BIRKHÄUSER

Mathematics
and Biosciences
in Interaction

Aspects of Mathematical Modelling

Applications in Science, Medicine, Economics and Management

Roger J. Hosking, Ezio Venturino (Editors)

BIRKHÄUSER

Mathematics and Biosciences in Interaction

Mathematics and Biosciences in Interaction is devoted to the publication of advanced textbooks, monographs, and multi-authored volumes on mathematical concepts in the biological sciences. It concentrates on truly interdisciplinary research presenting currently important biological fields and relevant methods being developed and refined in close relation to problems and results relevant for experimental bioscientists.

The series aims at publishing not only monographs by individual authors presenting their own results, but welcomes, in particular, volumes arising from collaborations, joint research programs or workshops. These can feature concepts and open problems as a result of such collaborative work, possibly illustrated with computer software providing statistical analyses, simulations or visualizations.

The envisaged readership includes researchers and advanced students in applied mathematics – numerical analysis as well as statistics, genetics, cell biology, neurobiology, bioinformatics, biophysics, bio(medical) engineering, biotechnology, evolution and behavioral sciences, theoretical biology, system theory.

Aspects of Mathematical Modelling

Applications in Science, Medicine, Economics and Management

Roger J. Hosking
Ezio Venturino

Editors

Birkhäuser
Basel · Boston · Berlin

Editors:

Roger J. Hosking
Universiti Brunei Darussalam
Dept. Mathematics
BE1410 Gadong
Brunei

Ezio Venturino
Università di Torino
Dipto. Matematica
Via Carlo Alberto 10
10123 Torino
Italy

Library of Congress Control Number: 2007935882

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

ISBN 978-3-7643-8590-3 Birkhäuser Verlag AG, Basel - Boston - Berlin

Part of Springer Science+Business Media
Printed on acid-free paper produced from chlorine-free pulp. TFC ∞
Cover illustration: See page 73. With the friendly permission of Q. J. A. Khan
Printed in Germany

ISBN 978-3-7643-8590-3
e-ISBN 978-3-7643-8591-0

9 8 7 6 5 4 3 2 1
www.birkhauser.ch

Preface

The construction of mathematical models is an essential scientific activity. Mathematics has long been associated with developments in the exact sciences and engineering, but more recently mathematical modelling has been used to investigate complex systems that arise in many other fields. Many chapters in this book discuss research where mathematics and the biosciences interact, and there are also chapters where mathematical modelling is applied elsewhere. The modern research topics discussed include ecology and environmental science, theoretical chemistry, medicine, phylogenetics and neural networks, economics and management.

This is an unusual book — not only because many of the authors are leaders in their respective fields, but also because the application of mathematical modelling and simulation is demonstrated in such an extensive array. The global reach of modern mathematical activity is more evident than usual too, with a geographical spread of contributions spanning four continents.

There are some invited reviews outlining current research directions in topics such as pattern formation in the first chapter by Malchow *et al.*, and in applications to medicine in the chapters by Quatember & Mayr and Motta *et al.*. There are also more targeted research papers on related topics, and in the various other disciplines represented. All of these contributions provide a background that may well inspire further research work on these subjects. The extensive relevant literature cited, particularly in some of the survey expository articles, is an important feature.

We expect that many established mathematical scientists will therefore find that this book provides useful information and further insights into interesting topics in their particular fields of expertise, or that it stimulates new work in less familiar areas. Moreover, we expect that many postgraduate students throughout the world will find that this book provides exceptionally helpful points of departure for their research endeavours.

September 2007

Roger J. Hosking
Ezio Venturino

Contents

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 1–26

Mathematical Models of Pattern Formation in Planktonic Predation-Diffusion Systems: A Review

Horst Malchow, Frank M. Hilker, Ivo Siekmann,
Sergei V. Petrovskii and Alexander B. Medvinsky

Abstract. Plankton form the basis of aquatic food webs. The mathematical modelling of plankton dynamics was initiated by fisheries in the early 20th century. Today, the significant role of plankton in the global carbon cycle and, hence, in climate control has been recognized. The main aim of modelling is to improve understanding of the functioning of food chains and webs and their dependence on internal and external conditions. Population-dynamical models have not only to account for growth and interactions but also for spatial processes like random or directed and joint or relative motion of species as well as the variability of the environment. Early attempts began with exponential growth, Lotka–Volterra type interactions and physico-chemical diffusion. These approaches have been continuously refined to more realistic descriptions of the development of natural populations. The aim of this paper is to give an introduction to the subject of equation-based modelling and the corresponding bibliography, based on and extending previous reviews [1–5]. The fascinating variety of temporal, spatial and spatio-temporal patterns in such systems and the governing mechanisms of their generation and further evolution are described and related to plankton dynamics.

Mathematics Subject Classification (2000). 35K55, 35K57, 35Q80, 37N25, 92B99, 60H15.

Keywords. Reaction-diffusion, Systems, Plankton dynamics, Pattern formation, Stability, Fronts, Waves, Epidemic spread, Bioinvasion.

1. Introduction

Ecosystems are complex adaptive systems. The exploration of pattern formation mechanisms in nonlinear complex systems is one of the central scientific problems. The development of the theory of self-organized temporal, spatial or functional structuring of nonlinear systems far from equilibrium has been one of the milestones of structure research [6, 7]. The occurrence of multiple steady states and transitions from one to another after critical fluctuations, the phenomena of excitability, oscillations, waves and, in general, the emergence of macroscopic order from microscopic interactions in various nonlinear nonequilibrium systems in nature and society has required and stimulated many theoretical and, where possible, experimental studies. Mathematical modelling has turned out to be one of the most useful methods to improve the understanding of such structure-generating mechanisms.

The aim of this paper is to give an introduction to and an overview of the mathematical modelling of biologically controlled temporal, spatial and spatiotemporal pattern formation in nonequilibrium plankton dynamics with a certain focus on conceptual models of prey-predator interactions of diffusing phytoplankton and zooplankton. Only deterministic and stochastic ordinary and partial differential equation-based models will be considered because they are the most appropriate tools for practical modelling. At first, the trophic level and importance of plankton in nature are briefly described. Then, the historical development of mathematical modelling of a growing number of properties of plankton dynamics in time and space is summarized. The models will be upgraded step by step, from the description of local processes like growth and interactions in uniform and variable environments to the consideration of environmental noise, as well as spatial processes like diffusion and advection.

2. Plankton and models of plankton dynamics

In the 17th century, the Dutch pioneer microscopist Anton van Leeuwenhoek was probably the first to see minute creatures in pond water which he called *animalcules* [8]. The German Victor Hensen, who organized Germany's first big oceanographic expedition in 1889 [9, 10], introduced the term *plankton* (derived from the Greek *planktos* = made to wander).

Phytoplankton are the plants in plankton. They drive all marine ecological communities and the life within them. Due to their photosynthetic growth, the world's phytoplankton generates half of the oxygen that mankind needs for maintaining life, and it absorbs half of the carbon dioxide that may be contributing to global warming. It is not only oxygen and carbon dioxide but also other substances and gases that are recycled by phytoplankton, *e.g.*, phosphorus, nitrogen and sulphur compounds [11–14]. Hence, phytoplankton is one of the main factors controlling the further development of the world's climate, a claim for which there is a vast supporting literature, cf. [15–17].

Zooplankton are the animals in plankton. In marine zooplankton both herbivores and predators occur; herbivores graze on phytoplankton and are eaten by zooplankton predators. Recently, reports have been published on the indirect role of zooplankton in climate control through grazing on the carbon dioxide absorbing phytoplankton and its transport to the deeper layers of the sea by sinking to depths where it can be deposited or distributed by higher predators [18, 19].

