0.769)	38.643 (0.830)	26.623 (1.000)
T	40.366 (0.775)	47.586 (0.489)	33.781 (0.999)

Table contains the smallest p value test statistics. Other test statistics are available upon request

4 Investigations of Expectation Theory Interest Transmission Mechanism Fisher Effect and Stability of Predictability: Generalized Impulse Response Analysis

In this section, we have computed a generalized impulse response analysis in order to detect the effects of structural changes and crises in predicting future output and inflation for Turkey. From this analysis, we will detect whether the predictability relationship variables mentioned are stable in the estimation period or not. Generalised Impulse Response Functions have advantages on their linear counter parts per [33]. Hence, we use GIRF as an indicator tool for visualizing the effects of structural changes and crises.

The TRIF has characteristic properties in case the model is linear. First, the TRIF then is symmetric. In the sense that a shock of $-\delta$ has exactly the opposite effect as a shock of size $+\delta$. Furthermore, it might be called linear, as the impulse response is proportional to the size of shock. Finally, the impulse response is history independent as it does not depend on the particular history w_{t-1} [26]. These properties do not carry over to nonlinear models. In nonlinear models, the impact of a shock depends on the sign and the size of the shock, as well as on the history of the process. The GRIF, introduced by Koop et al. [33] provides a natural solution to the problems involved in defining impulse responses in nonlinear models. The GIRF is a function of δ (shock) and w_{t-1} (history), which are the realizations of the random variables ε_t and Ω_{t-1} . Koop et al. [33] stress this; hence the GIRF is a realization of a random variable. Using this interpretation of the GIRF as a random variable, various conditional versions can be defined, which are of potential interest. In our case, to analyze the stability and source of the relationship, we give a shock to the estimated model for every specific year in the sample period.

During this analysis, we have considered whether “Expectations Theory” and “Interest Transmission Mechanism” operate simultaneously in the same direction or not, in order to investigate the source of the predictability relationship between

spread and real economic activity. In almost all empirical studies, these mechanisms are assumed to be the theoretical structure of real economic activity and the spread relationship. For example, if the Central Bank decides to exercise an expansionary monetary policy, the first effect of this policy will be a fall in the interest rate of all maturities. According to the Expectations Theory of the interest rate's term structure, long-term interest rates will be less than short-term interest rates. The simultaneously falling interest rates result in an increase in investment and, successively, an increase in real economic activity, whereas the higher fall in short-term interest rates relative to long-term interest rates causes an increase in the spread. As it can be seen from the above discussion, the mechanism is very simple. The spread and real economic activity move in the same direction with expansionary monetary policy. The move in the same direction can be demonstrated as the theoretical reason why the spread is an indicator of future real economic activity. On the other hand, Expectation Theory with Fisher Hypothesis is the key theoretical underpinning for the predictability of inflation. Moreover, there are two important assumptions; rational expectation and constancy of ex-ante real rates must be satisfied. Hence, we are also inspecting the spread and inflation relationships with a GIRF analysis. The second important relationship for the predictability of inflation, Fisher Hypothesis is analyzed by dynamic correlation analysis, which is also obtained by GIRF analysis. High volatility of the real interest rates relative to expected inflation, as well as the negative correlation between these two variables, produce significant and negative term structure coefficients as mentioned in [11] and [12]. Both of these arguments have revealed the predictability relationship of inflation for Turkey while we are investigating the relationship.

Among many authors who have investigated the relationship between monetary policy and the spread are [34]. Cook and Hahn [34], first firmly established the positive empirical relationships between target rates and long-term rates, and interpret their findings as supportive of the Expectations Theory of the term structure. The expectations theory says that a long-term interest rate should be equal to the sum of short-term interest rates over the same period of time plus a term premium; thus an increase in the first couple of short-term rates should drive up the long-term rate too, but by less. However, Romer [35] produced the opposite results to [34]. To them, positive movement in the long-term rate is inconsistent with standard monetary theory; this is, in short, a puzzle. According to received theory, they claim, an increase in short-term rates should reduce inflation, and hence reduce the level of long-term rates sufficiently. Romer [35] suggests that the puzzle can be resolved if the central bank has access to private information about economic fundamentals, but they do not develop their argument formally. In short, we can say that some authors argue that long-term rates should increase as monetary policy is tightened, mainly via the expectations hypothesis of the term structure. Others support the hypothesis that a monetary tightening should increase short rates but decrease long-term rates as inflation expectations fall.

Ellingsen and Söderström [36] expand Romer's [35] idea. According to [36], if monetary policy reveals information about economic developments interest rates of all maturities move in the same direction in response to policy innovation as if

monetary policy reveals the central bank's policy preferences interest rates short-term and long-term rates move in the opposite direction. So the first proposition of [36] supports the Expectation Hypothesis and they called this type of "policy endogenous". For both of the theoretical explanations, an increase in short-term interest rate (tight monetary policy) leads to a decrease in the spread variable. From Estrella's [1] argument, we can derive a new relationship which is revealing the Expectation Hypothesis; if the short-term rate increases (decreases) then the spread will decrease (increase), because short-term rates should drive up the long rate too, but by less; when monetary policy reveals the central bank's policy preferences then short rate and long rate move in opposite direction, which again leads to a decrease. Therefore, we can deduce Expectation Hypothesis by analyzing impulse response between the short-term interest rate (which can be the monetary policy tool) and the spread variable; giving positive one standard shock to the short-term interest rate leads to a negative response of the spread variable in impulse response analysis, which indicates that the Expectation Hypothesis is valid for that period for any type of theoretical explanation. In an opposite case, we will conclude that the Expectation Hypothesis is not valid.

We have given GIRF to nine different periods covering the sample period. From the impulse response given to extract the relationship of short-term interest rates and the spread, we have found out that six of them are negative, which indicates that the Expectation Hypothesis is valid for those periods. Besides, we have found two significant positive relationships covering the periods through 1998:5 to 2000:8 and one insignificant positive relationship covering the period through 2000:5 to 2001:8. These periods are important for the Turkish economy, which experienced heavy banking crises. The structural break analysis which is applied to univariated predictability equations shows the same dates as significant structural breaks. From these results, we can conclude that the GIRF findings are consistent with the earlier research in this study. The results of GIRF analysis can be seen from the Fig. 1 given below.

The relationships between monetary policy (or the short-term interest rate) and real economic activity can also be analyzed employing impulse response analysis in the estimation period. This analysis shows us whether the interest rate transmission channel of Turkey is working or not. When an increase (decrease) occurred in the short-term interest rate, this causes a decrease (increase) in real economic activity, e.g., by stimulated investments; this mechanism is called as "Interest Rate Transmission Mechanism" (see [37]). Hence, we can analyze this Transmission Channel by giving positive one standard impulse to short-term rates and tracing the response of this impulse on real economic activity. From the above argument, the short-term interest rate has to have a negative relationship with real economic activity; hence when we observe a positive response in terms of real economic activity, which indicates that the Transmission Channel is not working for that period.

We have given a GIRF to nine different periods covering the sample period. From the impulse response given to extract the relationships of short-term interest rate and real economic activity, we have found that six of them were positive, which indicates that the Interest Rate Transmission Channel was not working for these

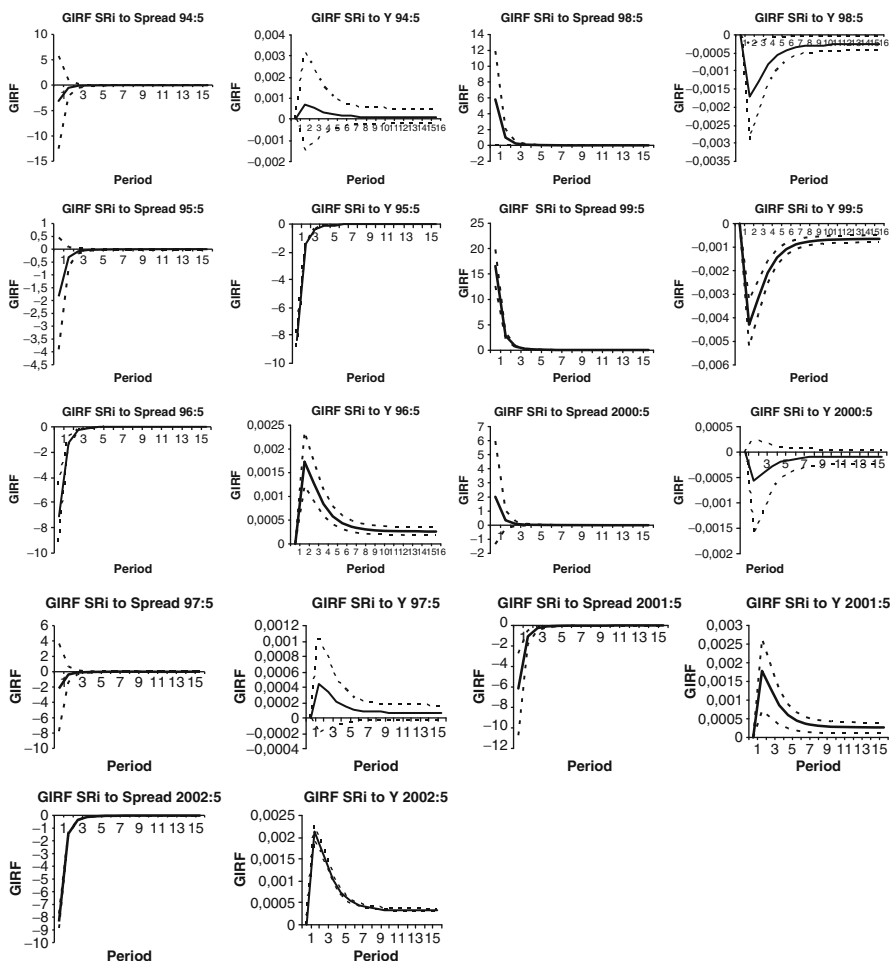


Fig. 1 Investigation of expectation theory and transmission channel: GIRF analysis

periods. Besides, we have found two significant negative relationships which cover the period through 1998:5 to 2000:8; one insignificant negative relationship which covers the period through 2000:5 to 2001:8. These periods are important periods in which the Turkish economy experienced banking crises, and which we mentioned above for the Expectation hypothesis. In the sample period, Turkey has a very high real interest rate which prevents domestic investors from making investments in real sector. Hence, this evidence shows the main obstacle of the Interest Transmission Channel to work properly for the sample period.⁴ From Fig. 1, we can trace the GIRF analysis of Interest Transmission Channel through the sample period.

⁴Conversely, during the crises the Interest Transmission Channel worked. What could be the cause of this abrupt phenomenon? This question is a subject of a further study.

One can easily recognize that generalized impulse response analysis shows an asymmetric relationship between Expectation Hypothesis and Interest Transmission Channel. When the Expectation Hypothesis is valid for Turkey, the Interest Rate Transmission Channel is not working and vice versa, along the sample period. On the other hand, this systematic and reverse symmetric movement explains the negative relationships between the spread and real economic activity. This systematic reverse (symmetric) movement explains why the spread predicts real economic activity negatively. For example, again if the Central Bank decides to exercise an expansionary monetary policy, the first effect of this policy will be a fall in the interest rate of all maturities. According to the Expectations Theory of the term structure interest rate, long-term interest rates will fall less than short-term interest rates. Unfortunately, the simultaneously falling interest rates do not result in an increase in investment and, successively, an increase in real economic activity, because of high real interest rates. As can be seen from this discussion, the mechanism does not lead to a move in the same direction between spread and real economic activity. Thus, we have obtained a negative predictability relationship between these variables. Alternatively, this result can be obtained by discussing an opposite argument.

Fama's [38] seminal paper found that real interest rates are constant over time, with fluctuations in nominal rates mainly reflecting changes in expected inflation (the so-called Fisher Effect). We have two possible methods to investigate the Fisher Hypothesis. First one is a direct approach of employing the GIRF analysis. In the first approach, the expected inflation must be specified which is an important component of Fisher equation:

$$i_t = \rho + \beta \pi_t^e, \quad (11)$$

i_t , nominal interest rate, ρ is constant, including constant real interest rate, and π_t^e is expected inflation rate. Barro and Gordon [39] and Lucas [40] suggest that the expected inflation rate, which is unobservable, may be systematically related to past rates of inflation. Hence, using the lag of past rates of inflation as a proxy for the expected rate of inflation, we can employ GIRF analysis to inspect Fisher Effect. The second approach is dynamic correlation approach, which is one of the methods used for investigating the long-run relationships between the nominal interest rate and inflation. In order to obtain dynamic correlation, a VAR analysis must be held in which the variables are filtered from structural breaks in level (see [41]). In our case, we are not in need of filtering data because of using nonlinear VAR technique, which considers the structural breaks while we are dealing with the data. The dynamic correlation is obtained by giving orthogonal shock in the inflation rate, which is taken as exogenous, and by tracing the effects of this shock in the interest rate and

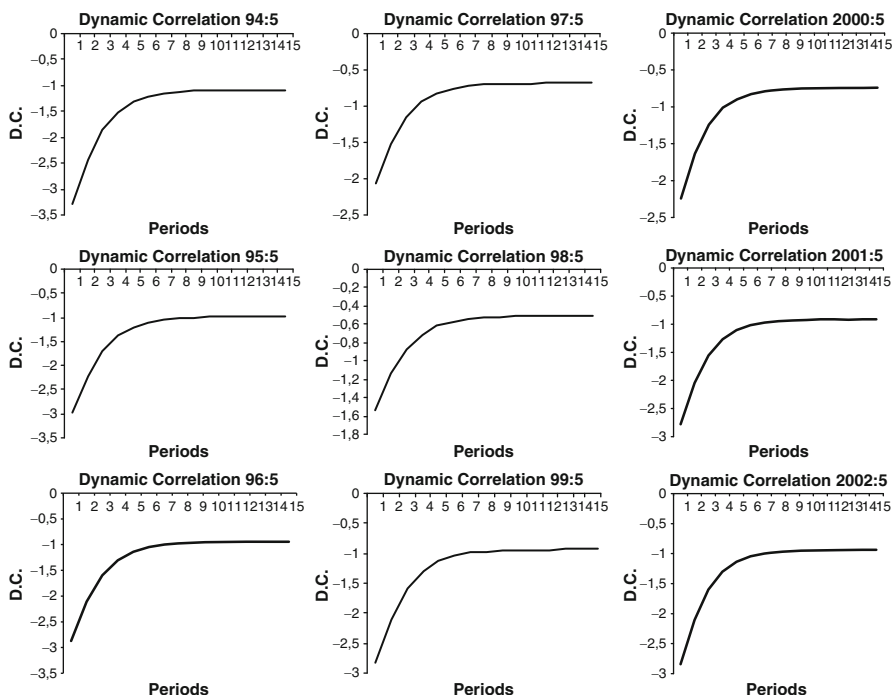


Fig. 2 Dynamic correlation analysis for long run Fisher effect

the inflation rate. The dynamic correlation is the ratio of the cumulative response function for the interest rate to that the inflation rate, calculated as:

$$\rho_s = \frac{\sum_{i=0}^s C_{11,i}}{\sum_{i=0}^s C_{21,i}}, \quad s = 1, 2, \dots, S. \quad (12)$$

If the long-run Fisher Effect is true, then the dynamic correlation has converged to 1.0 for large s (s is period for our GIRF analysis which is taken as 15 months⁵). Both analyses give the same results except for the 98:5, 99:5, and 2000:5 periods which are also sensitive for Expectation Theory and Transmission Mechanism. The first analysis is changing the sign negative to positive in these periods, but dynamic correlation is the same negative sign all through the sample period. Also dynamic correlation has changed in these periods as decreasing effect of the correlation. These periods are the significant structural breaks in the linear model so again these analyses are consistent with the other analyses, which we held before. The dynamic correlation analyses are given in below Fig. 2.

⁵Fifteen months is a very long period for an emerging market which has a high and volatile inflation rate.