Together, phyto- and zooplankton form the basis for all food chains and webs in the sea. In turn, the abundance of plankton species is affected by a number of environmental factors such as water temperature, salinity, sunlight intensity, biogen availability etc. [20,21]. Temporal variability of the species composition may be caused by seasonal changes and trophical prey-predator interactions between phyto- and zooplankton.

Because of its apparent importance, the dynamics of plankton systems have been under continuous investigation during more than a hundred years. It should be noted that, practically from the very beginning, regular plankton studies have combined field observations, laboratory experiments and mathematical modelling. It was in the 19th century that fisheries stimulated the interest in plankton dynamics because strong positive correlations between zooplankton and fish abundance were found. The already mentioned German plankton expedition of 1889 was mainly motivated by fishery's interests. At the same time, fishery science began to develop. In the beginning of the 20th century, the first mathematical models were developed in order to understand and to predict fish stock dynamics and its correlations with biological and physical factors and human interventions, cf. [22–24] for details and further references.

2.1. Physical and biological scales

Many mechanisms of the spatio-temporal variability of natural plankton populations are not known yet. The distinct spatial heterogeneity of the horizontal plankton distribution (patchiness) is found in many field observations [25–30]. This phenomenon takes place on all scales, from centimeters to thousands of kilometers. The field data show that, on a spatial scale of dozens of kilometers and more, the plankton patchy spatial distribution is mainly controlled by the inhomogeneity of underlying hydrophysical fields like temperature, nutrients etc. [31, 32]. Pronounced physical patterns like thermoclines, upwelling, fronts and eddies often set the frame for biological processes. On a scale less than a hundred meters, plankton patchiness is controlled by turbulence [33, 34]. However, on an intermediate scale, roughly, from a hundred meters to a dozen kilometers, the features of the plankton heterogeneous spatial distribution have little correlation with the environment. Phytoplankton behaves decreasingly like a simple passive quantity distributed by turbulence [34–37]. Similarly, the spatial variability of zooplankton abundance differs essentially from the environmental variability on scales less than a few dozen kilometers [32]. It has been observed that the direction of motion of plankton patches does not always coincide with the direction of the water flow [38, 39]. This distinction is usually considered as evidence of the biology's "prevailing" against hydrodynamics on this scale [40–42]. Sommer [21, 43] has emphasized the importance of biological dynamics during phytoplankton blooms. Daly and Smith [44] concluded "...that biological processes may be more important at smaller scales where behaviour such as vertical migration and predation may control the plankton production, whereas physical processes may be more important at larger scales in structuring biological communities ...".

Physical and biological processes may differ significantly not only on spatial but also on temporal scales. Plankton pattern formation is essentially dependent on the interference of various physical (light, temperature, hydrodynamics) and biological factors (nutrient supply, predation), cf. [25, 31, 33]. O'Brien and Wroblewski [45] introduced a dimensionless parameter, containing the characteristic water speed and the maximum specific biological growth rate, to distinguish parameter regions of biological and physical dominance, cf. also [46, 47].

Also under conditions of relative physical uniformity, the temporal and spatio-temporal variability can be a consequence of the coupled nonlinear biological and chemical dynamics [48, 49].

2.2. Local models

2.2.1. Constant conditions in a uniform environment. The first mathematical models of population growth were already known in the 17th and 18th century, cf. Graunt [50], Euler [51], Malthus [52], Gompertz [53] and Verhulst [54]. Mathematical models of population interactions were first introduced by Lotka [55] and Volterra [56]. The contemporary mathematical modelling of phytoplankton productivity has its roots in the work by Fleming [57], Ivlev [58], Riley [59], Odum [60] and others. A review of the developments has been given by [61]. The most frequently used models have been collected by Behrenfeldt and Falkowski [62]. Denman [63] has discussed the problem of increasing complexity and parametrizing of planktonic ecosystem models.

The control of phytoplankton blooming by zooplankton grazing was first modelled by Fleming [57], using a single ordinary differential equation for the temporal dynamics of phytoplankton biomass. Other approaches have been the construction of data fitted functions [59,64] and the application of standard Lotka–Volterra equations to describe the prey-predator relation of phytoplankton and zooplankton [48, 65–68]. More realistic descriptions of zooplankton grazing with functional responses to phytoplankton abundance have been introduced by Ivlev [58] with a certain modification by Mayzaud and Poulet [69]. Holling-type response terms [70] which are also known from Monod or Michaelis–Menten saturation models of enzyme kinetics [71, 72] are just as much in use, cf. [40, 49, 73–83]. Observed temporal patterns are the well-known stable prey-predator oscillations, as well as the oscillatory or monotonic relaxation to one of the possible multiple steady states. Excitable models are of special interest because their long-lasting relaxation to their stable resting state after an above-threshold external perturbation, such as a sudden temperature increase or nutrient inflow, is suitable to model red or brown tides [77, 78, 84–86]. These models were first introduced in neurodynamics to describe the firing of neurons after supercritical stimuli [87–89].

Concerning the temporal variability of plankton species abundance, the limits of its predictability are of particular interest. At early stages, the development of mathematical models of marine ecosystems was driven by the idea that the more species were explicitly included into the model the higher would be its predictive ability. As a result, a number of multi-species models appeared allowing for a detailed structure of the food web of the community, cf. [90–92]. However, the actual predictive ability of this class of models is not very high and rarely exceeds a few weeks. Moreover, an increasing number of model agents may sometimes even worsen the properties of the model. This apparent paradox can be explained in terms of dynamical chaos [93]. It should be noted that there appear stronger and stronger indications in favour of the existence of deterministic chaos in population dynamics [94–98]. Chaotic population dynamics essentially changes the approach to the system predictability, cf. [94], and makes conceptual few-species models of as much use as multi-species ones. Moreover, few-species models can sometimes be even more instructive since they take into account only the principal features of the community functioning, cf. [76, 99–103].

2.2.2. External forcing in a variable environment. Aquatic food chains, like all natural systems, are subject to environmental variability. This ideally periodic external forcing appears rather naturally due to daily, seasonal or annual cycles of photosynthetically active radiation, temperature, nutrient availability *etc.* [104–107]. A number of forced models for parts or the complete food chain from nutrients, phytoplankton and zooplankton to planktivorous fish have been investigated and many different routes to chaotic dynamics have been demonstrated [108–116].

The effect of external hydrodynamical forcing on the appearance and stability of nonequilibrium spatio-temporal patterns has been studied [117], making use of the separation of the different time scales of biological and physical processes. A channel under tidal forcing served as a hydrodynamical model system with a relatively high detention time of matter. Examples were provided on different time scales: The simple physical transport and deformation of a spatially nonuniform initial plankton distribution as well as the biologically determined formation of a localized spatial maximum of phytoplankton biomass.

However, the environmental variability is not purely deterministic but also subject to random perturbations. Therefore, the description by ordinary differential equations is always an approximation. One has to consider stochastic differential equations to account for the noise [118, 119]. Noise-induced regime shifts between alternative stable states in ecosystems are as possible [120–125] as counter-intuitive phenomena like quasi-deterministic oscillations [126–128], noise-enhanced stability, noise-delayed extinction, stochastic resonance [129] or noise-induced spatial pattern formation [130–133]. An introduction to stochastic processes with applications to biology has been published by Allen [134].

2.3. Spatially extended models

Mathematical models of plankton population dynamics have not only to account for growth and interactions but also for spatial processes like random or directed and joint or relative motion of species, as well as the variability of the environment. It is the interplay of phytoplankton and zooplankton growth, interactions and transport that yields the whole variety of spatio-temporal population structures, in particular the phenomenon of plankton patchiness, cf. [25,135,136]. A well-studied stripy plankton pattern is due to the trapping of populations of sinking microorganisms in Langmuir circulation cells [137,138]. Other physically determined plankton distributions like steep density gradients due to local temperature differences, nutrient upwelling, turbulent mixing or internal waves have also been reported [139–145].

On a small spatial scale of some tens of centimetres, or under relative physical uniformity, differences in the "diffusive" mobility of individuals and the ability of locomotion might create finer spatial structures, *e.g.* due to bioconvection and gyrotaxis. Bioconvection patterns of micro-organisms emerge through an interplay of upswimming due to photo- or chemotaxis and sinking due to gravity. This phenomenon has been known since the 19th century [146,147], but the theory was not developed until 100 years later, cf. [148–153]. Gyrotaxis is an even more complicated mechanism of pattern formation. It is a directed locomotion resulting from the orientation of the cell's axis by compensating gravitational and viscous torques in a flow [154–157]. Till now not for plankton but for certain bacteria, the mechanism of diffusion-limited aggregation [158] has been proposed and experimentally proven for the spatial fingering of colonies [159,160].

The mathematical modelling of biologically controlled pattern formation requires the use of reaction-diffusion and, if applicable, perhaps advection equations, sometimes even of stochastic partial differential equations [86,161–164]

$$\frac{\partial X_i(\vec{r},t)}{\partial t} = f_i\left(\mathbf{X}\right) - \vec{\nabla}\cdot\left[\vec{v}_i X_i - \sum_{j=0}^{N} D_{ij}\vec{\nabla}X_j\right] + F_i\left(\mathbf{X},\vec{r},t\right)\,, \qquad i = 0,1,2,\ldots,N\,; \tag{2.1}$$

with appropriate initial and boundary conditions. $\mathbf{X} = \{X_i\ ;\ i = 0,1,2,...,N\}$ is the density vector of the N species at time t and position $\vec{r} = \{x,y,z\}$. $\mathbf{f} = \{f_i\ ;\ i = 0,1,2,\ldots,N\}$ is the vector of functions, describing the species growth, death and interactions. $\vec{v}_i = \{v_{ix}, v_{iy}, v_{iz}\}\ ;\ i = 0,1,2,\ldots,N$; is the velocity vector of the i-th species. It stands for both the common passive advection with a surrounding transport medium such as water or air and the potential individual capacity of active locomotion. $\vec{\nabla} = \{\partial/\partial x, \partial/\partial y, \partial/\partial z\}$ is the Nabla operator. $\mathbf{D} = \{D_{ij}\ ;\ i,j = 0,1,2,\ldots,N\}$ is the matrix of self- and cross-diffusion coefficients. The self-diffusion coefficients describe the species dispersal, usually down their own gradient. Cross-diffusion is the dispersal of a species along the gradient of the others. The latter coefficients allow the simple description of some behavioural strategies like neutrality, attraction or repulsion [135, 165–169]. Cross-diffusion is well-known from electrolyte solutions and from the theory of pattern formation in electro-diffusion systems [170–172]. $\mathbf{F} = \{F_i \; ; \; i = 0, 1, 2, \ldots, N\}$ is the vector of density-dependent external stochastic forces with certain noise characteristics in time and space, modelling environmental variability. Noise increases with population densities and, usually, a linear density dependence is chosen as an approximation. A good overview of the use of partial as well as stochastic partial differential equations in ecological modelling has been provided for instance by [130, 135, 136, 173–176] and [177], respectively.

The spectrum of spatial and spatio-temporal patterns includes regular and irregular oscillations, propagating fronts, target patterns and spiral waves, pulses as well as stationary spatial patterns.

Diffusion-driven instabilities. Since the classic paper by Turing [178] on the role of nonequilibrium reaction-diffusion patterns in biomorphogenesis, dissipative mechanisms of spontaneous spatial and spatio-temporal pattern formation in a homogeneous environment have been of continuous interest in theoretical biology and ecology. Turing showed that the nonlinear interaction of at least two agents with considerably different diffusion coefficients can give rise to spatial structure. A spatially uniform population distribution which is stable against spatially uniform perturbations (or in the local model without diffusion) can be driven to diffusive instability against spatially heterogeneous perturbations, *e.g.*, a population wave or local outbreak, for sufficient differences of the diffusivities. Segel and Jackson [65] were the first to apply Turing's idea to a problem in population dynamics: the dissipative instability in the prey-predator interaction of phytoplankton and herbivorous copepods with higher herbivore motility. Levin and Segel [48] suggested this scenario of spatial pattern formation as a possible origin of planktonic patchiness.

Local bistability, predator-prey limit-cycle oscillations, plankton front propagation and the generation and drift of planktonic Turing patches were found in a Rosenzweig–MacArthur model [179] for phytoplankton-zooplankton interactions that was extended by Scheffer [73], accounting for the effects of nutrients and planktivorous fish on alternative local equilibria of the plankton community [75, 180]. Planktivorous fish may control also the spatial plankton dynamics. The latter has been studied in detail, using a hybrid model of equation-based plankton and rule-based fish school dynamics [161, 181, 182].

Differential-flow-induced instabilities. Conditions for the emergence of three-dimensional spatial and spatio-temporal patterns after differential-flow-induced instabilities [183] of spatially uniform populations were derived [172, 184, 185] and illustrated by patterns in Scheffer's model [73]. Instabilities of the spatially uniform distribution can appear if phytoplankton and zooplankton move with different velocities but regardless of which one is faster. This mechanism of generating patchy patterns is more general than the Turing mechanism which depends on the already mentioned strong conditions on the difference of the diffusion coefficients. The latter does not exist for micro-organisms in meso- and large-scale aquatic systems where the turbulent diffusion is relevant. Thus, one can expect a wider range of applications of the differential-flow mechanism in population dynamics [1, 186–188].

Diffusive fronts and spatial critical sizes. Skellam [189] and Kierstead and Slobodkin [190] were perhaps the first to think of the critical size problem for plankton patches, presenting their model nowadays called KISS (combining the initials of their surnames) with the coupling of exponential growth and diffusion of a single population. Of course, their patches are unstable because this coupling leads to an explosive spatial spread of the initial patch of species with the same diffusive front speed [191] as the asymptotic speed of a logistically growing population [192, 193]. Recently, the KISS model has been additionally extended by advection and applied to species distribution in streams [194].

Populations with a strong Allee effect [195–202], i.e., when the existence of a minimum viable population size yields two stable population states – extinction and survival at its carrying capacity, show a spatial critical size as well [203–210]. Population patches greater than the critical size will survive, while the others will go extinct. However, bistability and the emergence of a critical spatial size do not necessarily require an Allee effect, also logistically growing preys with a parametrized predator of type II or III functional response can exibit two stable steady states and the related hysteresis loops, cf. [211, 212].

Spiral waves. Many of these structures were first known from oscillating chemical reactions, cf. [213], but have never been observed as biologically controlled structures in natural plankton populations. However, spirals have been seen in the ocean as rotary motions of plankton patches on a kilometer scale [39]. Furthermore, they have been found important in models of parasitoid-host systems [214]. For other motile microorganisms, travelling waves like targets or spirals have been found in the cellular slime mold *Dictyostelium discoideum* [215–226]. These amoebae are chemotactic species, i.e., they move actively up the gradient of a chemical attractant and aggregate. Chemotaxis is a kind of density-dependent cross-diffusion [227, 228] and it is an interesting open question whether there is preytaxis in plankton or not. However, there is some evidence of chemotaxis in certain phytoplankton species [229]. Bacteria like *Escherichia coli* or *Bacillus subtilis* also show a number of complex colony growth patterns [230, 231], different from the already mentioned diffusion-limited aggregation patterns. Their emergence requires as well cooperativity and active motion of the species which has also been modelled as density-dependent diffusion and predation [232, 233].