Except 97:5, 98:5, and 2000:5 the dynamic correlation is converging to -1.0 . These negative correlations cannot be explained as inverted Fisher Effect. In inverted Fisher Effect the relationship is obtained by a real interest rate where as we are using a nominal interest rate. Hence, we can conclude that the relationships between a nominal interest rate and expected inflation is negative for the first analysis that we designed except the 98:5, 99:5, and 2000:5 periods and for long-run relationships it is negative all over the sample period. As we have mentioned before, the results are traced to a combination of the Fisher equation and the Expectation Hypothesis of the term structure of interest rate. The strength of the results, however, is seen to hinge on the nature of the variability of the unobserved real interest rates and on their relationships with the observed variables [1]. Both Expectation Hypothesis and Fisher equation are negative in our sample period. Negative results have been attributed to three factors: time-varying risk premia, the complexities of movement in the real rate, and the influences of monetary policy [1]. If we think about the predictability of the inflation equation (5), it is obtained by taking the difference of the Fisher equation for two interest rates with maturities m and n : $i_t^m = r_t^m + E_t \pi_{t+m}$, and $i_t^n = r_t^n + E_t \pi_{t+n}$, respectively. The results become an (5) if ex-ante real interest rates are constant and expectations are rational. Therefore, we can easily conclude that the negative predictability of inflation is caused by negative Fisher effect and Expectation Hypothesis.

From the linear bivariate single-equation analysis of inflation and real economic activity, we have found significant structural breaks with the recursive Chow test. From the GIRF analysis, we have found the same pattern in these relationships. As we have stated before, the GIRF analysis has its own advantages while analyzing data. Hence, we use these advantages in order to deduce stability relationship from the estimated model. Both generalized impulse responses of the spread to inflation and also real economic activity have significant and negative impacts. For the spread to inflation GIRF, the periods 94:5 to 95:8, 95:5 to 96:8, 96:5 to 97:8, 97:5 to 98:8, 2001:5 to 2002:8, and 2002:5 to 2003:8 have a significant negative impact of between -0.005 and -0.01 , whereas for periods 98:5 to 99:8 and 2000:5 to 2001:8 there is a significant negative impact around -0.01 . For the period 99:5 to 2000:8, there is a significant negative impact greater than -0.01 . These impulse response results are consistent with the recursive Chow test results. Again the most significant Chow test result date is the biggest impact period for the GIRF analysis. GIRF analysis shows us a more detailed picture of the structural break points without losing any degree of freedom. Moreover, we can follow the conditions of the relationships at every point of sample. For the spread to real economic activity GIRF, the periods 94:5 to 95:8, 95:5 to 96:8, 96:5 to 97:8, 97:5 to 98:8, 2001:5 to 2002:8, and 2002:5 to 2003:8 have a significant negative impact of between -0.006 and -0.008 , whereas for the periods 98:5 to 99:8 and 2000:5 to 2001:8 there was a significant negative impact around -0.01 and for period 99:5 to 2000:8 there was a significant negative impact greater than -0.01 . These impulse response results are consistent with the recursive Chow test results. Again the most significant Chow test result date is the biggest impact period for the GIRF analysis. From these results, we can conclude that the GIRF analysis can be used as a helpful tool for finding the

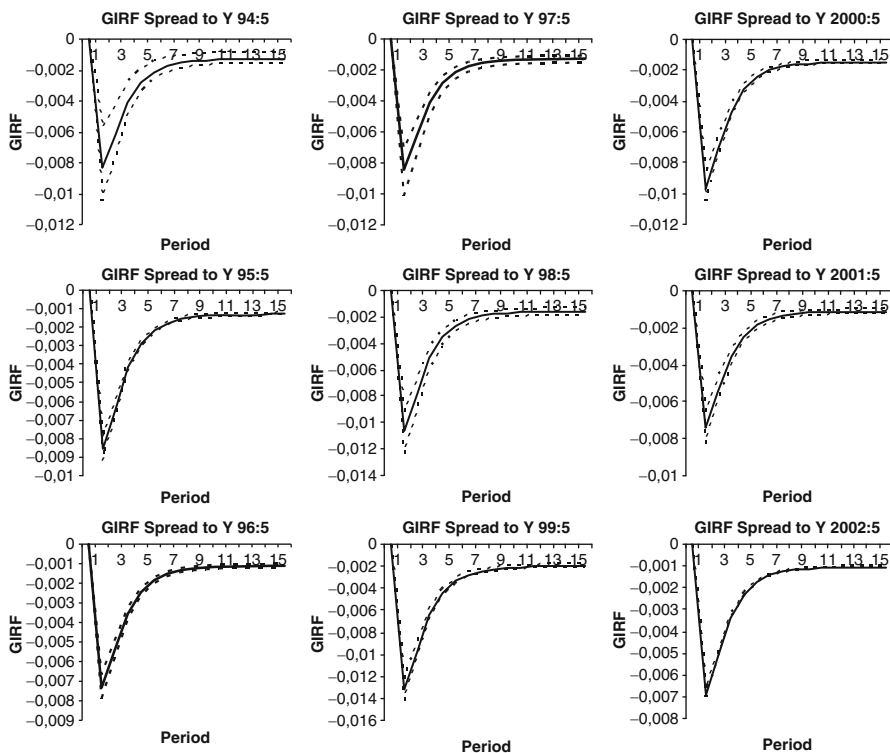


Fig. 3 Stability analysis of real economic activity predictability: GIRF analysis

break points of a relationship. On the other hand, there are numerous structural break points in the sample period, which cannot be analyzed by linear models, because of dividing them into subperiods. Most probably, subperiods have a very small sample size that cannot be analyzed by conventional statistical techniques. Especially, there are consecutive break points in the inflation predictability equation which leads to only one observation point for each subsample. Therefore, analyzing the changing structure of predictability can only be done by the GIRF method, which we have suggested in the beginning of this section (Fig. 3).

From the GIRF analysis, we have recognized that the negative predictability power increases in the periods of crises, when the Transmission Channel is working, providing that the Expectation Theory is not valid. In our opinion, this point needs further investigation. From this analysis, we can conclude that the Interest Rate Transmission Channel has more influential effects than the Expectation Hypothesis on predictability of the real economic activity. This point indicates that the information content of spread is more derived from monetary policy than Expectation Theory, which is the rival theory in the predictability literature. In a sense, then, this result supports Estrella's [1] conclusions, which we will analyze in the next section in order to prove our assessment (Fig. 4).

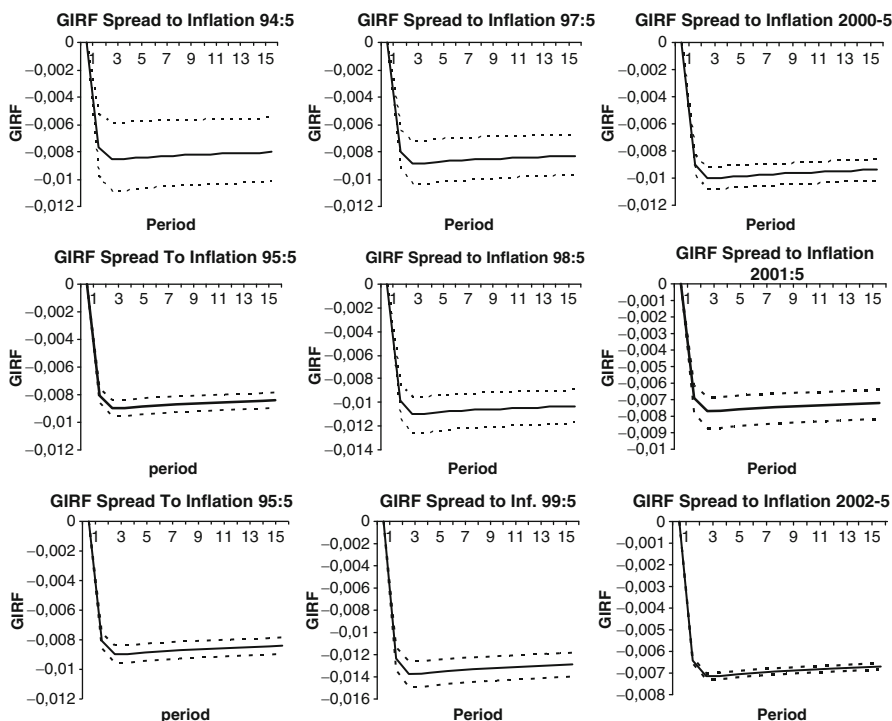


Fig. 4 Stability analysis of inflation predictability: GIRF analysis

5 Investigation of Effects of Monetary Policy Reaction Functions to Stability of Term Structure of Interest Rate

From Sect. 4, we have investigated the source of the negative relationship between real economic activity (and inflation) and the spread. Sections 3.2 and 4 has shed light on the stability of prediction relationships of variables mentioned for Turkey. On the other hand, from the Estrella's [1] discussion, it is clear that empirical relationships are not structural, and alternative monetary policy regimes could lead to very different outcomes. This result has the obvious but interesting empirical implication that shifts in the relationships between real output changes (or inflation) and the spread may provide evidence of shifts in the relative weights policy makers give to deviations of inflation or output from their targets (see [18]). In order to analyze the stability of these relationships then, a researcher should estimate a monetary policy reaction function and test whether the parameters of the monetary policy reaction function change or not. This feature can be examined by a parameter constancy test which is developed by Lin and Terasvirta [19]. In their specification test, they test linearity against multiple structural changes with STAR non-linearity.

Table 4 Estimation of policy reaction function

Model 1 $r = \beta + \beta r_{t-1} + \beta r_{t-2} + \beta \pi^* + \beta ygap$						
Model 2 $r = \beta + \beta r_{t-1} + \beta r_{t-2} + \beta \pi^* + \beta ygap + \beta \pi_{t-1}$						
Coefficient	Constant	$r(1)$	$r(2)$	π_{t+3}	$ygap$	$\pi(-1)$
Estimation	0.476	0.475 (0.195)	0.398	-0.564	0.811	-
Model 1	(0.638)		(0.225)	(0.534)	(0.475)	
Statistics	$R^2 = 0.336$	ARCH(4) = 0.691		$\hat{\sigma}_e = 0.297$	J.test = 0.999	
Estimation	0.859	0.427	0.320	-0.991	0.419	2.570
Model 2	(0.430)	(0.197)	(0.206)	(0.374)	(0.443)	(1.523)
Statistics	$R^2 = 0.349$	ARCH(4) = 0.691		$\hat{\sigma}_e = 0.293$	J.test = 0.999	

The values which are given under the parameter estimates are heteroscedastic consistent standard errors which are obtained by Newey-West (1987) procedure.

Berüment and Malatyalı [42] and Kalkan et al. [43], show that the inter-bank interest rate is the best variable indicating the policy stance of the Turkish Central Bank. Consequently, in this paper, we are using the inter-bank interest rate as the dependent variable of the policy reaction function estimation. Independent variables are forward looking inflation rates ($\pi_{t+3} = \log p_{t+3} - \log p_t$) and the output gap.⁶ The data before have been obtained again from the same source as before. The variables in the question are selected according to two criteria. First the consistency of the criteria in the investigation per previous literature and the second is Turkey’s specific economic environment. Although the most influential paper about the monetary policy reaction estimation is [44], suggesting the use of a forward-looking policy reaction function; we are using the backward-looking type of this function as well. We are using the forward-looking version in model 1 in order to be consistent with the previous studies, and we are using the backward-looking version in model 2 in order to be consistent with the Turkish economy’s special conditions. Turkey is a high inflationary and politically unstable country; hence these conditions let Central Bank use backward version of the function (see [42]). Besides these issues, the estimation is made by GMM technique, which is recommended by Clarida et al. [44] and Berüment and Malatyalı [42]. Estimated models of policy reaction functions are given below⁷ (Table 4).

With respect to model 1, the output gap has a significant coefficient estimate, whereas forward-looking inflation has an insignificant coefficient estimate. From these estimates, we can conclude that the Turkish Central Bank has targeted the output gap instead of the forward-looking inflation rate. When the lagged value of inflation was included into model 2, the output gap coefficient has an insignificant

⁶Output gap is obtained from seasonally adjusted industrial production index which is the deviation from the trend taht is calculated with Hodrick Prescott filter.

⁷The nonlinear estimation of the same sample periods of monetary reaction function can be found in [45] and [46].

statistic, whereas the lag value of inflation has a significant coefficient estimate. This result indicates that the Turkish Central Bank has targeted the backward-looking inflation rate. From now on we are only using model 2, because, this model describes the Turkish monetary authority behavior. In order to analyze the stability of this relationship, we have estimated a monetary policy reaction function and tested whether the parameters have changed or not. The test statistic for model 1 is $LM = 0.852(0.599)$, and model 2 is $LM = 2.743(0.001)$. From model 2, we have seen that the parameter estimates of the model are not constant. The recursive Chow structural test for model 2 shows that there are many structural breaks at 1% significance level, starting from $96 : 1 = 7.739(0.000)$ until the end of the sample period. The most significant one is $99 : 7 = 13.933(0.000)$. The first number is Chow test and the probabilities are given in parenthesis. The linear model of inflation predictability's structural dates and the structural dates of policy reaction function have a similar pattern; this shows there is a close relationship between these variables (inflation and monetary policy, which are shown by the short run interest rate) for Turkey. For the real economic activity, the same pattern cannot be observed; so we cannot make the same kind of argument with respect to these variables (real economic activity and monetary policy, which are shown by the short run interest rate). Estrella's [1] conclusion can be taken as a theoretical result for developed countries, but this conclusion may not be a valid result for some developing countries, especially, for a developing country which has multiple structural breaks in their data. Multiple structural breaks do not permit a researcher to come to a conclusion that the spread variable has information on monetary policy nor can it enable a prediction of real economic activity or inflation. In the case of inflation in Turkey, the above evidence is relatively weak. But for the real economic activity, the above structural break analysis is a contradiction for information content of spread. Besides, Peel and Ioannidis [18] have concluded that the ideal data for analyzing the theoretical prediction of Estrella's analysis are those of the United States and Canada. For the US and Canada data, one significant structural break has been found and this leads us to understand the predictability content of term structure. Hence, multiple structural breaks and the policy preference of Turkish money authority are not appropriate for analyzing Estrella's [1] theoretical prediction.

6 Concluding Remarks

This study investigates whether the term structure of interest rates contains useful information about future real economic activity and inflation for Turkey covering the period 1991–2004. We employ a recursive Chow Structural Break Test to the linear models and from this recursive test procedure, we have seen that the spread-real economic activity and spread-inflation relationships are not stable. This conclusion is also reaffirmed by the linearity test against STR nonlinearity, which is given as a parameter stability test. On the other hand, we have employed GIRF analysis to LSTVAR model in order to understand the source of the negative

relationship between these variables. This analysis shows us that the negative relationship between spread and real economic activity has occurred because of the negative symmetric relationship between Expectation Hypothesis and Interest Rate Transmission Channel. And, also this analysis shows us that the negative relationship between spread and inflation has occurred because of the negative Expectation Hypothesis and Fisher Effect. Furthermore, we have employed a GIRF analysis in order to see whether the predictability relationships are stable for Turkey. This analysis shows us that the stability relationship of these variables is not established for the period that we have analyzed. From this analysis, we have seen that the GIRF analysis is consistent with the recursive Chow test and parameter stability test at the same time. The GIRF analysis has its own advantages for analyzing the data which we have mentioned before. Especially, we have seen that the predictability power of these relationships can be analyzed by the GIRF without losing any degrees of freedom which, otherwise, cannot be obtained by linear bivariate equation estimation of subsamples as we have pointed out in Sect. 4. Finally, use of GIRF enables us to make a computation about the Expectation Hypothesis, Interest Rate Transmission Channel, and Fisher Effect, which cannot be obtained by TIRF analysis.

From the GIRF analysis, we have recognized that the negative predictability power is increased during periods of crisis, when the Interest Rate Transmission Channel is working, whereas the Expectation Theory is not valid. In order to further investigate this issue, we have estimated the policy reaction function of Turkey, covering the estimation period of the LSTVAR analysis. We have found out that the policy reaction function has multiple structural breaks that can prevent sound conclusions on the above issues. Hence Estrella's [1] theoretical prediction, "empirical relationships are not structural, and alternative monetary policy regimes could lead to very different outcomes", cannot be proven by the Turkish data [47].

Appendix A

GIRFs are obtained by making bootstrapping. Hence we have to construct their confidence band again by designing a bootstrapping instead of Monte-Carlo design. For GIRF, we handle 2 hundred of impulse responses in order to get one specific histories' impulse response. We have obtained these 2 hundred impulse responses in order to average the effect of intermediate shocks. For their confidence band again we design a bootstrap and handle the confidence band from this computation. For this purpose, we run 1,000 impulse responses which are the averaged from 200 impulse responses and create 10% confidence band for every histories impulse response.