New routes to spatiotemporal chaos. An important point is that the spatial dimensions of the plankton community functioning provide also new routes to chaotic dynamics. The emergence of diffusion-induced spatio-temporal chaos has been found along a linear nutrient gradient [76]. Chaotic oscillations behind propagating diffusive fronts are found in a prey-predator model [234, 235]. Recently, it has been shown that the appearance of chaotic spatio-temporal oscillations in a prey-predator system is a somewhat more general phenomenon and need not be attributed to front propagation or to an inhomogeneity of environmental parameters [99, 100, 102, 236, 237].

Virus infections and invasions. Not so much is known about marine viruses and their role in aquatic ecosystems and the species that they infect [238, 239]. It is said [240] that viruses may control the oceans and that "infection may be the spice of planktonic life". Suttle et al. [241] have experimentally shown that the viral disease can infect bacteria and phytoplankton in coastal water. There is some evidence that viral infection might accelerate the termination of phytoplankton blooms [242, 243]. However, despite the increasing number of reports, the role of viral infection in the phytoplankton population is still far from understood.

Mathematical models of the dynamics of virally infected phytoplankton populations are rare as well; the already classical publication is by [244]. More recent work by the authors of this review can be found in [86, 163, 237, 245, 246]. They observed regular and strange periodic oscillations as well as invading infection waves in a phytoplankton-zooplankton system with Holling-type II and III grazing under lysogenic viral infection and frequency-dependent transmission. The latter is also called proportionate mixing or standard incidence [247–249]. Other authors consider mass-action type transmission and lytic infections [250–252]. All these models exist without explicit virus dynamics and have the generic dimensionless growth and interaction functions

$$f_1 = (b_1 - m_1)(1-P)X_1 - \frac{a^n P^{n-1}}{1+b^n P^n} X_1 X_3 - \lambda \frac{X_1 X_2}{P^j} \,, \tag{2.2}$$

$$f_2 = (b_2 - m_2)(1-P)X_2 - \frac{a^n P^{n-1}}{1+b^n P^n} X_2 X_3 + \lambda \frac{X_1 X_2}{P^j} - m_2^* X_2 \,, \tag{2.3}$$

$$f_3 = \frac{a^n P^n}{1+b^n P^n} X_3 - m_3^q X_3^q - \frac{p^m P^m}{1+s^m P^m} X_3 - \frac{g^k X_3^k}{1+h^k X_3^k} X_4 \,. \tag{2.4}$$

$P = X_1 + X_2$ is the total phytoplankton density of susceptibles X_1 and infected X_2, X_3 is the density of zooplankton, X_4 that of a not explicitly modelled higher predator, *e.g.*, planktivorous fish. (b_1, b_2) are birth rates, (m_1, m_2, m_3) mortality rates, m_2^* stands for the additional disease-induced mortality of the infected (virulence). λ is the transmission rate of the disease. a, b, p, s, g, h are parameters characterizing the functional responses of predators X_3 and X_4, respectively. Different settings of parameters and exponents describe various dynamics, *e.g.*,

$b = 0$, $n = 1$	Lotka–Volterra dynamics;
$\lambda > 0$, $j = 0, 1$	Mass-action type and frequency-dependent transmission of infection, respectively;
$b_2 = 0, 1$	lytic and lysogenic infections, respectively;
$a > 0$, $n = 2$	Excitability [77, 78, 85];

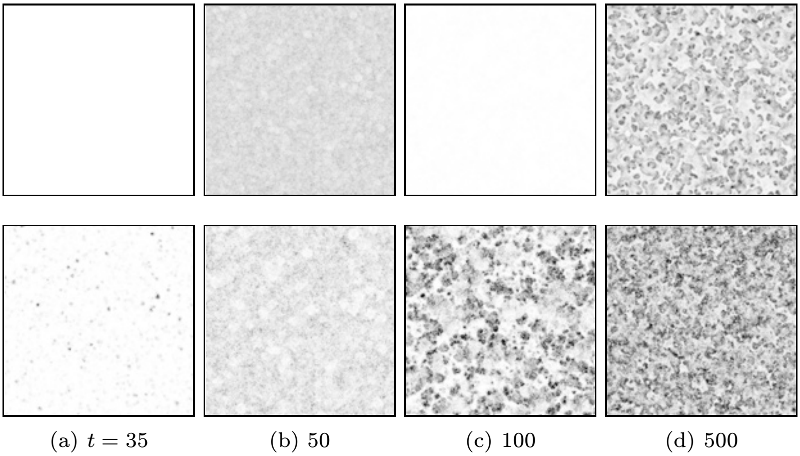
$1 < q \leq 2$	Intraspecific zooplankton competition [253];
$p > 0,\ m = 1, 2$	Increased zooplankton mortality through toxin-producing phytoplankton [254–256];
$g > 0,\ k = 1, 2$	Feeding of fish on zooplankton [73, 257].

Beretta and Kuang [258] introduced a local model with explicit viral dynamics, lytic infections and mass-action type of transmission but only susceptible and infected phytoplankton. It has been extended by Siekmann et al. [259] through consideration of Holling-type II grazing zooplankton, diffusion and multiplicative noise. The growth and interaction functions of the extended model read

$$
\begin{aligned}
f_0 &= -\lambda X_0 X_1 - m_0 X_0 + B m_2 X_2 \,, && (2.5)\\
f_1 &= -\lambda X_0 X_1 + (b_1 - m_1)(1 - P) X_1 - \frac{a}{1 + bP} X_1 X_3 \,, && (2.6)\\
f_2 &= +\lambda X_0 X_1 - m_2 X_2 - \frac{a}{1 + bP} X_2 X_3 \,, && (2.7)\\
f_3 &= \frac{aP}{1 + bP} X_3 - m_3 X_3 \,, && (2.8)
\end{aligned}
$$

where X_0 is the virus density and B the burst factor that stands for the number of virus particles that are set free during the lysis of an infected phytoplankton cell. Cross-diffusion is neglected, $D_{ij} \equiv 0 \ \forall\ i \neq j = 0, 1, 2, 3$, and equal (eddy) diffusion coefficients have been chosen, $D_{ii} = d \ \forall\ i = 0, 1, 2, 3$. The stochastic forces are

$$F_i(X_i, \vec{r}, t) = \omega X_i \xi_i(\vec{r}, t) \,;\ \ i = 0, 1, 2, 3 \,; \qquad (2.9)$$



(a) $t = 35$ (b) 50 (c) 100 (d) 500

FIGURE 1. Spatiotemporal dynamics of zooplankton X_3 in model (2.5–2.9) and with above parametrisation. Upper row: $\omega = 0.1$, lower row: $\omega = 0.2$. Spatially uniform initial conditions $X_0(\vec{r}, 0) = 0.1, X_1(\vec{r}, 0) = 0.5, X_2(\vec{r}, 0) = 0.6, X_3(\vec{r}, 0) = 0.1$. Neumann boundary conditions.

where $\xi_i(\vec{r}, t)$ is a spatiotemporal white Gaussian noise, i.e., a random Gaussian field with zero mean and delta correlation. ω is the constant noise intensity. The density dependence reflects the increase of noise with growing species numbers. In particular, such noise is originated by fluctuating mortalities. Furthermore, it accounts for the postulate of parenthood [260].

It has been proven that, besides trivial and semi-trivial local stationary solutions, all four populations may coexist only on a stable limit cycle [259]. For illustration of spatiotemporal pattern formation, model (2.5–2.9) is simulated with parameters that support coexistence:

$$\lambda = 1, B = 35, a = b = 5, b_1 = 1, m_0 = 1.1, m_1 = 0, m_2 = 1.07, m_3 = 0.2, d = 0.05.$$

The results are shown in Fig. 1. Without noise, stable spatially uniform oscillations would appear. The noise generates the spatial heterogeneity, the stronger the faster and the finer. Distinguished wavy structures appear in a uniform environment with initial heterogeneities.

3. Concluding remarks

This paper gave an introduction to the equation-based mathematical modelling of plankton dynamics in continuous time and space. The presented models produce a wide spectrum of real-world structures, such as steady-state multiplicity, regular and irregular oscillations, propagating fronts, target patterns and spiral waves, pulses as well as stationary spatial patterns, on various temporal as well as spatial scales and provide insight into the underlying mechanisms that can generate these patterns. Other modelling tools, such as integro-differential and difference equations, metapopulation models, cellular automata, individual-based models and further rule-based tools as well as combinations of different methods, also show promising results and need attention and development.

Acknowledgements

This work has been partially supported by the Deutsche Forschungsgemeinschaft, grant no. 436 RUS 113/821, by the Russian Foundation for Basic Research as well as by the VIGONI Programme of the German Academic Exchange Service, grant no. D/05/54280.

References

[1] H. Malchow, Nonequilibrium spatio-temporal patterns in models of nonlinear plankton dynamics, Freshwater Biology 45 (2000) 239–251.

[2] H. Malchow, S. V. Petrovskii, A. B. Medvinsky, Pattern formation in models of plankton dynamics. A synthesis, Oceanologica Acta 24 (5) (2001) 479–487.

[3] H. Malchow, S. V. Petrovskii, F. M. Hilker, Models of spatiotemporal pattern formation in plankton dynamics, Nova Acta Leopoldina NF 88 (332) (2003) 325–340.

[4] S. V. Petrovskii, H. Malchow, Mathematical models of marine ecosystems, in: J. Filar (Ed.), Mathematical Models, In: *The Encyclopedia of Life Support Systems (EOLSS)*, EOLSS Publishers, Oxford UK, 2004, [`http://www.eolss.net`].

[5] H. Malchow, F. M. Hilker, Pattern formation in models of nonlinear plankton dynamics: a minireview, in: B. Schröder, H. Reuter, B. Reineking (Eds.), GfÖ Arbeitskreis Theorie in der Ökologie 2005: Multiple Skalen und Skalierung in der Ökologie, Peter Lang Verlag, Frankfurt/M., 2007, in press.

[6] G. Nicolis, I. Prigogine, Self-organization in nonequilibrium systems, Wiley-Interscience, New York, 1977.

[7] H. Haken, Synergetics. An introduction, Vol. 1 of Springer Series in Synergetics, Springer, Berlin, 1978.

[8] G. Hallegraeff, Plankton. A microscopic world, E. J. Brill, Leiden, 1988.

[9] V. Hensen (Ed.), Ergebnisse der in dem Atlantischen Ocean von Mitte Juli bis Anfang November 1889 ausgeführten Plankton-Expedition der Humboldt-Stiftung, Verlag von Lipsius & Tischer, Kiel und Leipzig, 1892.

[10] R. Porep, Der Physiologe und Planktonforscher Victor Hensen (1835-1924). Sein Leben und Werk, Vol. 9 of Kieler Beiträge zur Geschichte der Medizin und Pharmazie, Karl Wachholtz Verlag, Neumünster, 1970.

[11] R. C. Bain Jr., Predicting DO variations caused by algae, Journal of the Sanitary Engineering Division, Proceedings of the American Society of Civil Engineers (October 1968) 867–881.

[12] J. Duinker, G. Wefer, Das CO_2-Problem und die Rolle des Ozeans, Naturwissenschaften 81 (1994) 237–242.

[13] G. Malin, Sulphur, climate and the microbial maze, Nature 387 (1994) 857–859.

[14] R. L. Ritschard, Marine algae as a CO_2 sink, Water, Air and Soil Pollution 64 (1992) 289–303.

[15] R. Charlson, J. Lovelock, M. Andreae, S. Warren, Oceanic phyto-plankton, atmospheric sulphur, cloud albedo and climate, Nature 326 (1987) 655–661.

[16] P. Williamson, J. Gribbin, How plankton change the climate, New Scientist 1760 (1991) 48–52.

[17] G. C. Hays, A. J. Richardson, C. Robinson, Climate change and marine plankton, Trends in Ecology & Evolution 20 (6) (2005) 337–344.

[18] T. Kobari, A. Shinada, A. Tsuda, Functional roles of interzonal migrating mesozooplankton in the western subarctic Pacific, Progress in Oceanography 57 (2003) 279–298.

[19] P. J. Harrison, F. A. Whitney, A. Tsuda, H. Saito, K. Tadokoro, Nutrient and plankton dynamics in NE and NW gyres of the subarctic Pacific ocean, Journal of Oceanography 60 (2004) 93–117.

[20] J. E. G. Raymont, Plankton and productivity in the oceans, Pergamon Press, Oxford, 1980.

[21] U. Sommer, Planktologie, Springer, Berlin, 1994.

[22] D. H. Cushing, Marine ecology and fisheries, Cambridge University Press, Cambridge, 1975.

[23] J. A. Gulland, Fish population dynamics, Wiley, New York, 1977.

[24] J. H. Steele (Ed.), Fisheries mathematics, Academic Press, London, 1977.

[25] M. J. R. Fasham, The statistical and mathematical analysis of plankton patchiness, Oceanography and Marine Biology: an Annual Review 16 (1978) 43–79.

[26] J. H. Steele (Ed.), Spatial patterns in plankton communities, Vol. 3 of NATO Conf. Series IV (Marine Sciences), Plenum Press, New York, 1978.

[27] D. L. Mackas, C. M. Boyd, Spectral analysis of zooplankton spatial heterogeneity, Science 204 (1979) 62–64.

[28] C. H. Greene, E. A. Widder, M. J. Youngbluth, A. Tamse, G. E. Johnson, The migration behavior, fine structure, and bioluminescent activity of krill sound–scattering layer, Limnology and Oceanography 37 (1992) 650–658.

[29] M. Abbott, Phytoplankton patchiness: ecological implications and observation methods, in: S. A. Levin, T. M. Powell, J. H. Steele (Eds.), Patch Dynamics, Vol. 96 of Lecture Notes in Biomathematics, Springer, Berlin, 1993, pp. 37–49.

[30] R. W. Sterner, D. O. Hessen, Algal nutrient limitation and the nutrition of aquatic herbivores, Annual Review of Ecology and Systematics 25 (1994) 1–29.

[31] K. L. Denman, Covariability of chlorophyll and temperature in the sea, Deep-Sea Research 23 (1976) 539–550.

[32] L. H. Weber, S. Z. El-Sayed, I. Hampton, The variance spectra of phytoplankton, krill and water temperature in the Antarctic ocean south of Africa, Deep-Sea Research 33 (1986) 1327–1343.

[33] T. Platt, Local phytoplankton abundance and turbulence, Deep-Sea Research 19 (1972) 183–187.

[34] T. M. Powell, P. J. Richerson, T. M. Dillon, B. A. Agee, B. J. Dozier, D. A. Godden, L. O. Myrup, Spatial scales of current speed and phytoplankton biomass fluctuations in Lake Tahoe, Science 189 (1975) 1088–1090.

[35] K. Nakata, R. Ishikawa, Fluctuation of local phytoplankton abundance in coastal waters, Japanese Journal of Ecology 25 (1975) 201–205.

[36] T. M. Powell, A. Okubo, Turbulence, diffusion and patchiness in the sea, Proceedings of the Royal Society of London B 343 (1994) 11–18.

[37] L. Seuront, F. Schmitt, Y. Lagadeuc, D. Schertzer, S. Lovejoy, Universal multifractal analysis as a tool to characterize multiscale intermittent patterns: example of phytoplankton distribution in turbulent coastal waters, Journal of Plankton Research 21 (1999) 877–922.

[38] T. Wyatt, Production dynamics of *Oikopleura dioica* in the Southern North Sea, and the role of fish larvae which prey on them, Thalassia Jugoslavica 7 (1971) 435–444.