Appendix B

Figure 5

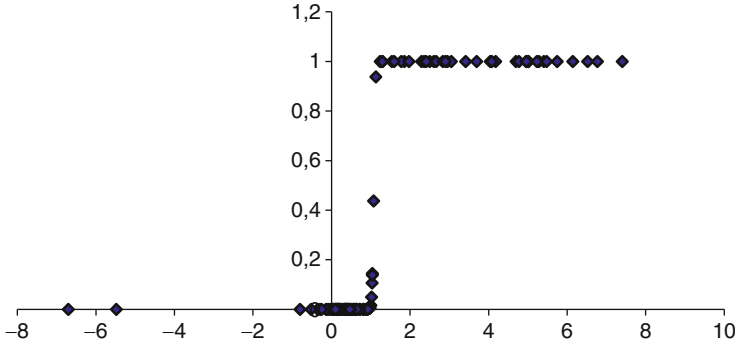


Fig. 5 Transition function

References

1. Estrella, A. (1997), "Why Do Interest Rates Predict Macro Outcomes? A Unified Theory of Inflation, Output, Interest, and Policy," *Federal Reserve Bank of New York Research Paper* No. 9717.
2. Harvey, C. R. (1988) "The Real Term Structure and Consumption Growth," *Journal of Financial Economics*, 22, 305–333.
3. Chen, N. (1991), "Financial Investment Opportunities and the Macroeconomy," *Journal of Finance* 46, 529–554.
4. Estrella, A., ve G. A. Hardouvelis. (1991), "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance*, 46, 555–576.
5. Plosser, C., ve K. G. Rounwenhorst. (1994), "International Term Structure and Real Economic Growth," *Journal of Monetary Economics*, 33, 133–155.
6. Haubrich, J. G. and A. M. Dombrosky. (1996), "Predicting Real Growth Using the Yield Curve," *Federal Reserve Bank of Cleveland, Economic Review*, 32, 26–35.
7. Dotsey, M. (1998), "The Predictive Content of Interest Rate Term Spread for Future Economic Growth," *Federal Reserve Bank of Richmond Economic Quarterly*, 84, Summer 31–51.
8. Moody, A., ve M.P. Taylor (2000), "The High Yield Spread as a Predictor of real economic Activity: Evidence of a Financial Accelerator for United States," *IMF Staff Paper*.
9. Hamilton, J. D. and D. H. Kim. (2002), "A Re-examination of the Predictability of Economic Activity Using the Spread," *Journal of Money, Credit, and Banking*, 34, 2–10.
10. Venetis, I. A., I. Paya and D. A. Peel (2003), "Reexamination of the Predictability of Economic Activity Using the Yield Spread: A nonlinear approach" *International Review of Economics and Finance*, 2,
11. Mishkin, F. (1990a), "What Does the Term Structure Tell Us about Future Inflation?" *Journal of Monetary Economics*, 25, 77–95.

12. Mishkin, F. (1990b), "The Information in the Longer Maturity Term Structure About Inflation." *Quarterly Journal of Economics*, 55, 815–828.
13. Mishkin, F. (1991), "A Multi-Country Study of the Information in the Shorter Term Structure About Future Inflation." *Journal of International Money and Finance*, 10, 2–22.
14. Frankel, J. A. ve C. S. Lown (1994), "An Indicator of Future Inflation Extracted from the Steepness of the Interest Rate Yield Curve Along its Entire Length." *The Quarterly Journal of Economics*, 517–530.
15. Estrella, A. and F. S. Mishkin. (1995), "The Term Structure of Interest Rates and Its Role in Monetary Policy for the European Central Bank." *National Bureau of Economic Research Working Paper*, 5279.
16. Telatar, E., F. Telatar and R.A. Ratti (2003), "On the Predictive Power of the Term Structure of Interest Rates for Future Inflation Changes in the Presence of Political Instability: The Turkish Economy." *Journal of Policy Modeling*, 25, 931–946.
17. Jardet, C., (2004), "Why Did the Term Structure of Interest Rates Lose its Predictive Power?" *Economic Modeling*, 21: 509–524.
18. Peel, D. A. ve C. Ioannidis (2003), "Empirical Evidence on the Relationship Between the Term Structure of Interest and Future Real Output Changes When There are Changes in Policy Regime." *Economic Letters*, 78, 147–152.
19. Lin, H. and T. Terasvirta (1994), "Testing the Constancy of Regression Parameters against Continuous Structure Change" *Journal of Econometrics*, 62, 211–228.
20. Kim K. A., ve P. Limpaphayom (1997), "The Effect of Economic Regimes on the Relation between the Term Structure and Real Activity in Japan" *Journal of Economics and Business*, 49, 379–392.
21. Şahinbeyoğlu, G. and C. Yalçın (2000), "The Term Structure of Interest Rates: Does it tell about inflation?" *TCMB Discussion Paper* 2000:2.
22. Akyıldız, K., (2003), "Getiri Farkı Ekonomik Aktivitenin Tahmininde Öncü Gösterge İşlevi Görebilir Mi? Türkiye Örneği" *Hazine Dergisi*, Sayı: 16, 1–20.
23. Fuhrer, J.C. ve G.R. Moore (1995) "Monetary Policy Trade offs and the Correlation Between nominal interest rates and real output", *American Economic Review* 85, 219–239.
24. Taylor, J. B. (1993), "Discretion Versus Policy Rules in Practice," *Carnegie-Rochester Conference Series On Public Policy* 39:195–214.
25. Newey, W. ve K. West (1987), "A Simple, Positive Definite, Heteroskedasticity and Autocorrelation Consistent Variance Matrix" *Econometrica*, 55, 703–708.
26. van Dijk D. (1999), *STR Models Extensions and Outlier Robust Inference*. Tinbergen Institute Research Series no. 200.
27. Luukkonen, R., P. Saikkonen and T. Terasvirta (1988), Testing Linearity against STAR Models." *Biometrika*, 75, 491–499.
28. Terasvirta, T. (1994), "Specification, Estimation and Evaluation of STAR Models." *Journal of American Statistical Association*, 89, 208–218.
29. White, H. and Domowitz, I. (1984) "Nonlinear regression with dependent observations", *Econometrica*, 52, pp. 143–161
30. Pötscher, B.M. and Prucha, I.V. (1997) *Dynamic Nonlinear Econometric Models – Asymptotic Theory*, Berlin, Springer-Verlag
31. Anderson, M. H., and Yahid, F. (1998), "Testing multiple equation systems for common nonlinear component", *Journal of Econometrics*, Volume 84, Issue 1, May 1998, Pages 1–36
32. Eitrheim, Q. ve T. Terasvirta (1996), "Testing the Adequacy of STAR Models." *Journal of Econometrics*, 74, 59–76.
33. Koop, G., M.H. Pesaran and S.M. Porter (1996) "Impulse response analysis in nonlinear multivariate models", *Journal of Econometrics*, 74, 119–147.
34. Cook, T. and Hahn, T., (1989), "The effect of changes in the federal funds rate target on market interest rates in the 1970's" *Journal of Monetary Economics*, 19, 31–72.
35. Romer, D. (1996), *Advanced Macroeconomics*, New York: McGraw-Hill.
36. T. Ellingsen and U. Söderström "Why are Long Rates Sensitive to Monetary Policy?" Riksbank Working Paper No:5

37. Mishkin, F. (1995), "Symposium on the Monetary Transmission Mechanism." *Journal of Economic Perspective*, 9, 3–10.
38. Fama, E. F. (1975), "Short-term Interest Rates as a Predictors of Inflation." *The American Economic Review*, 65, 269–282.
39. Barro R.J. and Gordon D.B.(1981) "Rules, Discretion and Reputation in a Model of Monetary Policy." *Journal Of Monetary Economics*, 12:101–121
40. Lucas R.E. J. (1972) "Expectations and the Neutrality of Money." *Journal of Economics Theory*, 4: 103–124.
41. Chen, N. (1991), "Financial Investment Opportunities and the Macroeconomy," *Journal of Finance* 46, 529–554.
42. Berüment H. and Malatyalı K., (2000), "The implicit reaction function of th Central Bank of Repuclic of Turkey", *Applied Economics Letters*, 7, 425–430.
43. Kalkan, M., Kipici, A. and Peker, A., (1997), "Monetary policy and leading indicators of inflation in Turkey", *Irvin Fisher Committee Bulletin*, 1.
44. Clarida, R., J. Gali ve M. Gertler, (2000), "Monetary Policy Rules and Macroeconomic stability: Evidence and Some Theory", *Quarterly Journal of Economics*, 115, 147–180
45. Omay T. and Hasanov M. (2006) "Türkiye için Reaksiyon Fonksiyonunun Doğrusal Olmayan Modelle Tahmin Edilmesi" MPRA Paper 20154.
46. Hasanov, M. and Omay, T. (2008), "Monetary Policy Rules in Practice: Re-examining the Case for Turkey. *Physica A Statistical Mechanics and Its Applications*. 387(17), p. 4309–4318
47. Omay T. (2006) Türkiye'de Fazin Vade Yapısı ile Reel Ekonomik Aktivite Arasındaki İlişki. İktisadi Araştırmalar Vakfı Yayınları.

Licensing in an International Competition with Differentiated Goods

Fernanda A. Ferreira

1 Introduction

The effect of entry on social welfare has been studied by Collie [1], Cordella [2], and Klemperer [8]. In a closed economy, Klemperer [8] shows that entry reduces social welfare if the cost of the entrant is sufficiently higher than that of the incumbent (see also Lahiri and Ono [9]). Collie [1] examines this issue in an open economy and shows that entry of a foreign firm reduces domestic welfare unless the cost of the foreign firm is sufficiently lower than that of the incumbent. Cordella [2] also considers this issue in an open economy, showing the effects of the number of firms. Technological difference is an important reason for cost differences between firms, which may encourage them to share their technological information through licensing. Faul-Oller and Sandonis [3] show that higher welfare under entry is more likely in the presence of licensing by the technologically efficient incumbent compared with no licensing. Their results suggest that entry always increases welfare if there is licensing with output royalty but licensing with fixed fee only increases the likelihood of higher welfare under entry rather than eliminating the possibility of lower welfare under entry. While they have considered the situation of a closed economy, Mukherjee and Mukherjee [10] show the welfare implications of entry in the presence of technology licensing in an open economy. If either the entrant or the incumbent has a relatively superior technology,¹ it creates the possibility of technology licensing. Mukherjee and Mukherjee [10] show that if there is licensing with an upfront fixed fee, entry of a foreign firm not only increases

¹In our analysis, technology is defined by the marginal cost of production. Lower marginal cost implies better technology.

F.A. Ferreira (✉)

ESEIG - Polytechnic Institute of Porto and CMUP, Rua D. Sancho I, 981,

4480-876 Vila do Conde, Portugal

e-mail: fernandaamelia@eu.ipp.pt

domestic welfare when the foreign firm is sufficiently technologically superior to the domestic firm, it also increases domestic welfare if the foreign firm's technological inferiority is neither very small nor very large. However, if there is licensing with output royalty, foreign entry increases domestic welfare when the foreign firm is either sufficiently technologically superior or sufficiently technologically inferior to the domestic firm. So, the presence of technology licensing significantly affects the result of Collie [1], which considers the welfare effect of foreign entry without licensing. In this paper, we follow Mukherjee and Mukherjee [10], by doing a similar study for differentiated goods, in the case of a technologically superior entrant. Since differentiation of the goods reduces competition between firms, it increases the possibility of licensing. These results have important implications for competition policies and show that the policymakers need to be concerned about the technological efficiency of the foreign firm and the type of licensing contract (i.e. fixed-fee or royalty licensing) available to the firms.

We should mention that issues related to those of this paper have been studied by Ferreira [4, 5], Ferreira and Ferreira [6], and Ferreira et al. [7].

2 The Model and the Results

Consider a country, called the domestic country, in which there is a monopolist, called incumbent. To study the implications of entry, we will consider the following two situations in our analysis: (a) where the incumbent is a monopolist in the domestic country; and (b) where a foreign firm, called entrant, enters the market and competes with the incumbent. We suppose that the entrant is technologically superior to the incumbent. This situation may be consistent for trade between the developed countries with technological leapfrogging by the technologically lagging country.

2.1 The Case of a Monopoly

Let us first consider the situation where the incumbent is a monopolist in the domestic country, where the inverse market demand function is given by $p = a - q$, where p is the price, q the quantity in the market, and $a > 0$ the demand intercept. We assume that the incumbent can produce a good with the constant marginal production cost c_1 . For simplicity, we assume that there is no other production cost. The incumbent maximizes the following objective function to maximize its profit: $\max_q (a - q - c_1)q$. Optimal output of the incumbent is $q^* = (a - c_1)/2$ and its profit and consumer surplus are, respectively,

$$\pi^* = \frac{(a - c_1)^2}{4} \quad \text{and} \quad CS = \frac{(a - c_1)^2}{8}.$$

Therefore, in the monopoly case, welfare W^m of the domestic country, which is the summation of consumer surplus and profit of the incumbent, is given by

$$W^m = \frac{3(a - c_1)^2}{8}.$$

2.2 Entry Without Licensing

To show the implications of licensing, let us first consider entry of a foreign firm without licensing. Assume that there is a foreign firm, called entrant, who can produce the good at the constant marginal production cost c_2 less than the marginal cost c_1 of the domestic firm. So, we consider the following assumption:

Assumption 1. $c_2 < c_1$.

Assumption 1 can be interpreted as the foreign firm being technologically superior to the domestic firm. We also assume that there is no other production cost of the entrant, namely we assume that there is no transportation costs and/or tariff. In our stylized framework, we assume that the entrant exports its product to the domestic country and the firms (the incumbent and the entrant) compete like Cournot duopolists with differentiated products. The inverse demands are given by

$$p_i = a - q_i - \gamma q_j,$$

with $\alpha > 0$ and $0 < \gamma \leq 1$, where p_i is the price and q_i the amount produced of good i , for $i, j \in \{1, 2\}$. We note that the two products are substitutes, and, since $\gamma \leq 1$, “cross effects” are dominated by “own effect.” The value of γ expresses the degree of product differentiation. When γ is equal to one, the goods are homogeneous, and when γ tends to zero, we are close to independent goods. In what follows, we restrict the parameters of the model to satisfy the following assumption:

Assumption 2.

$$\max \left\{ 0, \frac{2c_1 - (2 - \gamma)a}{\gamma} \right\} < c_2 < \frac{(2 - \gamma)a + \gamma c_1}{2}.$$

This assumption requires that, in case of entry, both firms always produce positive outputs.

The incumbent and the entrant choose their outputs to maximize, respectively, their profits, i.e.,

$$\max_{q_1} (a - q_1 - \gamma q_2 - c_1)q_1 \quad \text{and} \quad \max_{q_2} (a - \gamma q_1 - q_2 - c_2)q_2,$$

where q_1 and q_2 are the outputs of the incumbent and the entrant, respectively. Optimal outputs of the incumbent and the entrant are, respectively,

$$q_1^* = \frac{(2 - \gamma)a - 2c_1 + \gamma c_2}{4 - \gamma^2}$$

and

$$q_2^* = \frac{(2 - \gamma)a - 2c_2 + \gamma c_1}{4 - \gamma^2}.$$

Profits of the incumbent, the entrant, and consumer surplus are, respectively,

$$\pi_1^* = \frac{((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2},$$

$$\pi_2^* = \frac{((2 - \gamma)a - 2c_2 + \gamma c_1)^2}{(4 - \gamma^2)^2}$$

and

$$CS = \frac{2(2 - \gamma)(2 + \gamma - \gamma^2)a^2 - 2(4 - 3\gamma^2 + \gamma^3)(c_1 + c_2)a}{2(4 - \gamma^2)^2} + \frac{(4 - 3\gamma^2)c_1^2 + (2\gamma^3c_1 + (4 - 3\gamma^2)c_2)c_2}{2(4 - \gamma^2)^2}.$$

So, in the case of entry without licensing, domestic welfare W_{nl}^e is given by

$$W_{nl}^e = \frac{2(2 - \gamma)a^2 - 2((3 - \gamma)c_1 + (1 - \gamma)c_2)a + 3c_1^2 - (2\gamma c_1 - c_2)c_2}{2(4 - \gamma^2)}. \quad (1)$$

Theorem 1. Assume that there is no possibility of technology licensing between the firms. If

$$c_1 > \frac{(3\gamma - 2)a}{3\gamma} \quad \text{and} \quad c_2 < \frac{3\gamma c_1 - (3\gamma - 2)a}{2},$$

entry increases domestic welfare.

It reduces welfare otherwise.