[39] T. Wyatt, The biology of *Oikopleura dioica* and *Fritillaria borealis* in the Southern Bight, Marine Biology 22 (1973) 137–158.

[40] J. H. Steele, E. W. Henderson, A simple plankton model, The American Naturalist 117 (1981) 676–691.

[41] S. A. Levin, Physical and biological scales and the modelling of predator-prey interactions in large marine ecosystems, in: K. Sherman, L. M. Alexander, B. Gold (Eds.), Large marine ecosystems: patterns, processes and yields, American Association for the Advancement of Science, Washington, 1990, pp. 179–187.

[42] T. M. Powell, Physical and biological scales of variability in lakes, estuaries and the coastal ocean, in: T. M. Powell, J. H. Steele (Eds.), Ecological Time Series, Chapman & Hall, New York, 1995, pp. 119–138.

[43] U. Sommer, Algen, Quallen, Wasserfloh. Die Welt des Planktons, Springer, Berlin, 1996.

[44] K. L. Daly, W. O. Smith Jr., Physical-biological interactions influencing marine plankton production, Annual Review of Ecology and Systematics 24 (1993) 555–585.

[45] J. J. O'Brien, J. S. Wroblewski, On advection in phytoplankton models, Journal of Theoretical Biology 38 (1973) 197–202.

[46] J. S. Wroblewski, J. J. O'Brien, T. Platt, On the physical and biological scales of phytoplankton patchiness in the ocean, Mémoires de la Société Royale des Sciences de Liège, Série 6, Tome VII (1975) 43–57.

[47] J. S. Wroblewski, J. J. O'Brien, A spatial model of phytoplankton patchiness, Marine Biology 35 (1976) 161–175.

[48] S. A. Levin, L. A. Segel, Hypothesis for origin of planktonic patchiness, Nature 259 (1976) 659.

[49] J. H. Steele, E. W. Henderson, The role of predation in plankton models, Journal of Plankton Research 14 (1992) 157–172.

[50] J. Graunt, Natural and political observations made upon the bills of mortality, Martyn, London, 1662.

[51] L. Euler, Recherches générales sur la mortalité et la multiplication du genre humain, Mémoires de l'Académie Royale des Sciences et Belles-Lettres 16 (1760) 144–164.

[52] T. R. Malthus, An essay on the principle of population, J. Johnson in St. Paul's Churchyard, London, 1798.

[53] B. Gompertz, On the nature of the function expressive of the law of human mortality, and a new mode of determining the value of life contengencies, Philosophical Transactions of the Royal Society of London 115 (1825) 513–585.

[54] P. F. Verhulst, Notice sur la loi que la population suit dans son accroissement, Correspondance Mathématique et Physique Publiée par A. Quételet 10 (1838) 113–121.

[55] A. J. Lotka, Elements of physical biology, Williams and Wilkins, Baltimore, 1925.

[56] V. Volterra, Variazioni e fluttuazioni del numero d'individui in specie animali conviventi, Atti della Reale Accademia Nazionale dei Lincei, Memorie della Classe di Scienze Fisiche, Matematiche e Naturali, Serie 6, Volume II (3) (1926) 31–113.

[57] R. Fleming, The control of diatom populations by grazing, Journal du Conseil Permanent International pour l'Exploration de la Mer 14 (1939) 210–227.

[58] V. S. Ivlev, Biologicheskaya produktivnost' vodoemov, Uspekhi Sovremennoi Biologii XIX (1945) 98–120.

[59] G. A. Riley, Factors controlling phytoplankton populations on Georges Bank, Journal of Marine Research 6 (1946) 54–73.

[60] H. Odum, Primary production in flowing waters, Limnology and Oceanography 1 (1956) 102–117.

[61] M. Droop, 25 years of algal growth kinetics, Botanica Marina XXVI (1983) 99–112.

[62] M. J. Behrenfeldt, P. G. Falkowski, A consumer's guide to phytoplankton primary productivity models, Limnology and Oceanography 42 (1997) 1479–1491.

[63] K. L. Denman, Modelling planktonic ecosystems: parametrizing complexity, Progress in Oceanography 57 (2003) 429–452.

[64] G. A. Riley, Theory of food-chain relations in the ocean, in: M. Hill (Ed.), The Sea, Vol. 2, Wiley, New York, 1963, pp. 438–463.

[65] L. A. Segel, J. L. Jackson, Dissipative structure: an explanation and an ecological example, Journal of Theoretical Biology 37 (1972) 545–559.

[66] D. Dubois, A model of patchiness for prey-predator plankton populations, Ecological Modelling 1 (1975) 67–80.

[67] M. E. Vinogradov, V. V. Menshutkin, Modeling open-sea systems, in: E. D. Goldberg (Ed.), The Sea: Ideas and Observations on Progress in the Study of the Seas, Vol. 6, Wiley, New York, 1977, pp. 891–921.

[68] M. Mimura, J. D. Murray, On a diffusive prey-predator model which exhibits patchiness, Journal of Theoretical Biology 75 (1978) 249–262.

[69] P. Mayzaud, S. A. Poulet, The importance of the time factor in the response of zooplankton to varying concentrations of naturally occuring particulate matter, Limnology and Oceanography 23 (1978) 1144–1154.

[70] C. S. Holling, Some characteristics of simple types of predation and parasitism, The Canadian Entomologist 91 (7) (1959) 385–398.

[71] L. Michaelis, M. Menten, Die Kinetik der Invertinwirkung, Biochemische Zeitschrift 49 (1913) 333–369.

[72] J. Monod, F. Jacob, General conclusions: Teleonomic mechanisms in cellular metabolism, growth and differentiation, Cold Spring Harbor Symposia on Quantitative Biology 26 (1961) 389–401.

[73] M. Scheffer, Fish and nutrients interplay determines algal biomass: a minimal model, Oikos 62 (1991) 271–282.

[74] J. H. Steele, E. W. Henderson, A simple model for plankton patchiness, Journal of Plankton Research 14 (1992) 1397–1403.

[75] H. Malchow, Spatio-temporal pattern formation in nonlinear nonequilibrium plankton dynamics, Proceedings of the Royal Society of London B 251 (1993) 103–109.

[76] M. Pascual, Diffusion-induced chaos in a spatial predator-prey system, Proceedings of the Royal Society of London B 251 (1993) 1–7.

[77] J. E. Truscott, J. Brindley, Equilibria, stability and excitability in a general class of plankton population models, Philosophical Transactions of the Royal Society of London A 347 (1994) 703–718.

[78] J. E. Truscott, J. Brindley, Ocean plankton populations as excitable media, Bulletin of Mathematical Biology 56 (1994) 981–998.

[79] A. M. Edwards, J. Brindley, Oscillatory behaviour in a three-component plankton population model, Dynamics and Stability of Systems 11 (1996) 347–370.

[80] J. W. Pitchford, J. Brindley, Intratrophic predation in simple predator–prey models, Bulletin of Mathematical Biology 60 (1998) 937–953.

[81] M. Scheffer, Ecology of shallow lakes, Vol. 22 of Population and Community Biology Series, Chapman & Hall, London, 1998.

[82] A. M. Edwards, Adding detritus to a nutrient-phytoplankton-zooplankton model: a dynamical-systems approach, Journal of Plankton Research 23 (4) (2001) 389–413.

[83] G. A. Gibson, D. L. Musgrave, S. Hinckley, Non-linear dynamics of a pelagic ecosystem model with multiple predator and prey types, Journal of Plankton Research 27 (5) (2005) 427–447.

[84] E. Beltrami, A mathematical model of the brown tide, Estuaries 12 (1989) 13–17.

[85] E. Beltrami, Unusual algal blooms as excitable systems: The case of "brown-tides", Environmental Modeling & Assessment 1 (1996) 19–24.

[86] H. Malchow, F. M. Hilker, R. R. Sarkar, K. Brauer, Spatiotemporal patterns in an excitable plankton system with lysogenic viral infection, Mathematical and Computer Modelling 42 (9-10) (2005) 1035–1048.

[87] A. L. Hodgkin, A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, Journal of Physiology 117 (1952) 500–544.

[88] R. Fitzhugh, Impulses and physiological states in theoretical models of nerve membrane, Biophysical Journal 1 (6) (1961) 445–466.

[89] J. Nagumo, S. Arimoto, S. Yoshizawa, An active pulse transmission line simulating nerve axon, Proceedings of the Institute of Radio Engineers 50 (1962) 2061–2070.

[90] D. DeAngelis, Dynamics of nutrient cycling and food webs, Vol. 9 of Population and Community Biology Series, Chapman & Hall, London, 1992.

[91] S. E. Jørgensen, Fundamentals of ecological modelling, Vol. 19 of Developments in Environmental Modelling, Elsevier, Amsterdam, 1994.

[92] P. Yodzis, The trophodynamics of whole ecological communities, in: S. Levin (Ed.), Frontiers in Mathematical Biology, Vol. 100 of Lecture Notes in Biomathematics, Springer, Berlin, 1994, pp. 443–453.

[93] R. M. May, Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos, Science 186 (1974) 645–647.

[94] M. Scheffer, Should we expect strange attractors behind plankton dynamics – and if so, should we bother?, Journal of Plankton Research 13 (1991) 1291–1305.

[95] R. F. Costantino, R. A. Desharnais, J. M. Cushing, B. Dennis, Chaotic dynamics in an insect population, Science 275 (1997) 389–391.

[96] J. Huisman, F. Weissing, Biodiversity of plankton by oscillations and chaos, Nature 402 (1999) 407–410.

[97] J.M. Cushing, R. Costantino, B. Dennis, R.A. Desharnais, S. Henson, Chaos in ecology. Experimental nonlinear dynamics, Theoretical Ecology Series, Academic Press, Amsterdam, 2003.

[98] L. Becks, F.M. Hilker, H. Malchow, K. Jürgens, H. Arndt, Experimental demonstration of chaos in a microbial food web, Nature 435 (2005) 1226–1229.

[99] S.V. Petrovskii, H. Malchow, A minimal model of pattern formation in a prey-predator system, Mathematical and Computer Modelling 29 (1999) 49–63.

[100] S.V. Petrovskii, H. Malchow, Wave of chaos: new mechanism of pattern formation in spatio-temporal population dynamics, Theoretical Population Biology 59 (2) (2001) 157–174.

[101] S.V. Petrovskii, A.Y. Morozov, E. Venturino, Allee effect makes possible patchy invasion in a predator-prey system, Ecology Letters 5 (2002) 345–352.

[102] S.V. Petrovskii, B.-L. Li, H. Malchow, Quantification of the spatial aspect of chaotic dynamics in biological and chemical systems, Bulletin of Mathematical Biology 65 (3) (2003) 425–446.

[103] S.V. Petrovskii, B.-L. Li, H. Malchow, Transition to spatiotemporal chaos can resolve the paradox of enrichment, Ecological Complexity 1 (1) (2004) 37–47.

[104] G.T. Evans, J.S. Parslow, A model of annual plankton cycles, Biological Oceanography 3 (3) (1985) 327–347.

[105] J.E. Truscott, Environmental forcing of simple plankton models, Journal of Plankton Research 17 (1995) 2207–2232.

[106] E.E. Popova, M.J.R. Fasham, A.V. Osipov, V.A. Ryabchenko, Chaotic behaviour of an ocean ecosystem model under seasonal external forcing, Journal of Plankton Research 19 (1997) 1495–1515.

[107] V.A. Ryabchenko, M.J.R. Fasham, B. Kagan, E. Popova, What causes short-term oscillations in ecosystem models of the ocean mixed layer?, Journal of Marine Systems 13 (1997) 33–50.

[108] Y.A. Kuznetsov, S. Muratori, S. Rinaldi, Bifurcations and chaos in a periodic predator-prey model, International Journal of Bifurcation and Chaos 2 (1992) 117–128.

[109] F.A. Ascioti, E. Beltrami, T.O. Carroll, C. Wirick, Is there chaos in plankton dynamics?, Journal of Plankton Research 15 (1993) 603–617.

[110] F. Doveri, M. Scheffer, S. Rinaldi, S. Muratori, Y. Kuznetsov, Seasonality and chaos in a plankton-fish model, Theoretical Population Biology 43 (1993) 159–183.

[111] S. Rinaldi, S. Muratori, Conditioned chaos in seasonally perturbed predator-prey models, Ecological Modelling 69 (1993) 79–97.

[112] S. Rinaldi, S. Muratori, Y. Kuznetsov, Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities, Bulletin of Mathematical Biology 55 (1993) 15–35.

[113] E. Steffen, H. Malchow, Chaotic behaviour of a model plankton community in a heterogeneous environment, in: F. Schweitzer (Ed.), Selforganisation of complex structures: From individual to collective dynamics, Gordon and Breach, London, 1996, pp. 331–340.

[114] E. Steffen, H. Malchow, Multiple equilibria, periodicity, and quasiperiodicity in a model plankton community, Senckenbergiana maritima 27 (1996) 137–143.

[115] M. Scheffer, S. Rinaldi, Y. A. Kuznetsov, E. H. van Nes, Seasonal dynamics of daphnia and algae explained as a periodically forced predator-prey system, Oikos 80 (1997) 519–532.

[116] E. Steffen, H. Malchow, A. B. Medvinsky, Effects of seasonal perturbation on a model plankton community, Environmental Modeling & Assessment 2 (1997) 43–48.

[117] H. Malchow, N. Shigesada, Nonequilibrium plankton community structures in an ecohydrodynamic model system, Nonlinear Processes in Geophysics 1 (1) (1994) 3–11.

[118] C. W. Gardiner, Handbook of stochastic methods, Vol. 13 of Springer Series in Synergetics, Springer, Berlin, 1985.

[119] V. S. Anishenko, V. V. Astakov, A. B. Neiman, T. Vadivasova, L. Schimansky-Geier, Nonlinear dynamics of chaotic and stochastic systems. Tutorial and modern developments, Springer Series in Synergetics, Springer, Berlin, 2003.

[120] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, B. Walker, Catastrophic shifts in ecosystems, Nature 413 (2001) 591–596.

[121] M. Scheffer, S. R. Carpenter, Catastrophic regime shifts in ecosystems: linking theory to observation, Trends in Ecology & Evolution 18 (12) (2003) 648–656.

[122] J. S. Collie, K. Richardson, J. H. Steele, Regime shifts: can ecological theory illuminate the mechanisms?, Progress in Oceanography 60 (2004) 281–302.

[123] M. Rietkerk, S. C. Dekker, P. C. de Ruiter, J. van de Koppel, Self-organized patchiness and catastrophic shifts in ecosystems, Science 305 (2004) 1926–1929.

[124] J. H. Steele, Regime shifts in the ocean: reconciling observations and theory, Progress in Oceanography 60 (2004) 135–141.

[125] J. A. Freund, S. Mieruch, B. Scholze, K. Wiltshire, U. Feudel, Bloom dynamics in a seasonally forced phytoplankton-zooplankton model: Trigger machanisms and timing effects, Ecological Complexity 3 (2006) 129–139.

[126] H. Hempel, L. Schimansky-Geier, J. Garcia-Ojalvo, Noise-sustained pulsating patterns and global oscillations in subexcitable media, Physical Review Letters 82 (18) (1999) 3713–3716.

[127] A. Neiman, L. Schimansky-Geier, A. Cornell-Bell, F. Moss, Noise-enhanced phase synchronization in excitable media, Physical Review Letters 83 (23) (1999) 4896–4899.