Proof. The inequality $W_{nl}^e > W^m$ is equivalent to $c_1 > (3\gamma - 2)a/(3\gamma)$ and $c_2 < (3\gamma c_1 - (3\gamma - 2)a)/2$. Noting that $(3\gamma c_1 - (3\gamma - 2)a)/2 < ((2 - \gamma)a + \gamma c_1)/2$, we get the result. \square

2.3 Entry with Licensing

Now, we are going to analyze the case of the entry under licensing. We consider two important types of licensing contracts (see, for example, Wang [11]): (a) fixed-fee licensing, where the licensor charges an upfront fixed fee for its technology; and (b) licensing with output royalty, where the licensor charges royalty per unit of output.

We consider the following game under entry. At stage 1, the technologically efficient entrant decides whether to license its technology to the incumbent, and the incumbent accepts the licensing contract, if it is not worse off under licensing compared with no licensing. At stage 2, the firms compete *à la* Cournot.

2.3.1 Fixed-Fee Licensing

We have seen above that profits of the incumbent and the entrant under no licensing are, respectively,

$$\pi_{1, \text{nl}}^* = \frac{((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2}$$

and

$$\pi_{2, \text{nl}}^* = \frac{((2 - \gamma)a - 2c_2 + \gamma c_1)^2}{(4 - \gamma^2)^2}.$$

Now, consider the situation under licensing. If licensing occurs, both firms produce with c_2 , since the entrant charges an upfront fixed fee for its technology. Profits of the incumbent and the entrant are, respectively,

$$\frac{(a - c_2)^2}{(2 - \gamma)^2} - F \quad \text{and} \quad \frac{(a - c_2)^2}{(2 - \gamma)^2} + F,$$

where F is the optimal licensing fee charged by the entrant. So, licensing is profitable, if the following two conditions are satisfied for the incumbent and the entrant, respectively (with at least one strict inequality):

$$\frac{(a - c_2)^2}{(2 + \gamma)^2} - F \geq \frac{((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2}$$

and

$$\frac{(a - c_2)^2}{(2 + \gamma)^2} + F \geq \frac{((2 - \gamma)a - 2c_2 + \gamma c_1)^2}{(4 - \gamma^2)^2}.$$

Since the entrant gives a take-it-or-leave-it offer to the incumbent, the fixed fee makes the incumbent indifferent between licensing and no licensing, i.e.

$$F = \frac{(a - c_2)^2}{(2 + \gamma)^2} - \frac{((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2}.$$

So, licensing occurs if, and only if,

$$\frac{2(a - c_2)^2}{(2 + \gamma)^2} > \frac{((2 - \gamma)a + \gamma c_1 - 2c_2)^2 + ((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2},$$

which is satisfied for $c_2 > \frac{2(2 - \gamma)^2 a - (4 + \gamma^2)c_1}{\gamma^2 - 8\gamma + 4}$ with $\gamma \neq 4 - 2\sqrt{3}$. Thus, we have proved the following lemma.

Lemma 1. *Licensing occurs if, and only if,*

$$c_2 > \frac{2(2 - \gamma)^2 a - (4 + \gamma^2)c_1}{\gamma^2 - 8\gamma + 4} \text{ with } \gamma \neq 4 - 2\sqrt{3}.$$

Under fixed-fee licensing, the profit of the incumbent and consumer surplus are, respectively,

$$\pi_{1,lf}^* = \frac{((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2}$$

and

$$CS_{lf} = \frac{(1 + \gamma)(a - c_2)^2}{(2 + \gamma)^2},$$

if $c_2 > \frac{2(2 - \gamma)^2 a - (4 + \gamma^2)c_1}{\gamma^2 - 8\gamma + 4}$. So, domestic welfare W_{lf}^e under fixed-fee licensing is given by

$$W_{lf}^e = \frac{((2 - \gamma)a - 2c_1 + \gamma c_2)^2}{(4 - \gamma^2)^2} + \frac{(1 + \gamma)(a - c_2)^2}{(2 + \gamma)^2}.$$

Theorem 2. *Consider the possibility of fixed-fee licensing. Entry increases domestic welfare in the following situations:*

(i) For $0 < \gamma < 2/3$, if

$$c_2 < \frac{(8\gamma + (2 - \gamma)\sqrt{\Theta})c_1}{4(\gamma^3 - 2\gamma^2 + 4)} - \frac{(2 - \gamma)(\sqrt{\Theta} - 4(2 - \gamma^2))a}{4(\gamma^3 - 2\gamma^2 + 4)};$$

(ii) For $2/3 \leq \gamma \leq 1$, if

$$c_1 > \frac{(2-\gamma)(\sqrt{\Theta}-4(2-\gamma^2))a}{8\gamma+(2-\gamma)\sqrt{\Theta}}$$

and

$$c_2 < \frac{(8\gamma+(2-\gamma)\sqrt{\Theta})c_1}{4(\gamma^3-2\gamma^2+4)} - \frac{(2-\gamma)(\sqrt{\Theta}-4(2-\gamma^2))a}{4(\gamma^3-2\gamma^2+4)},$$

where $\Theta = 6\gamma^5 + 12\gamma^4 - 24\gamma^3 - 24\gamma^2 + 32\gamma + 32$.

It reduces welfare otherwise.

Proof. We have that $W_{lf}^e > W^m$ is equivalent to

$$\begin{aligned} c_2 < & \frac{(8\gamma+(2-\gamma)\sqrt{36\gamma^5+12\gamma^4-24\gamma^3-24\gamma^2+32\gamma+32})c_1}{4(\gamma^3-2\gamma^2+4)} \\ & - \frac{(2-\gamma)(\sqrt{36\gamma^5+12\gamma^4-24\gamma^3-24\gamma^2+32\gamma+32}-4(2-\gamma^2))a}{4(\gamma^3-2\gamma^2+4)} \end{aligned} \quad (2)$$

or

$$\begin{aligned} c_2 > & \frac{(8\gamma+(2-\gamma)\sqrt{36\gamma^5+12\gamma^4-24\gamma^3-24\gamma^2+32\gamma+32})c_1}{4(\gamma^3-2\gamma^2+4)} \\ & + \frac{(2-\gamma)(\sqrt{36\gamma^5+12\gamma^4-24\gamma^3-24\gamma^2+32\gamma+32}-4(2-\gamma^2))a}{4(\gamma^3-2\gamma^2+4)}. \end{aligned} \quad (3)$$

Since inequality (3) contradicts Assumption 1, we only have inequality (2). The result follows by noting that

(i) For $0 < \gamma \leq 2/3$, the expression

$$\begin{aligned} & \left(8\gamma+(2-\gamma)\sqrt{36\gamma^5+12\gamma^4-24\gamma^3-24\gamma^2+32\gamma+32}\right)c_1 \\ & - (2-\gamma)\left(\sqrt{36\gamma^5+12\gamma^4-24\gamma^3-24\gamma^2+32\gamma+32}-4(2-\gamma^2)\right)a \end{aligned}$$

is always positive.

(ii) For $2/3 \leq \gamma \leq 1$, we have that

$$\begin{aligned} & \left(8\gamma + (2 - \gamma) \sqrt{36\gamma^5 + 12\gamma^4 - 24\gamma^3 - 24\gamma^2 + 32\gamma + 32} \right) c_1 \\ & - (2 - \gamma) \left(\sqrt{36\gamma^5 + 12\gamma^4 - 24\gamma^3 - 24\gamma^2 + 32\gamma + 32} - 4(2 - \gamma^2) \right) a > 0 \end{aligned}$$

if, and only if,

$$c_1 > \frac{(2 - \gamma) \left(\sqrt{36\gamma^5 + 12\gamma^4 - 24\gamma^3 - 24\gamma^2 + 32\gamma + 32} - 4(2 - \gamma^2) \right) a}{8\gamma + (2 - \gamma) \sqrt{36\gamma^5 + 12\gamma^4 - 24\gamma^3 - 24\gamma^2 + 32\gamma + 32}}. \quad \square$$

2.3.2 Licensing with Output Royalty

Now, consider licensing with per-unit output royalty, where the entrant charges a per-unit output royalty for its technology. In that case of licensing, the effective marginal cost of the incumbent is $c_2 + r$, where r is the optimal per-unit output royalty. The optimal outputs of the incumbent and the entrant are, respectively,

$$q_{1,lr}^* = \frac{(2 - \gamma)a - (2 - \gamma)c_2 - 2r}{4 - \gamma^2}$$

and

$$q_{2,lr}^* = \frac{(2 - \gamma)a - (2 - \gamma)c_2 + \gamma r}{4 - \gamma^2}.$$

So, their profits are, respectively,

$$\pi_{1,lr}^* = \frac{((2 - \gamma)a - (2 - \gamma)c_2 - 2r)^2}{(4 - \gamma^2)^2}$$

and

$$\pi_{2,lr}^* = \frac{(2 - \gamma)^2(a - c_2)^2 + (2 - \gamma)(4 + 2\gamma - \gamma^2)(a - c_2)r - (8 - 3\gamma^2)r^2}{(4 - \gamma^2)^2}.$$

The entrant solves the problem

$$\max_r \pi_{2,lr}^* \quad (4)$$

subject to the constraint $r \leq c_1 - c_2$, to determine the optimal royalty rate.

Lemma 2. *The optimal output royalty r^* is as follows:*

(i) *If*

$$c_2 \geq \frac{2(8-3\gamma^2)c_1 - (2-\gamma)(4+2\gamma-\gamma^2)a}{8-2\gamma^2-\gamma^3},$$

then

$$r^* = c_1 - c_2;$$

(ii) *If*

$$c_2 < \frac{2(8-3\gamma^2)c_1 - (2-\gamma)(4+2\gamma-\gamma^2)a}{8-2\gamma^2-\gamma^3},$$

then

$$r^* = \frac{(2-\gamma)(4+2\gamma-\gamma^2)(a-c_2)}{2(8-3\gamma^2)}. \quad (5)$$

Proof. The solution of (4) is $r = (2-\gamma)(4+2\gamma-\gamma^2)(a-c_2)/(2(8-3\gamma^2))$. The result follows, since we have that $(2-\gamma)(4+2\gamma-\gamma^2)(a-c_2)/(2(8-3\gamma^2)) \geq c_1 - c_2$, for $c_2 \geq (2(8-3\gamma^2)c_1 - (2-\gamma)(4+2\gamma-\gamma^2)a)/(8-2\gamma^2-\gamma^3)$, and we note that, for all $0 < \gamma \leq 1$, we have that

$$\frac{2c_1 - (2-\gamma)a}{\gamma} \leq \frac{2(8-3\gamma^2)c_1 - (2-\gamma)(4+2\gamma-\gamma^2)a}{8-2\gamma^2-\gamma^3} \leq \frac{(2-\gamma)a + \gamma c_1}{2}. \quad \square$$

Theorem 3. *Suppose that there is royalty licensing.*

(i) *If*

$$c_2 \geq \frac{2(8-3\gamma^2)c_1 - (2-\gamma)(4+2\gamma-\gamma^2)a}{8-2\gamma^2-\gamma^3},$$

welfare implications of entry remain the same under licensing and no licensing.

(ii) *If*

$$c_2 < \frac{2(8-3\gamma^2)c_1 - (2-\gamma)(4+2\gamma-\gamma^2)a}{8-2\gamma^2-\gamma^3}, \quad (6)$$

then entry always increases welfare.

Proof. In case (i), by Lemma 2, the optimal per-unit output royalty is $r^* = c_1 - c_2$. Therefore, under entry, the incumbent's profit and consumer surplus remain the same under licensing and no licensing. Hence, domestic welfare under entry with royalty licensing is given by (1), and so we get the result.

In case (ii), by Lemma 2, the optimal per-unit output royalty is given by (5). Therefore, under entry, the incumbent's profit, consumer surplus and domestic welfare are, respectively, given by

$$\pi_{1,lr}^* = \frac{4(1-\gamma)^2(a-c_2)^2}{(8-3\gamma^2)^2},$$

$$CS_{lr} = \frac{(80-76\gamma^2+12\gamma^3+9\gamma^4)(a-c_2)^2}{8(8-3\gamma^2)^2}$$

and

$$W_{lr}^e = \frac{(112-64\gamma-44\gamma^2+12\gamma^3+9\gamma^4)(a-c_2)^2}{8(8-3\gamma^2)^2}.$$

The inequality $W_{lr}^e > W^m$ is equivalent to

$$c_2 < a - \frac{\sqrt{3}(8-3\gamma^2)(a-c_1)}{\sqrt{112-64\gamma-44\gamma^2+12\gamma^3+9\gamma^4}}$$

$$\vee c_2 < a + \frac{\sqrt{3}(8-3\gamma^2)(a-c_1)}{\sqrt{112-64\gamma-44\gamma^2+12\gamma^3+9\gamma^4}}.$$

This, together with condition (6), implies the result. \square

3 Conclusions

We showed the effects of foreign entry on social welfare in the presence of licensing, when the firms produce differentiated goods. We considered both fixed-fee licensing and royalty licensing, in the case where the foreign firm (the entrant) is technologically superior to the domestic firm (the incumbent). We found that the welfare implications of entry depend upon the type of licensing contract, and since differentiation of the goods reduces competition between firms, it increases the possibility of licensing.

Acknowledgments We are grateful to Alberto A. Pinto for a number of very fruitful and useful discussions on this work and for his friendship and encouragement. We thank ESEIG – Instituto Politécnico do Porto, Centro de Matemática da Universidade do Porto and the Programs POCTI and POCI by FCT and Ministério da Ciência, Tecnologia e do Ensino Superior for their financial support.

References

1. Collie D (1996) Gains and losses from unilateral free trade under oligopoly. *Recherches Economiques de Louvain* 62: 191–202
2. Cordella T (1993) Trade liberalization and oligopolistic industries: a welfare appraisal. *Recherches Economiques de Louvain* 59: 355–363

3. Faul-Oller R, Sandonis J (2003). To merge or to license: implications for competition policy. *International Journal of Industrial Organization* 21: 655–672
4. Ferreira FA (2009) Privatization and entry of a foreign firm with demand uncertainty. In: Simos TE et al. (eds) *Numerical Analysis and Applied Mathematics*, AIP Conference Proceedings 1168: 971–974, American Institute of Physics, New York
5. Ferreira FA (2008) Licensing in an international market. In: Simos TE et al. (eds) *Numerical Analysis and Applied Mathematics*, AIP Conference Proceedings 1048: 201–204, American Institute of Physics, New York
6. Ferreira FA, Ferreira F (2008) Welfare effects of entry into international markets with licensing. In: Todorov D (ed) *Applications of Mathematics in Engineering and Economics*, AIP Conference Proceedings 1067: 321–327, American Institute of Physics, New York
7. Ferreira FA, Ferreira F, Ferreira M, Pinto AA (2008) Quantity competition in a differentiated duopoly. In: Tenreiro JA et al. (eds) *Intelligent Engineering Systems and Computational Cybernetics* 365–374. Springer Science+Business Media B.V., New York
8. Klemperer P (1988) Welfare effects of entry into markets with switching costs. *Journal of Industrial Economics* 37: 159–165
9. Lahiri S, Ono T (1988) Helping minor firms reduces welfare. *Economic Journal* 98: 1199–1202
10. Mukherjee A, Mukherjee S (2005) Foreign competition with licensing. *The Manchester School* 73: 653–663
11. Wang XH (1988) Fee versus royalty licensing in a Cournot duopoly model. *Economics Letters* 60: 55–62

Multidimensional Scaling Analysis of Stock Market Indexes

Gonçalo M. Duarte, J. Tenreiro Machado, and Fernando B. Duarte

1 Introduction

Economical indexes measure the performance of segments of the stock market and are normally used to benchmark the performance of stock portfolios. This chapter proposes a descriptive method that analyzes possible correlations in international stock markets. The study of the correlation of international stock markets may have different motivations. Economic motivations to identify the main factors that affect the behavior of stock markets across different exchanges and countries. Statistical motivations to visualize correlations in order to suggest some potentially plausible parameter relations and restrictions. The understanding of such correlations would be helpful to the design good portfolios [1–4].

Bearing these ideas in mind, the outline of our chapter is as follows. In Sect. 2, we give the fundamentals of the multidimensional scaling (MDS) technique, which is the core of our method, and we discuss the details that are relevant for our specific application. In Sect. 3, we apply our method for daily data on 25 stock markets, including major American, Asian/Pacific, and European stock markets. In Sect. 4, we conclude the chapter with some final remarks and potential topics for further research.

2 Fundamental Concepts

MDS is a set of data analysis techniques for analysis of similarity or dissimilarity data. It is used to represent (dis)similarity data between objects by a variety of distance models.

J.T. Machado (✉)

Department of Electrical Engineering of Engineering, Porto, Portugal

e-mail: jtm@isep.ip.pt

The term similarity is used to indicate the degree of “likeness” between two objects, while dissimilarity indicates the degree of “unlikeness”. MDS represents a set of objects as points in a multidimensional space in such a way that the points corresponding to similar objects are located close together, while those corresponding to dissimilar objects are located far apart. The researcher then attempts to “make sense” of the derived object configuration by identifying meaningful regions and/or directions in the space.

In this article, we introduce the basic concepts and methods of MDS. We then discuss a variety of (dis)similarity measures and the kinds of techniques to be used. The main objective of MDS is to represent these dissimilarities as distances between points in a low dimensional space such that the distances correspond as closely as possible to the dissimilarities.

Let n be the number of different objects and let the dissimilarity for objects i and j be given by δ_{ij} . The coordinates are gathered in an $n \times p$ matrix \mathbf{X} , where p is the dimensionality of the solution to be specified in advance by the user. Therefore, row i from \mathbf{X} gives the coordinates for object i . Let d_{ij} be the Euclidean distance between rows i and j of \mathbf{X} defined as

$$d_{ij} = \sqrt{\sum_{s=1}^p (x_{is} - x_{js})^2} \quad (1)$$

that is, the length of the shortest line connecting points i and j . The objective of MDS is to find a matrix \mathbf{X} such that d_{ij} matches δ_{ij} as closely as possible. This objective can be formulated in a variety of ways but here we use the definition of raw-Stress σ^2 , that is,

$$\sigma^2 = \sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} (\delta_{ij} - d_{ij})^2 \quad (2)$$

by Kruskal [5] who was the first one to propose a formal measure for doing MDS. This measure is also referred to as the least-squares MDS model. Note that due to the symmetry of the dissimilarities and the distances, the summation only involves the pairs i, j where $i > j$. Here, w_{ij} is a user defined weight that must be nonnegative. The minimization of σ^2 is a complex problem. Therefore, MDS programs use iterative numerical algorithms to find a matrix \mathbf{X} for which σ^2 is a minimum. In addition to the raw stress measure there exist other measures for doing stress. One of them is normalized raw stress, which is simply raw stress divided by the sum of squared dissimilarities. The advantage of this measure over raw stress is that its value is independent of the scale and the number of dissimilarities. The second measure is Kruskal’s stress-1 which is equal to the square root of raw stress divided by the sum of squared distances. A third measure is Kruskal’s stress-2, which is similar to stress-1 except that the denominator is based on the variance of the distances instead of the sum of squares. Another measure that seems reasonably popular is called S-stress and it measures the sum of squared error between squared distances and squared dissimilarities.

Because Euclidean distances do not change under rotation, translation, and reflection, these operations may be freely applied to MDS solution without affecting the raw-stress. Many MDS programs use this indeterminacy to center the coordinates so that they sum to zero dimension wise. The freedom of rotation is often exploited to put the solution in so-called principal axis orientation. That is, the axis are rotated in such a way that the variance of \mathbf{X} is maximal along the first dimension, the second dimension is uncorrelated to the first and has again maximal variance, and so on.

In order to assess the quality of the MDS solution we can study the differences between the MDS solution and the data. One convenient way to do this is by inspecting the so-called Shepard diagram. A Shepard diagram shows both the transformation and the error. Let p_{ij} denote the proximity between objects i and j . Then, a Shepard diagram plots simultaneously the pairs (p_{ij}, d_{ij}) and (p_{ij}, δ_{ij}) . By connecting the (p_{ij}, δ_{ij}) points a line is obtained representing the relationship between the proximities and the disparities. The vertical distances between the (p_{ij}, δ_{ij}) points and (p_{ij}, d_{ij}) are equal to $\delta_{ij} - d_{ij}$, that is, they give the errors of representation for each pair of objects. Hence, the Shepard diagram can be used to inspect both the residuals of the MDS solution and the transformation. Outliers can be detected as well as possible systematic deviations.

Measuring and predicting human judgment is an extremely complex and problematic task. There have been many techniques developed to deal with such type of problems. These techniques fall under a generic category called Multidimensional Scaling (MDS). Generally speaking, MDS techniques develop spatial representations of psychological stimuli or other complex objects about which people make judgments (e.g., preference, relatedness), that is, they represent each object as a point in an n -dimensional space. What distinguishes MDS from other similar techniques (e.g., factor analysis) is that in MDS there are no preconceptions about which factors might drive each dimension. Therefore, the only data needed are measures for the similarity between each possible pair of objects under study. The result is the transformation of the data into similarity measures that can be represented by Euclidean distances in a space of unknown dimensions [6]. The greater the similarity of two objects, the closer they are in the n -dimensional space. After having the distances between all the objects, the MDS techniques analyze how well they can be fitted by spaces of different dimensions. The analysis is normally made by gradually increasing the number of dimensions until the quality of fit (measured for example by the correlation between the data and the distance) is little improved with the addition of a new dimension. In practice, a good result is normally reached well before the number of dimensions theoretically needed to a perfectly fit is reached (i.e., $N - 1$ dimensions for N objects) [7–10].

In the MDS method, a small distance between two points corresponds to a high correlation between two stock markets and a large distance corresponds to low or even negative correlation [11, 12]. A correlation of one should lead to zero distance between the points representing perfectly correlated stock markets. MDS tries to estimate the distances for all pairs of stock markets to match the correlations as close as possible. MDS may thus be seen as an exploratory technique without any distributional assumptions on the data. The distances between the points in the MDS

maps are generally not difficult to interpret and thus may be used to formulate more specific models or hypotheses. Also, the distance between two points should be interpreted as being the distance conditional on all the other distances. One possibility to obtain such an approximate solution is given by minimizing the stress function. The obtained representation of points is not unique in the sense that any rotation or translation of the points retains the distances [13].

3 Dynamics of Financial Indexes

In this section, we study numerically the 25 selected stock markets, including seven American, eleven European, and seven Asian/Pacific markets.

Our data consist of the n daily close values of $S = 25$ stock markets, listed in Table 1, from January 2, 2000, up to December 31, 2009, to be denoted as $x_i(t)$, $1 \leq t \leq n$, $i = 1, \dots, S$.

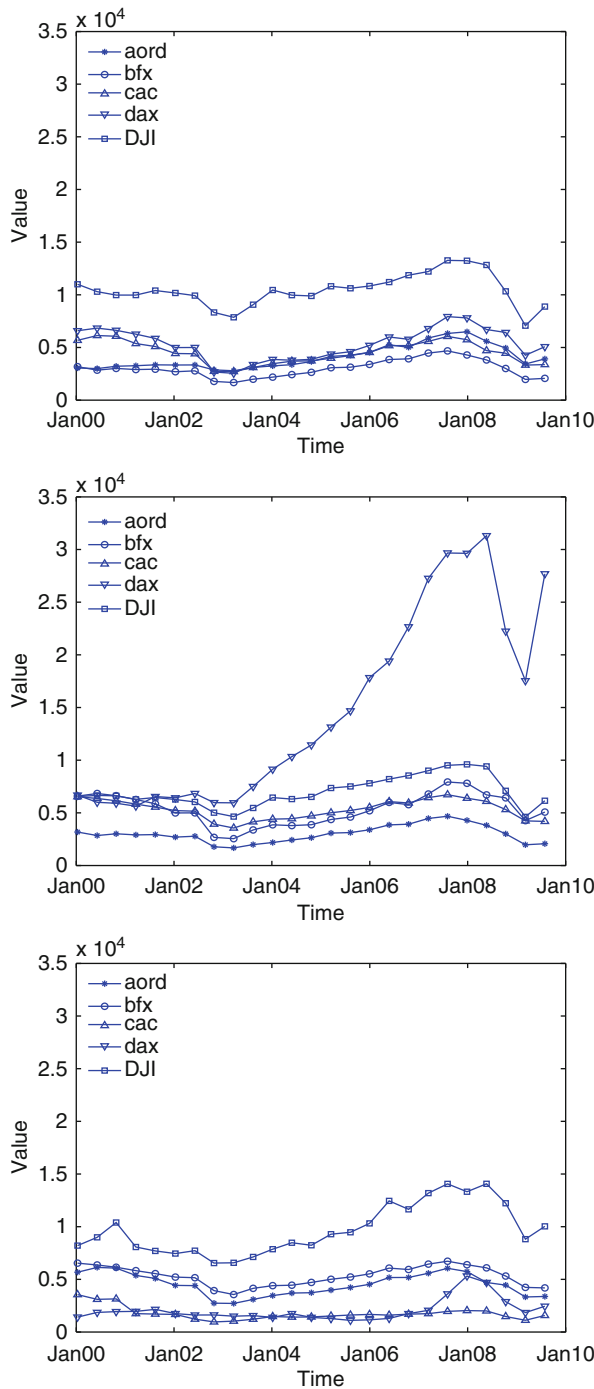
The data are obtained from data provided by Yahoo Finance web site [14] and [15], and they measure indexes in local currencies.

Figure 1 depicts the time evolution, of daily, closing price of the 25 stock markets versus year with the well-known noisy and “chaotic-like” characteristics.

Table 1 Twenty-five stock markets

i	Stock market index	Abbrev.	Country
1	Dutch Euronext Amsterdam	aex	Netherlands
2	Index of the Vienna Bourse	atx	Austria
3	EURONEXT BEL-20	bfm	Belgium
4	Bombay Stock Exchange Index	bse	India
5	So Paulo (Brazil) Stock	bvsp	Brazil
6	Budapest Stock Exchange	bux	Hungary
7	Dow Jones Industrial	dji	USA
8	Cotation Assistée en Continu	cac	France
9	Footsie	ftse	United Kingdom
10	Deutscher Aktien Index	dax	Germany
11	Standard & Poor's	sp500	USA
12	Toronto Stock Exchange	tsx	Canada
13	Stock Market Index in Hong Kong	hsi	Hong Kong
14	Iberia Index	ibex	Spain
15	Jakarta Stock Exchange	jkse	Indonesia
16	Stock Market Index of South Korea	ks11	South Korea
17	Italian Bourse	mibtel	Italy
18	Bolsa Mexicana de Valores	mxx	Mexico
19	Tokyo Stock Exchange	nikkei	Japan
20	NASDAQ	ndx	USA
21	New York Stock Exchange	nyse	USA
22	Stock Exchange of Portugal	psi20	Portugal
23	Shanghai Stock Exchange	ssec	China
24	Swiss Market Index	ssmi	Switzerland
25	Straits Times Index	sti	Singapore

Fig. 1 Time series for the 25 indexes from January 2000, up to December 2009



The section is organized into three subsections, the first adopts an analysis based on a Fourier transform (FT) “distance measure”, the second adopts an analysis based on the correlation of the time evolution, and the third adopts a metrics based on histogram distances.

3.1 MDS Analysis Based on Fourier Transform

We calculate the Fourier transform (FT) for each one of the indexes, leading to the values $\text{Re}[\mathcal{F}\{x_k(t)\}] + j \text{Im}[\mathcal{F}\{x_k(t)\}]$, $k = \{1, \dots, S\}$. It is adopted the “distance measure” defined as $(i, j = 1, \dots, S)$:

$$s(i, j) = \frac{\sum_{\Omega} (|\text{Re}_i - \text{Re}_j|^2 + |\text{Im}_i - \text{Im}_j|^2)}{\sum_{\Omega} (|\text{Re}_i|^2 + |\text{Im}_i|^2) \cdot \sum_{\Omega} (|\text{Re}_j|^2 + |\text{Im}_j|^2)}, \quad (3)$$

where Ω is the set of sampling frequencies for the FT calculation, and Re_i , Im_i , Re_j and Im_j are the values of real and imaginary parts of the FT. We get the matrix \mathbf{M} , where each cell represents the distance between a pair of FTs.

In order to reveal possible relationships between the signals the MDS [7] technique is used and several distance criteria are tested. The Sammon criterion [16, 17], that tries to optimize a cost function that describes how well the pairwise distances in a data set are preserved, revealed good results and is adopted in this work.

Figures 2 and 3 show the 2D and 3D MDS locus of index positioning in the perspective of the expressions (3), respectively. Figure 4 depicts the stress as function of the dimension of the representation space, revealing that, as usual, a high dimensional space would probably describe slightly better the “map” of the 25 indexes. However, the three dimensional representations were adopted because the graphical representation is easier to analyze while yielding a reasonable accuracy. Moreover, the resulting Sheppard plot, represented in Fig. 5 shows that a good distribution of points around the 45° line is obtained.

Analyzing Figs. 2 and 3 is visible that the MDS indexes forms an arc shapes in all the stock market indexes under analysis. How can we interpret these regularities? This is an interesting question for further research and a matter of decision of the stock market handlers.

3.2 MDS Analysis Based on Time Correlation

In this section, we apply the MDS method to the time correlation of all the stock markets.

For the 25 markets, we consider the time correlations between the daily close values. We first compute the correlations among the 25 stock markets obtained a $S \times S$ matrix and then apply MDS. In this representation, points represent the stock markets.

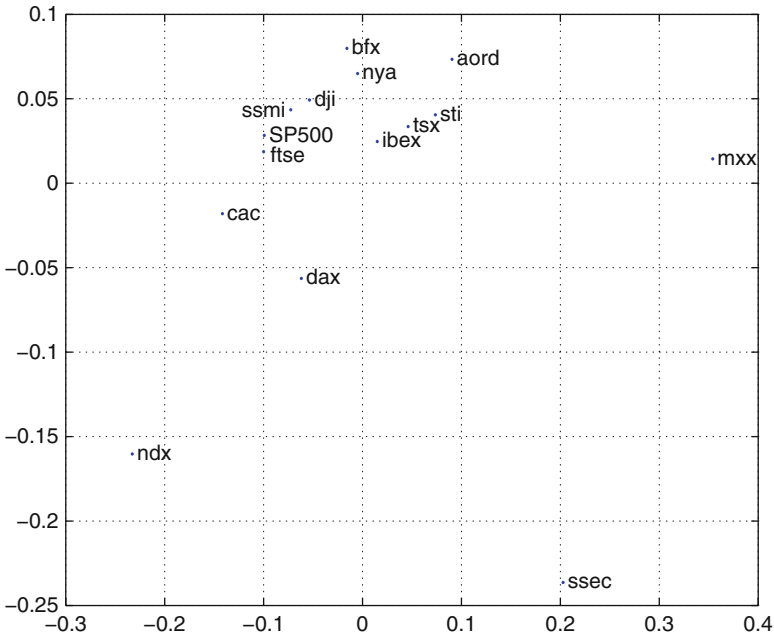


Fig. 2 Two-dimensional MDS graph for the 25 indexes based on the FT distance (3)

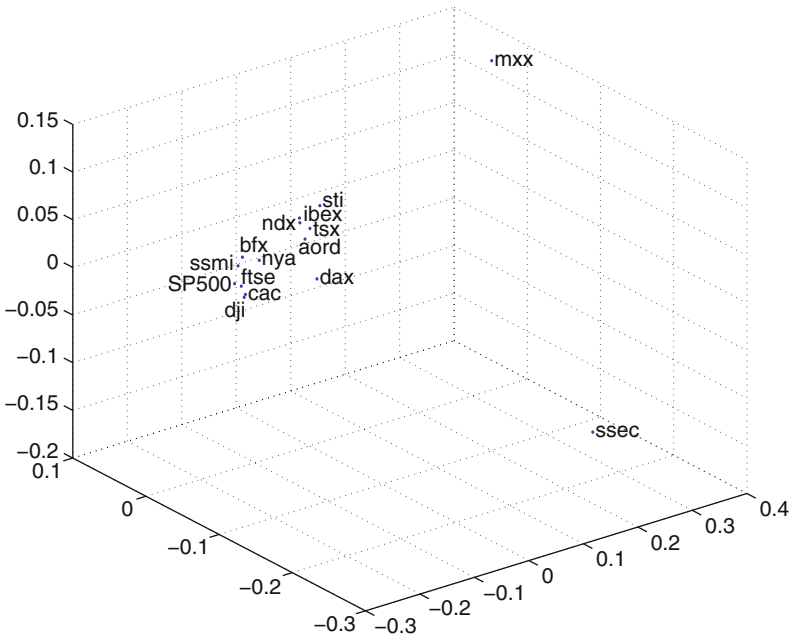


Fig. 3 Three-dimensional MDS graph for the 25 indexes based on the FT distance (3)

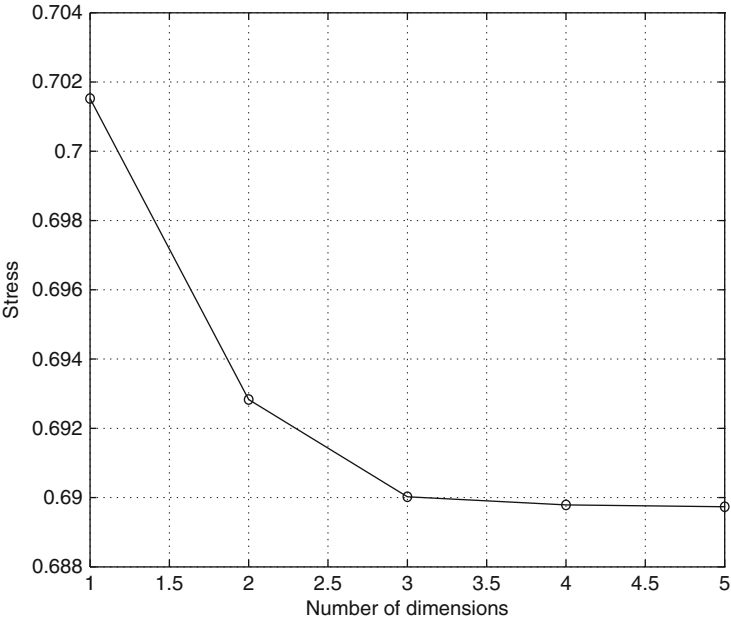


Fig. 4 Stress plot of the MDS representation, for the 25 indexes, vs number of dimensions, using the FT distance (3)

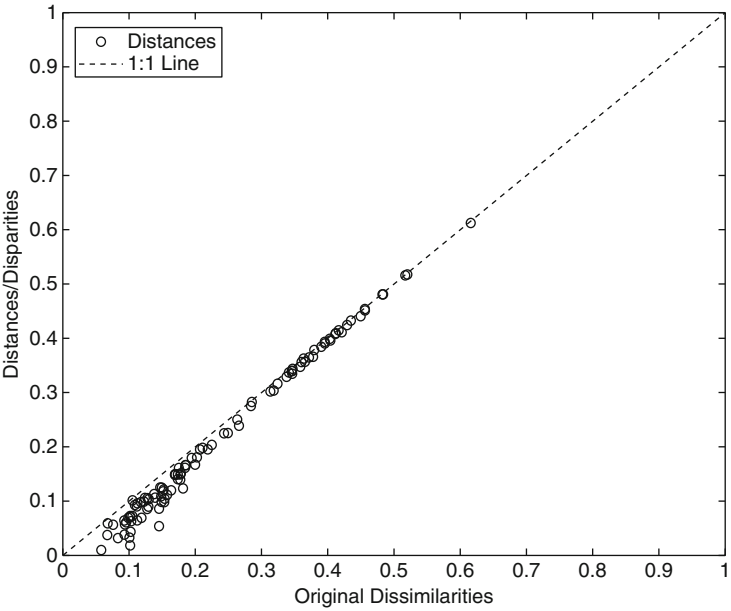


Fig. 5 Shepard plot of the three-dimensional MDS representation, for the 25 indexes, using the FT distance (3)

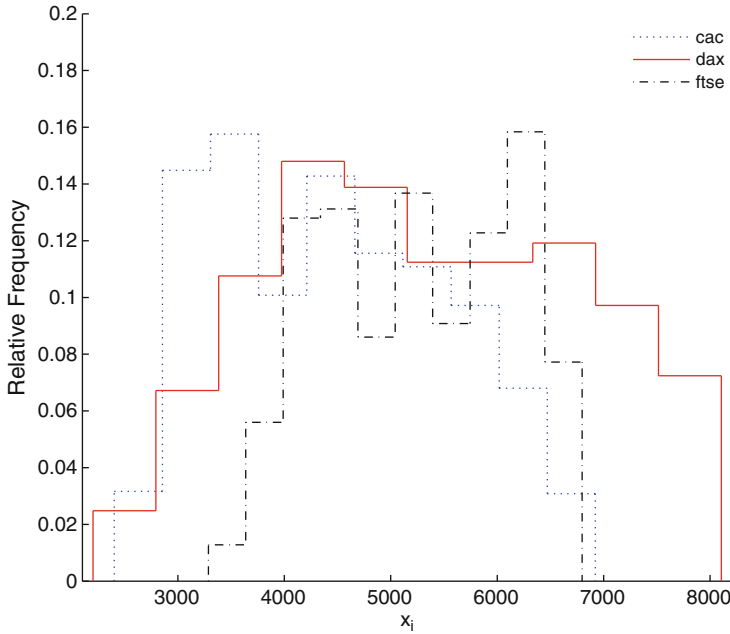


Fig. 6 Two-dimensional MDS graph for the 25 indexes using time correlation, according (4)

In order to reveal possible relationships between the market stocks index, the MDS technique is used. In this perspective, several MDS criteria are tested. The Sammon criterion revealed good results and is adopted in this work [18]. For this purpose, we calculate 25×25 matrix \mathbf{M} based on a correlation coefficient $c(i, j)$, that provides a measurement of the similarity between two indexes and is defined in (4). In matrix \mathbf{M} , each cell represents the time correlation between a pair of indexes, $i, j = 1, \dots, S$.

$$c(i, j) = \left(\frac{\frac{1}{n} \sum_{t=1}^n x_i(t) \cdot x_j(t)}{\sqrt{\frac{1}{n} \sum_{t=1}^n (x_i(t))^2 \cdot \frac{1}{n} \sum_{t=1}^n (x_j(t))^2}} \right)^2 \quad (4)$$

Figures 6 and 7 show the 2D and 3D locus of each index positioning in the perspective of expression (4), respectively. Figure 8 depicts the stress as function of the dimension of the representation space, revealing that a three-dimensional space describe a with reasonable accuracy the “map” of the 25 signal indexes. Moreover, the resulting Shepard plot, represented in Fig. 9, shows that a good distribution of points around the 45° line is obtained.

There are several empirical conclusions one can draw from the graphs in Figs. 6 and 7, and we will mention just a few here. We can clearly observe that there seem

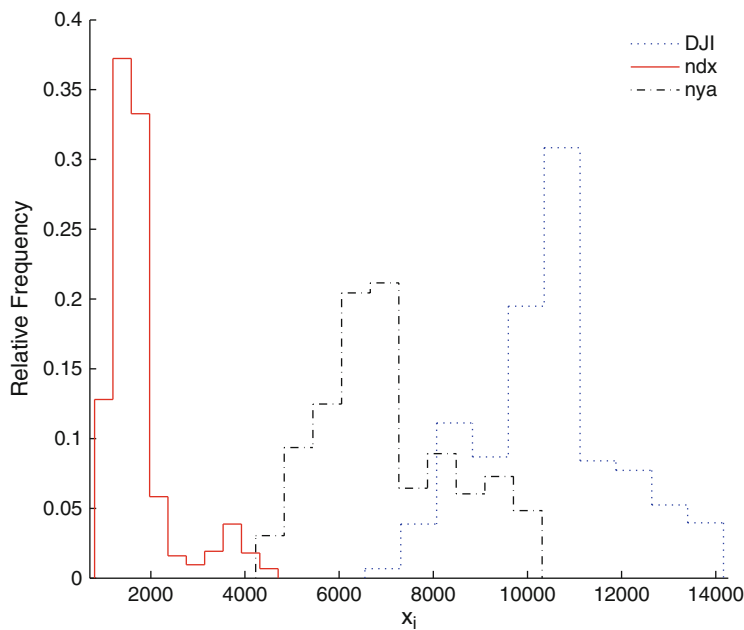


Fig. 7 Three-dimensional MDS graph for the 25 indexes using time correlation, according (4)

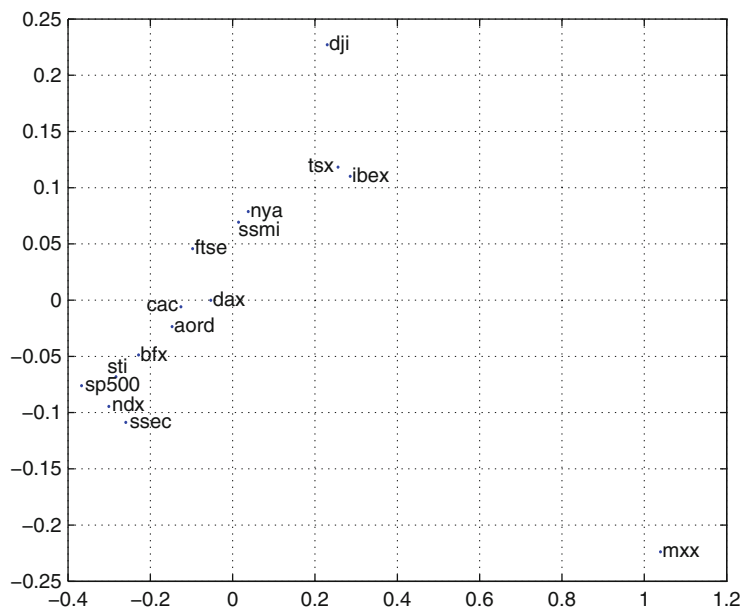


Fig. 8 Stress plot of MDS representation of the 25 indexes vs number of dimension using time correlation, according (4)

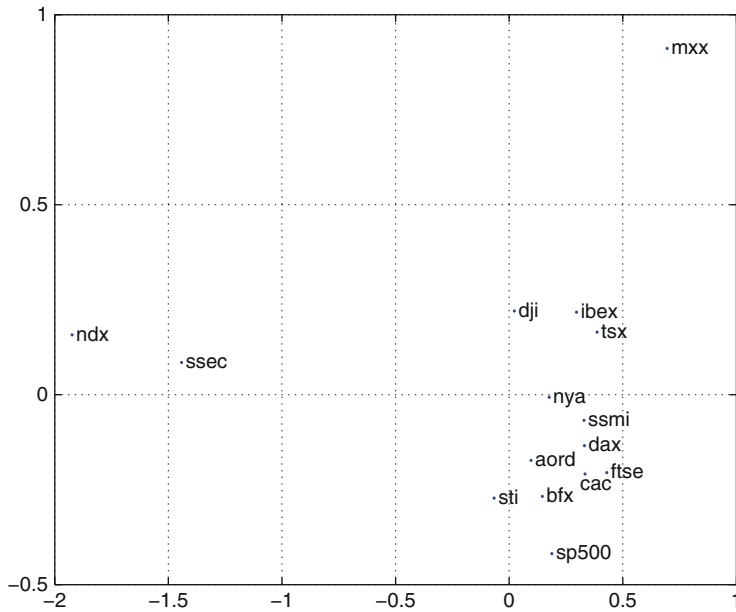


Fig. 9 Shepard plot for MDS with a three-dimensional representation of the 25 indexes using time correlation, according (4)

to emerge clusters, which show similar behavior [19]. Hence, there does not seem to be a single world market, but perhaps there are several important regional markets. This last observation would match with standard financial theory which tells us that higher (lower) volatility corresponds with higher (lower) returns. Indeed, if this would be the case, one would expect to see similar patterns over time across returns and volatility.

3.3 MDS Analysis Based on Histograms

For each of the 25 indexes, we draw the corresponding histogram of relative frequency and we calculate statistical descriptive parameters like the arithmetic mean (μ_i), the standard deviation (σ_i) and the Pearson's Kurtosis coefficient γ_i .

For all the 25 indexes, we calculate the "histogram's distance" [20, 21], d_h , using (5):

$$d_h(i, j) = \sqrt{\frac{(\mu_i - \mu_j)^2}{A} + \frac{(\sigma_i - \sigma_j)^2}{B} + \frac{(\gamma_i - \gamma_j)^2}{C}}, \quad (5)$$

where $i, j = 1, \dots, S$, $A = [\max(\mu_i - \mu_j)]^2$, $B = [\max(\sigma_i - \sigma_j)]^2$, $C = [\max(\gamma_i - \gamma_j)]^2$.

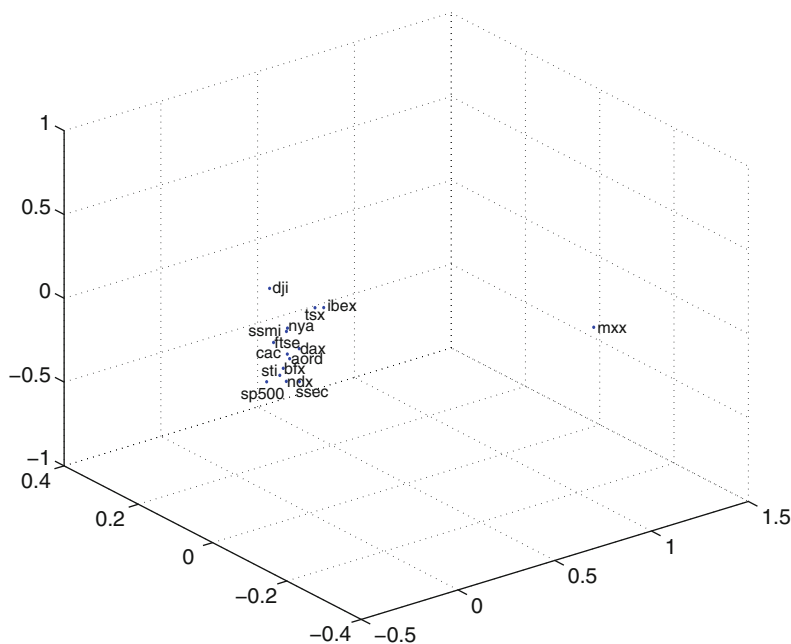


Fig. 10 Two-dimensional MDS graph for the 25 indexes using histogram's distance d_h

We get the matrix \mathbf{M} , where each cell represents the histogram's distance between a pair of indexes.

Figures 10 and 11 show, respectively, the 2D and 3D locus of each index positioning in the perspective of the expression (5), demonstrating differences between the corresponding MDS plots.

Figure 12 depicts the stress vs the dimension of the representation space, for d_h , revealing that a three-dimensional space describes with reasonable accuracy the “map” of the 25 indexes. The resulting Sheppard plots, represented in Fig. 13, shows that a good distribution of points around the 45° line is obtained for the indices [22].

Curiously in the chart corresponding to the MDS based on correlation (Fig. 6), we can see a parabolic shape with the S&P500 index at the vertex, and the NDX, and ATX at the top. However, in the chart corresponding to the MDS based on the histogram distances (Fig. 10) and in the chart corresponding to the MDS based on the FT distance (Fig. 2), such parabolic shape form cannot be found.

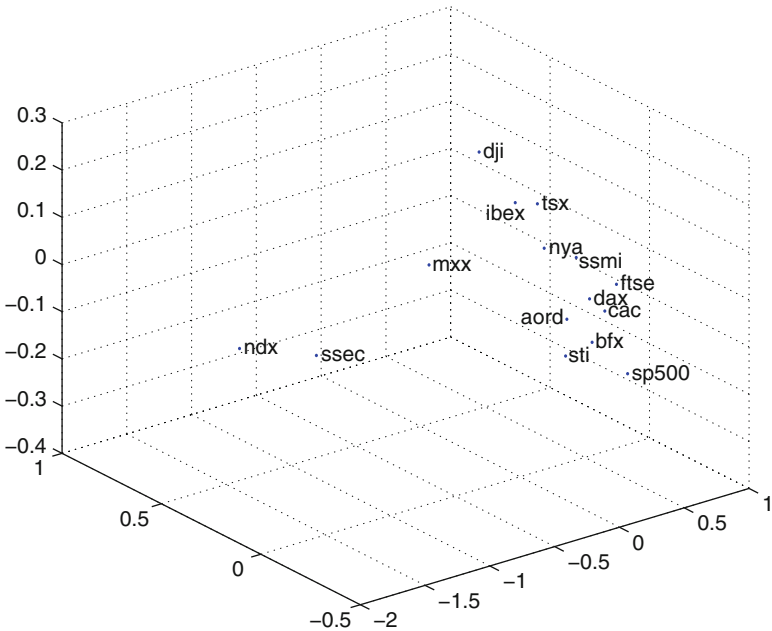


Fig. 11 Three-dimensional MDS graph for the 25 indexes using histogram’s distance d_h

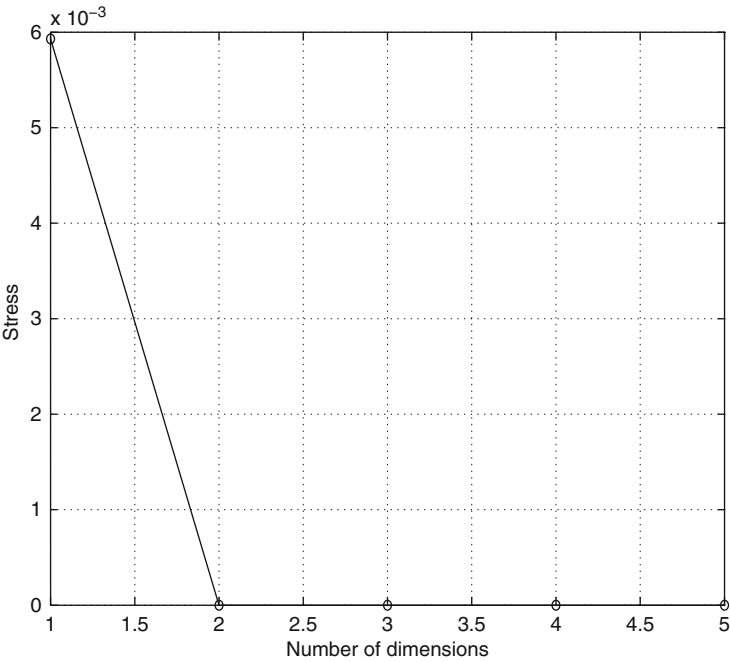


Fig. 12 Stress plot of MDS representation of the 25 indexes vs number of dimension using histogram’s distance d_h

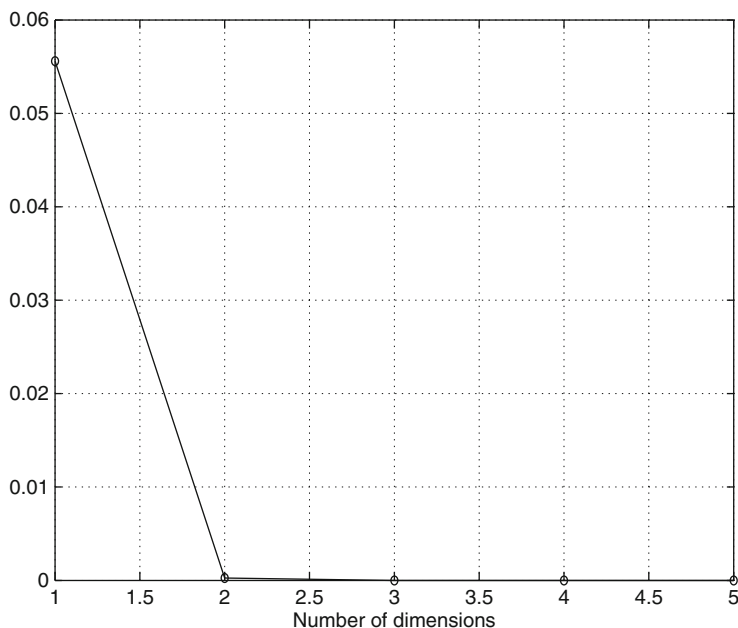


Fig. 13 Shepard plot for MDS with a three-dimensional representation of the 25 indexes, using histogram's distance d_h

4 Conclusion

In this chapter, we proposed simple graphical tools to visualize time-varying correlations between stock market behavior. We illustrated our MDS-based method daily close values of 25 stock markets. There are several issues relevant for further research. A first issue concerns applying our method to alternative data sets, with perhaps different sampling frequencies or returns and absolute returns, to see how informative the method can be in other cases. A second issue concerns taking the graphical evidence seriously and incorporating it in an econometric time series model to see if it can improve empirical specification strategies.

References

1. R. Nigmatullin, Communications in Nonlinear Science and Numerical Simulation **15**(3), 637 (2010)
2. R.V. Mendes, M.J. Oliveira, Economics: The Open-Access, Open-Assessment E-Journal **2**(2008-22) (2008). URL <http://www.economics-ejournal.org/economics/discussionpapers/2008-22>
3. R.V. Mendes, Nonlinear Dynamics **55**(4), 395 (2009)
4. R.B.A.L.S.H. Plerou V., Gopikrishnan P., Physica A **279**, 443 (2000)

5. J. Kruskal, Psychometrika **29**(1), 1 (1964). URL <http://dx.doi.org/10.1007/BF02289565>
6. I. Borg, P. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer, New York, 2005)
7. T. Cox, M. Cox, (Chapman & Hall/Crc, New York, 2001)
8. J. Kruskal, M. Wish, *Multidimensional Scaling* (Sage Publications, Inc., Newbury Park, CA, 1978)
9. J. Woelfel, G.A. Barnett, Quality and Quantity **16**(6), 469 (1982)
10. J.O. Ramsay, Psychometrika **45**(1), 139 (1980)
11. S. Nirenberg, P.E. Latham, Proc. National Academy of Sciences **100**(12), 7348 (2003)
12. F.B. Duarte, J.T. Machado, G.M. Duarte, Nonlinear Dynamics **61**(4), 691 (2010)
13. A. Buja, D. Swayne, M. Littman, N. Dean, H. Hofmann, L. Chen, Journal of Computational and Graphical Statistics **17**(85), 444 (2008)
14. <http://finance.yahoo.com>
15. <http://www.bse.hu>
16. J.W. Sammon, IEEE Trans. Comput. **18**(5), 401 (1969)
17. G.A.F. Seber, *Multivariate Observations* (J. Wiley & Sons, New York, 1986)
18. B. Ahrens, Hydrology and Earth System Sciences (10), 197 (2006)
19. J.T. Machado, G.M. Duarte, F.B. Duarte, Nonlinear Dynamics
20. F. Serratos, A. Sanroma, G. and Sanfeliu, Lecture Notes in Computer Science **4756**(1), 115 (2008)
21. B. Sierra, E. Lazkano, E. Jauregi, I. Irigoien, Decis. Support Syst. **48**(1), 180 (2009). DOI <http://dx.doi.org/10.1016/j.dss.2009.07.010>
22. J.T. Machado, F.B. Duarte, G.M. Duarte, in *Proceedings of 3rd Conference on Nonlinear Science and Complexity*, ed. by D. Baleanu (Cankaya University, 2010)

Index

A

- Adaptive remeshing
 - bang–bang example
 - delta- v , functional minimization, 75
 - iterations, 75, 76
 - reconfiguration computation, 75
 - spacecraft, 74
 - considerations, value v , 77–78
 - FEFF methodology
 - functional minimization, 71
 - goal, 70
 - spacecraft, equation, 70
 - formation flying, 69
 - low-thrust example
 - iterations and elements, TPF, 77
 - reconfiguration example, collision risk, 76
 - methodology, trajectories, 80
 - nonlinearities
 - bang–bang case, 78–79
 - low thrust case, 79–80
 - objective, 78
 - optimal mesh, 70
 - reconfigurations
 - elements number, new mesh, 73–74
 - Li and Bettess remeshing strategy, 73
 - objective, 71–72
 - procedure schema, 72
 - strategies, 72
 - spacecraft reconfigurations computation, 74
 - technologies, 69
- Adsorption, Ru-PC K314
 - cycle 1, 2 and 3
 - estimated and observed data, 104–105
 - parameter estimation, 103–104
 - start, end, max and length values, 101, 103
 - Langmuir model, 99, 101
 - observed data, 102, 103
 - QCM electrode and RH humidity sensor, 98
- Antarctica. *See* Deception Island
- Asteroids
 - global dynamics, 89–92
 - main belt and Trojans, 89, 90
- Asthma
 - children, 222, 223
 - described, 218
- Autonomous mobile waste sorter robot
 - geometric design
 - inverted slider crank arm mechanism, 141
 - structural members, 144
 - turning block mechanism, 141–143
 - wheel shaft, 143–144
 - properties
 - control mechanism, 139
 - degrees of freedom, 137, 138
 - hopper design, 138
 - inverted slider crank arm mechanism, 140
 - motor selection, 140–141
 - recycling, 135–136
 - sorting and working principles
 - hopper, 136–137
 - logic definition, 136
- Autoregressive conditional heteroscedasticity (ARCH), 275
- Axial next-nearest neighbor Ising (ANNNI) model
 - drawback, 117
 - T - p space, 118

B

Ballistic capture, 15

Bethe–Peierls theory, 118

Bicircular four-body model (BCRFBP)

Earth-to-halo transfers, 41–42

$N-1$ intervals, 46

spatial, 45

Bicircular model (BCP)

BCRFBP, 41–42

box covering, 64–65

coordinates changes

inertial and the synodical coordinates,
57

Kepler's law, 58

particle velocity and position, relations,
57

second integration, relations, 58

coupled CR3BP approximation

intersection, 64

transfer trajectories, design procedures,
63

CR3BP (*see* Circular restricted three-body
problem)

different integrations, 66

mass ratios, 56

n -body problem, 53

paper plan, 54

Poincaré section, 53

prevalence regions, 62–63

primaries positions, 55, 56

restricted problems, 54

spacecraft, 56

Sun–Earth–Moon system, 55

time-space system, 55–56

trajectories

box covering technique, 44

Earth escape stage, 42–43

halo capture stage, 43

trajectory example and error analysis, 66,
67

transfer trajectory and related errors ΔSE ,
 ΔEM , 66, 67

zero ΔV connections, 65–66

Boron, 127. *See also* CaX_6

Box covering

n -dimensional box, 64

Poincaré map, 65

Box covering technique, 44

C

CaX_6

band structure calculation, 131

computational method

cohesive energy, 128, 129

elastic constants, 128

hardness, 129

mechanical properties, 129

plane-wave calculation, 128

crystal structure, 127, 128

micro-hardness, 130

partial density calculation, 132

structural parameters, 130

Cayley tree, 118, 119

CF. *See* Cystic fibrosis

Chaotic behaviour

asteroids, resonant perturbations, 89

chaotic motion, 89

FLI, orbit, 87–88

strong motions, 93

Chaotic transport, 6, 94

Chow structural break test, 275

Circular bicircular four-body model
(CRTBP)

Earth-to-halo transfers

Jacobi integral of motion, 40–41

Lagrange points, 41

quantity Ω , 21

Circular restricted three-body problem
(CR3BP)

vs. BCP

coupled CR3BP approach, 59

vs. CR3BP_{EM}, 61–62

vs. CR3BP_{SE}, 59–60

features, 58–59

weak stability boundary theory, 59

description, 54

Jacobi constant and phase space, 55

massless particle motion, differential
equation, 54

potential Ω , five critical points, 55

Climate change, Turkey

center-based clustering, 173–175

data, 172

distance measures, 172–173

temperature variables, 172

Collision avoidance

important constraint, 71

trajectory discovery, spacecraft, 70

Correlation function

Dysthe model and nonlinear simulations,
152

Fourier phases, 151

laboratory wave modelling, 154–155

NLS equation simulations, BFI, 154

nonlinear simulations, BFI, 153, 154

steeper wave conditions, 152–153

Coupled three body problems approximation,
17, 63–64

- Cournot competition
 - firms profit, home and foreign, 259
 - output levels, equilibrium, 259, 263
- Cystic fibrosis (CF)
 - children, 222, 224
 - described, 218–219

D

- Deception Island
 - described, 193
 - hydrodynamic modeling, Port Foster
 - application, 196–199
 - approach, 195–196
 - results, 199–202
 - water mass vs. Bransfield Strait water mass, 194
- Desorption, Ru-PC K314
 - cycles, 101–102
 - Langmuir model, 99
 - nonlinear regression, 100
 - oscillation frequency shift, 100–101
 - parameter estimation, 102–105
 - QCM electrodes and RH humidity sensor, 98
- Differential method
 - Gabor filters approach, 185
 - Lucas and Kanade, 184–185
- Direct current motor
 - arm mechanism, 141
 - control mechanism, 139
 - robot movement, 138
- Duopoly market
 - Cournot, 257
 - optimal trade, demand uncertainty, 257
- Dynamical system theory, 15, 30, 58
- Dysthe model, 152

E

- Electron-electron interaction via bosons, 161
- Expectation maximization algorithm
 - distance norm, expression, 233
 - Gaussian membership functions, 233
 - partition matrix, 232
 - proposed algorithm, 234

F

- Fast Lyapunov indicator (FLI)
 - advantages, 85
 - behavior, orbits, 87–88
 - chaotic and regular orbits, 86
 - definition, 86
 - dynamical maps computation, 88, 90–92

- LCI, 85–86
- MLE, 85
- stability, extrasolar planets, 84
- Feature selectivity
 - “bumps”, 207
 - model
 - mirror bump, spindle torus, 210
 - neurons, torus topology, 208, 216
 - nontrivial dynamics, 208–209
 - ring torus, 209
 - spindle torus, 209–210
 - ring torus equilibrium solutions
 - inhomogeneous input absence, 211–213
 - possible sets, 213–215
 - \emptyset -solution, 211
 - π -solution, 211
 - α -solutions, 211
 - tristability, 215
- Financial indexes
 - MDS analysis
 - Fourier transform (FT), 312
 - histograms, 315, 318–320
 - time correlation, 312, 315
 - stock markets, 310
 - time evolution, 310, 311
- Financial mathematics. *See* Nature-inspired optimization methods, financial mathematics
- Finite element method
 - formation reconfiguration, 70
 - procedure FEFF, 71
 - time interval partition, 71
- First-principles
 - CaC₆ compound, 127
 - CaX₆, 130
- Fixed-fee licensing
 - domestic welfare, 300–301
 - inequality, 301–302
 - occurrence, 300
 - profits
 - consumer surplus and incumbent, 300
 - incumbent and entrant, 299
- FLI. *See* Fast Lyapunov indicator
- Flushing time
 - computed, 199
 - deception bay, 200–202
 - defined, 196
- Forced oscillation technique (FOT)
 - device setup, 219
 - recorded airway pressure signal, 220
- Forecasting
 - applications, fuzzy logic, 235–239
 - expectation maximization algorithm, 232–234

Forecasting (*Cont.*)

fuzzy

- clustering algorithm, 231–232
- C-Means algorithm, 232
- sets theory, 231

objectives, 239

validation

- indices, 234
- performance error, 235

Formation reconfiguration

- FEFF methodology, 70
- optimization problem, 70

FOT. *See* Forced oscillation technique

Fourier transform (FT), MDS analysis

- “distance measure”, 312
- stress and Shepard plot, 314
- two and three-dimensional graphs, 312, 313

Fractal dynamics, respiratory system

- breathing signal, 225
- exponential decay, volume, 224
- fractional order mathematical models, 217
- lung function tests, 217
- materials and methods
 - FOT, 219–220
 - fractal dimension and power law model, 221–222
 - patient database, 218–219
 - PV curve, 221
- PV relationship and plots, 225–226
- resistance and compliance, 223–224
- results
 - fitted power-law models and plots, 222, 224, 225
 - healthy, asthma and CF children, 222–224
 - PV loop, healthy child, 222

Fractional-order system, 217

Freak waves

- description, 147
- events, wave coherence
 - NLS equation, 148
 - numerical simulations, modulationally unstable, 149–150
 - phase evidence, 151–155

Friction velocity, 108, 112

Fuzzy clustering, 231–232

Fuzzy k-means

- clustering graph, 174–175
- description, 174

Fuzzy logic applications

- clustering
 - map, 236
 - results, 236, 237

- validity indices, PC and PE, 236, 237
- values, centres, 236, 237

data, 235

development, model, 235

model simulation

- membership functions, 237, 238
- predicted, learning phase, 236, 238
- rules, 237, 238

testing model

- rules results, 237, 239
- simulated values plot, 237, 239

G

Generalized impulse response function (GIRF), 270

Global dynamics, 84, 89

Graphite intercalation compounds (GICs), 127

H

Halo orbits

- Earth-escaping trajectories, 40
- EM, 39

Hamiltonian systems

- autonomous, 89
- and canonical variables, 88
- KAM theorem, 88

Hardness

- calculation, Simunek’s method, 129
- micro-hardness, 130

Hydrodynamic modeling, Port Foster

- application
 - model setup, 196–197
 - tidal and current measurements, 197–199
 - tidal flushing, 199
- approach
 - equations, 195
 - residence times and water exchange, 195–196

Bransfield Strait tides, 193

Deception Island, 193, 194

results

- flushing time, deception bay, 200–202
- harmonic analysis, 198, 199
- tidal elevation data, 200, 201
- tidal velocities, 199–200
- total sea water volume, 200
- water mass vs. Bransfield Strait water mass, 194

I

- Images processing, rainy cloud
 - data bank, 180–181
 - identification and tracking, 181–183
 - motion estimation, optical flow, 183–185

Inflation

- predictability, stability analysis, 288
- term structure, interest rates (*see* Term structure, interest rates)

Information visualization, 307

Invariant manifold

- PCR3B libration points, 4, 5
- stable and unstable, 4
- Sun–Jupiter PCR3BP, 10–11
- Sun–planet PC3BPs, 5

Invariant manifolds

- mission requirements class, 59
- Poincaré maps, 53

Inverted slider crank arm mechanism

- geometric design, 140
- kinematic modelling, 142
- position analysis, 141

K

k-means

- clustering graph, 174–175
- SSE vs. number of clusters, 173–174
- steps, 173

L

Lamb shift

- Fröhlich fraction, 160–162
- Hamiltonian interaction, 159–160
- three-dimensional space
 - Darwin term, 165
 - free state, 162–163
 - radial wave function, 164–165
 - Schrödinger equations, 163–164
 - spin-orbit coupling, 164
- two-dimensional space, 165–169

Langmuir model, parameter identification

- adsorption and desorption data, 101
- data analysis, 101–103
- experimental, 98–99
- nonlinear regression, 100
- oscillation frequency shift, QCM, 100
- parameter estimation, 102–105
- surface adsorption kinetics and frequency shift, 99

Law of the wake, 107, 113–115

Law of the wall

- expression, 108

law of the wake, 111

Libration points

- formation flying missions, 69
- orbit, 69, 70

Licensing, international competition

- domestic country, 296
- entry with licensing
 - fixed-fee licensing, 299–302
 - output royalty, 302–304

entry, without licensing

- domestic welfare, 298
- foreign firm, 297
- incumbent and entrant, 298
- profits, 298

monopoly

- optimal output, incumbent, 296
- welfare, domestic country, 297

output royalty, 296

social welfare, 295

technological difference, 295

Lobe dynamics, 6, 11–13

Logistic smooth transition vector

- autoregressive (LSTVAR) model

estimation results, 279

GIRF analysis, 270, 290

impulse response, 270–271

multiple regime, 280

Long-time natural transport

- fixed points and manifolds, 8–11

lobe dynamics, 11–13

Mars-to-Earth transport, 6

Poincaré map, Sun–Jupiter PCR3BP, 7–8

Low-energy Earth-to-halo transfers,

- Earth–Moon scenario

BCRTBP, 41–42

CRTBP (*see* Circular bicircular four-body model)

optimization trajectory

trajectory design

- box covering technique, 44

capture stage, 43

Earth escape stage, 42–43

LEO-to-halo, 42

- patched restricted three-body problem, 42

trajectory optimization

BCRFBP, 45–46

description, 44–45

dynamics, 46

optimization problem (OP), 46

single-impulse problem statement, 48

two-impulse problem statement,

- 47–48

Low-energy trajectories. *See also* Moon,
 low-energy and low-thrust transfer
 Earth–Moon transfers, 25
 four-body dynamics, 16
 Moon, transfer, 30–31
 two-impulse transfers, LLOs, 16–17
 Low-thrust propulsion
 attainable set defined, 23–24
 capture stage, Moon
 admissible final state domains, 27
 coupled RTBPs approximation, Earth
 escape stage, 26–27
 Moon ballistic capture stage
 EM model, 24–25
 impulsive capture solution, 26
 RTBPs approximation, global transfer,
 25–26
 two-impulse transfer, 25
 motion, controlled model, 22–23
 orbit expression, 23
 perturbation, 22
 tangential trajectory, 23
 LSTVAR model. *See* Logistic smooth
 transition vector autoregressive

M

Mechanical properties, 225
 Metaheuristics
 application, 249–251
 described, 244
 Meteorological radar, echoes
 fixed, 180, 181
 rainfall, 181, 182
 Microbalance. *See* Quartz crystal microbalance
 Modulational instability, stochastic wave fields
 BFI, 149–150
 laboratory measurements, 150
 nonlinear time, 149
 wave realizations, initial, 149
 Monetary policy. *See also* Term structure,
 interest rates
 central bank's policy preferences, 281–282
 Estrella's model, 273
 expansionary, 281
 real economic activity, 282
 tight, 282
 Moon, low-energy and low-thrust transfer
 design
 definition, 16–17
 eccentricity and periapsis/apoapsis, 17
 planar circular restricted three-body
 problem, 17–18
 Poincaré section, 16

earth escape stage
 Earth escape trajectory, 18, 19
 Earth–Moon, 19–20
 Lyapunov orbit, 18
 Moon–Perturbed Sun–Earth restricted
 three-body problem, 21–22
 parking orbit, 20
 Poincaré sections, RTBP, 18–19
 Sun–Earth PCRTBP computation, 22
 zero radial velocity, Earth, 20
 Earth–Venus transfers, 15
 Lyapunov orbits, 15
 propulsion (*see* Low-thrust propulsion)
 solution, optimized
 LLOs, 32–33
 low orbits trajectories, 33–36
 trajectory optimization
 BRFBP version, 28–29
 four-body potential, 28
 low-thrust problem statement, 31–32
 low-thrust propulsion and gravitational
 attractions, 28
 optimal control problem (OCP), 29–30
 Sun perturbation, 28
 two-impulse problem statement, 30–31

Motion estimation

differential method, 184–185
 optical flow equation, 183
 Multidimensional scaling (MDS) technique
 correlations, 309–310
 description, 309
 Euclidean distance, 308
 financial indexes
 Fourier transform (FT) analysis, 312
 histograms, 315, 318–320
 time correlation, 312, 315
 normalized raw and Kruskal's stress-1,
 308
 objective, 308
 principal axis orientation, 309
 Shepard diagram, 309
 similarity/dissimilarity data, 307–308

N

Natural transport
 description, 3
 long-time
 fixed points and manifolds, 8–11
 lobe dynamics, 11–13
 Mars-to-Earth transport, 6
 Poincaré map, Sun–Jupiter PCR3BP,
 7–8
 PCR3BP and coupled PCR3BPs, 3

- short-time
 - heliocentric distances, 5
 - inequality, 5–6
 - LPO and PC3BP, 4
 - Sun-Earth and Sun-Mars PCR3BP, 4–5
 - Nature-inspired optimization methods,
 - financial mathematics
 - algorithms, 248–249
 - ant colony, 248, 249
 - capital markets management and problems, 241
 - differential evolution, 246–247
 - genetic algorithm, 245, 246
 - genetic programming, 245
 - metaheuristics
 - application, 249–251
 - described, 244
 - multiobjective problems, 244–245
 - particle swarm
 - 2-D problem, 247–248
 - multiobjective, 248
 - portfolio optimization, 242–244
 - n*-body problem, 17–18, 53
 - Network topology
 - ring torus, 209
 - spindle torus, 209–210
 - torus, 208–209
 - Neural dynamics, 211
 - Nonlinear dynamics, 217
 - Numerical simulations, 149–150
- O**
- Optical flow. *See* Motion estimation
 - Optimal control theory, 29
 - Optimization
 - definition, 231
 - nature-inspired methods, 244–249
 - PC and PE values, 236
 - portfolio (*see* Portfolio optimization)
- P**
- Parameter identification. *See* Langmuir model, parameter identification
 - Partition coefficient (PC)
 - clustering, 236
 - overlapping amount measurement, 234
 - Partition entropy (PE), 234
 - Persen's theory, 108–109
 - Persistent states, 207
 - Phase space. *See* Planar elliptic restricted three-body problem
 - Planar elliptic restricted three-body problem (PERTBP)
 - autonomous Hamiltonian, 89
 - Delaunay variables, 88
 - dynamical maps computation, FLI, 88, 90–92
 - FLI (*see* Fast Lyapunov indicator)
 - invariant tori, diophantine frequencies, 88
 - mean motion resonances, low order, 93–94
 - resonances, 89
 - S (Sun) and J (Jupiter), 86–88
 - stability region, 92–93
 - strong chaotic motions, 93
 - PLC. *See* Programmable logic controller
 - Poincaré section
 - parallel shooting method, 8–9
 - Sun-Jupiter PCR3BP, 6–7
 - Portfolio optimization
 - Markowitz's approach, 242
 - metaheuristics application
 - constraints, 250
 - genetic programming, 251
 - researchers and publications, 249–250
 - transaction costs, 250
 - models
 - mean absolute deviation, 243
 - mean-variance, 242–243
 - semi-variance, 243
 - variance skewness, 243–244
 - Potts model, ternary and binary interactions
 - Cayley tree, 119
 - equation
 - partition function, 120–121
 - variables, 121–122
 - Kronecker symbol, 119
 - magnetization, average, 123–124
 - phase diagram morphology
 - periodic and aperiodic, 122–123
 - recursion relations, 122
 - Power-law dynamics, 221–222
 - Presen's wake function, 110–111
 - Pressure-volume (PV) loops. *See* Fractal dynamics, respiratory system
 - Programmable logic controller (PLC)
 - box, 144
 - control mechanism, 136, 139
 - Proximity maneuvers, 69–70
 - Pseudo phase plots, 217, 225
- Q**
- Quartz crystal. *See* Quartz crystal microbalance
 - Quartz crystal microbalance (QCM)
 - adsorption kinetics and frequency shift, 99
 - electrochemical (EQCM), 98
 - electrode preparation, 98–99

Quartz crystal microbalance (QCM) (*Cont.*)
 Langmuir model, 99
 ruthenium polypyridyl complex film,
 101–102

R

Rainy cloud

data bank, 180–181, 182
 identification and tracking
 classification, 183
 structure, 181–182
 motion estimation, 183–185

Real economic activity

Expectation Hypothesis, 287
 Indian, 272
 and inflation
 forecast horizon, 274
 LM test, ARCH, 275
 monthly industrial production index,
 274
 spread and future activity, relation,
 275–276
 stability issue, 275, 276
 monetary policy, 282
 short-term interest rate, 282–283
 and spread relationship, 272, 281
 stability analysis, 287
 structural break analysis, 290

Regions of prevalence

coordinates system, function definition, 62
 zero level set, ΔE , 62–63

Regression analysis, 97, 100

Residence time

residual circulation, 202
 water exchange, 195–196

Resonance

mean motion, low order, 89, 93–94
 overlap criterion, 93

Respiratory system. *See* Fractal dynamics,
 respiratory system

Restricted three-body problems

BCRTBP, 41–42
 CRTBP

Jacobi integral of motion, 40–41
 Lagrange points, 41

S (Sun) and J (Jupiter)

FLI behavior, orbits, 87–88
 heliocentric equations and infinitesimal
 body, 86–87
 truncation error, 87

Ring torus

equilibrium solutions
 inhomogeneous input absence, 211–213
 \emptyset -solution, 211

possible sets, 213–215
 π -solution, 211
 α -solutions, 211
 invariant weight function, 209
 Rogue waves. *See* Freak waves

S

Satellite optimal control, 76–77

Sensors

capacitive, 136, 139
 inductive, 136, 137–139
 X-ray transmission, 135

Shepard diagram, 309

Short-time natural transport

heliocentric distances, 5
 inequality, 5–6
 LPO and PC3BP, 4
 Sun-Earth and Sun-Mars PCR3BP, 4–5

Simulation

with adaptive remeshing
 bang–bang example, 74–76
 low-thrust example, 76–77
 value of v , 77–78
 Fuzzy logic to cost estimation, 236–237
 stochastic numerical, 149–150

Simunek's method, 129

Smooth transition vector auto-regression model (STVAR)

k -dimensional, 276–277
 LM-type statistic, 278
 multiple regime LSTVAR, 279, 280
 nonlinear optimization procedure, 279
 Taylor approximation, transition function,
 277

vector autoregressive model, 277

vector time series, 276

Solar system. *See* Natural transport

Sorting

glass and plastic wastes, 138
 hopper, 136–137
 sensor, logic, 136

Spindle torus, 209–210

Stackelberg leader

vs. Cournot firm, 262
 foreign firm, 267

Statistical techniques

MDS analysis, histograms
 “histogram's distance”, 315
 Shepard plot, 320
 stress plot, 319
 two and three-dimensional graph, 315,
 318
 subperiods, 287

- Stochastic nonlinear wave fields. *See*
Modulational instability, stochastic
wave fields
- Stock market daily values
close values, 310
closing price, 310, 311
time correlations, 312, 315
- Subgame perfect Nash equilibrium, 258, 259,
261, 263–265
- Suspended sediment, turbulent boundary layer
Coleman's data, 111–115
Coles' wake function, 109–110
law of the wall, 108
Persen's theory, 108–109
Presen's wake function, 110–111
regions, 107–108
Spalding's formulation, 108
- T**
- Tariff. *See* Uncertainty, revenue-maximizing
tariff model
- Term structure, interest rates
data, 274
equation system, 269
expectations theory and interest
transmission mechanism
dynamic correlation, 284, 285
expectation hypothesis, 282
Fisher hypothesis, 281, 284
GIRF analysis, 282
inflation predictability, stability
analysis, 287, 288
interest rate transmission mechanism,
282
inverted Fisher effect, 286
investigation, 283
linear model, 280
long- and short-term interest rates, 281
negative relationship, 283–284
“policy endogenous”, 281–282
recursive Chow test, 286
stability analysis, real economic activity
predictability, 287
- GIRF analysis
advantages, 291
bootstrapping, 291
- literature and model
Estrella's rational expectation's model,
273
inflation and spread, 272
parameter restrictions, equation system,
273
shorter maturities, 272
time-varying parameter model, 272
yield spread, 271
- LSTVAR model, 270
- monetary policy reaction function, stability
estimation, 289
inter-bank interest rate, 289
multiple structural breaks, 290
output gap coefficient, 289–290
parameters, 288
- policymakers, 269
predictability of inflation, 270
reduced form relationships, 269–270
regression, predictability
forecast horizon, 274
LM test, ARCH, 275
monthly industrial production index,
274
spread and future activity, relation,
275–276
stability issue, 275, 276
STVAR and linearity tests, 276–280
transition function, 292
- Terrestrial Planet Finder (TPF) model
formation, 79
swap, 79
- Three-body problem
circular restricted, 40–41, 54–55
planar circular restricted, 17–18
restricted, 86–88
- Tides
Bransfield Strait, 193
elevation data, 200, 201
flushing, 199
measurements
amplitude and phase errors, 199
harmonic analysis results, 197–198
- Time-varying correlation
Sammon criterion, 315
Sheppard plot, 317
stock markets, 310, 311
stress plot, 317
two and three-dimensional plots, 316
- Traditional impulse response function (TIRF),
270, 291
- Trajectory optimization
BCRFBP, 45–46
description, 44–45
dynamics, 46
optimization problem (OP), 46
single-impulse problem statement, 48
two-impulse problem statement
collision, inequality constrain,
47–48
NLP problem, 48

Trajectory optimization (*Cont.*)

variable vector, 47

velocity, 48

Turbulent boundary layer

Coleman's data

concentration functions, 113–114

correlation coefficient, 113

kinematic viscosity and

concentration, 111

locus of ξ , 114–115

theoretical profile and measured,

112–113

wall shear stress, 112

water temperature, 111

Coles' wake function

definition, 109

sediment concentration effect,

109–110

law of the wall, 108

Persen's theory, 108–109

Presen's wake function, 110–111

regions, 107–108

Spalding's formulation, 108

Turning block mechanism

after kinematic synthesis, 142

description, 141

side view, robot, 143

Two dimensional space

fine structure constant

Bessel function, 166

binding energy, 167

Coulomb interaction,

166–167

Lamb shift

Darwin term, 169

expression, 167–168

spin-orbit coupling, 168

U

Uncertainty, revenue-maximizing tariff model

benchmark model

comparisons, 262

firms' profits, 258

follower, home firm, 260–262

leader, home firm, 259–260

simultaneous decision, 258–259

two-stage game, 258

demand function

comparison, 266–267

follower, home firm, 265–266

leader, home firm, 264–265

simultaneous decision, 263–264

two-stage game, 263

W

Waste treatment. *See* Autonomous mobile

waste sorter robot

Wave coherence

evidence, phase

correlation estimator, 151–155

Fourier phases, correlation function,
151

modulational instability, 149–150

Weather forecasting. *See* Rainy cloud