[128] H. Malchow, L. Schimansky-Geier, Coherence resonance in an excitable prey-predator plankton system with infected prey, in: T. Pöschel, H. Malchow, L. Schimansky-Geier (Eds.), Irreversible Prozesse und Selbstorganisation, Logos Verlag, Berlin, 2006, pp. 293–301.

[129] J. A. Freund, L. Schimansky-Geier, B. Beisner, A. Neiman, D. F. Russell, T. Yakusheva, F. Moss, Behavioral stochastic resonance: How the noise from a Daphnia swarm enhances individual prey capture by juvenile paddlefish, Journal of Theoretical Biology 214 (2002) 71–83.

[130] J. García-Ojalvo, J. M. Sancho, Noise in spatially extended systems, Institute for Nonlinear Science, Springer, New York, 1999.

[131] B. Lindner, J. García-Ojalvo, A. Neiman, L. Schimansky-Geier, Effects of noise in excitable systems, Physics Reports 392 (2004) 321–424.

[132] B. Spagnolo, D. Valenti, A. Fiasconaro, Noise in ecosystems: a short review, Mathematical Biosciences and Engineering 1 (1) (2004) 185–211.

[133] M. Sieber, H. Malchow, L. Schimansky-Geier, Constructive effects of environmental noise in an excitable prey-predator plankton system with infected prey, Ecological Complexity (2007), submitted.

[134] L. J. S. Allen, An introduction to stochastic processes with applications to biology, Pearson Education, Upper Saddle River NJ, 2003.

[135] A. Okubo, Diffusion and ecological problems: Mathematical models, Vol. 10 of Biomathematics Texts, Springer, Berlin, 1980.

[136] A. Okubo, S. Levin, Diffusion and ecological problems: Modern perspectives, Vol. 14 of Interdisciplinary Applied Mathematics, Springer, New York, 2001.

[137] H. Stommel, Trajectories of small bodies sinking slowly through convection cells, Journal of Marine Research 8 (1948) 24–29.

[138] S. Leibovich, Spatial aggregation arising from convective processes, in: S. A. Levin, T. M. Powell, J. H. Steele (Eds.), Patch dynamics, Vol. 96 of Lecture Notes in Biomathematics, Springer, Berlin, 1993, pp. 110–124.

[139] J. A. Yoder, S. G. Ackleson, R. T. Barber, P. Flament, W. M. Balch, A line in the sea, Nature 371 (1994) 689–692.

[140] P. J. S. Franks, Spatial patterns in dense algal blooms, Limnology and Oceanography 42 (5, part 2) (1997) 1297–1305.

[141] E. R. Abraham, The generation of plankton patchiness by turbulent stirring, Nature 391 (1998) 577–580.

[142] V. N. Biktashev, I. V. Biktasheva, J. Brindley, A. V. Holden, N. A. Hill, M. A. Tsyganov, Effects of shear flows on nonlinear waves in excitable media, Journal of Biological Physics 25 (2) (1999) 101–113.

[143] A. P. Martin, Phytoplankton patchiness: the role of lateral stirring and mixing, Progress in Oceanography 57 (2003) 125–174.

[144] I. Scheuring, G. Károlyi, Z. Toroczkai, T. Tél, A. Péntek, Competing populations in flows with chaotic mixing, Theoretical Population Biology 63 (2003) 77–90.

[145] E. Hernández-García, C. López, Sustained plankton blooms under open chaotic flows, Ecological Complexity 1 (2004) 253–259.

[146] C. Nägeli, Ortsbewegungen der Pflanzenzellen und ihrer Theile (Strömungen), Beiträge zur Wissenschaftlichen Botanik 2 (1860) 59–108.

[147] H. Wager, On the effect of gravity upon the movements and aggregation of *Euglena viridis*, Ehrb., and other micro-organisms, Philosophical Transactions of the Royal Society of London B 201 (1911) 333–390.

[148] J. R. Platt, “Bioconvection patterns” in cultures of free-swimming organisms, Science 133 (1961) 1766–1767.

[149] H. Winet, T. L. Jahn, On the origin of bioconvective fluid instabilities in *Tetrahymena* culture systems, Biorheology 9 (1972) 87–104.

[150] S. Childress, M. Levandowsky, E. A. Spiegel, Pattern formation in a suspension of swimming micro-organisms: equations and stability theory, Journal of Fluid Mechanics 63 (1975) 591–613.

[151] M. Levandowsky, W. S. Childress, E. A. Spiegel, S. H. Hutner, A mathematical model of pattern formation by swimming microorganisms, Journal of Protozoology 22 (1975) 296–306.

[152] J. O. Kessler, Co-operative and concentrative phenomena of swimming microorganisms, Contemporary Physics 26 (2) (1985) 147–166.

[153] T. J. Pedley, J. O. Kessler, Hydrodynamic phenomena in suspensions of swimming microorganisms, Annual Review of Fluid Mechanics 24 (1992) 313–358.

[154] T. J. Pedley, N. A. Hill, J. O. Kessler, The growth of bioconvection patterns in a uniform suspension of gyrotactic micro-organisms, Journal of Fluid Mechanics 195 (1988) 223–237.

[155] J. G. Mitchell, A. Okubo, J. A. Fuhrman, Gyrotaxis as a new mechanism for generating spatial heterogeneity and migration in microplankton, Limnology and Oceanography 35 (1) (1990) 123–130.

[156] U. Timm, A. Okubo, Gyrotaxis: A plume model for self-focusing micro-organisms, Bulletin of Mathematical Biology 56 (2) (1994) 187–206.

[157] U. Timm, A. Okubo, Gyrotaxis: Interaction between algae and flagellates, Bulletin of Mathematical Biology 57 (5) (1995) 631–650.

[158] T. A. Witten, L. M. Sander, Diffusion-limited aggregation, a kinetic critical phenomenon, Physical Review Letters 47 (1981) 1400–1403.

[159] M. Matsushita, H. Fujikawa, Diffusion-limited growth in bacterial colony formation, Physica A 168 (1990) 498–506.

[160] E. Ben-Jacob, H. Shmueli, O. Shochet, A. Tenenbaum, Adaptive self-organization during growth of bacterial colonies, Physica A 87 (1992) 378–424.

[161] H. Malchow, S. V. Petrovskii, A. B. Medvinsky, Numerical study of plankton-fish dynamics in a spatially structured and noisy environment, Ecological Modelling 149 (2002) 247–255.

[162] H. Malchow, F. M. Hilker, S. V. Petrovskii, Noise and productivity dependence of spatiotemporal pattern formation in a prey-predator system, Discrete and Continuous Dynamical Systems B 4 (3) (2004) 707–713.

[163] H. Malchow, F. M. Hilker, S. V. Petrovskii, K. Brauer, Oscillations and waves in a virally infected plankton system. Part I: The lysogenic stage, Ecological Complexity 1 (3) (2004) 211–223.

[164] R. R. Sarkar, J. Chattopadhyay, Occurence of planktonic blooms under environmental fluctuations and its possible control mechanism – mathematical models and experimental observations, Journal of Theoretical Biology 224 (2003) 501–516.

[165] J. G. Skellam, The formulation and interpretation of mathematical models of diffusionary processes in population biology, in: M. S. Bartlett, R. Hiorns (Eds.), The mathematical theory of the dynamics of biological populations, Academic Press, New York, 1973, pp. 63–85.

[166] J. Jorné, The diffusive Lotka-Volterra oscillating system, Journal of Theoretical Biology 65 (1977) 133–139.

[167] N. Shigesada, E. Teramoto, A consideration on the theory of environmental density (in Japanese), Japanese Journal of Ecology 28 (1978) 1–8.

[168] N. Shigesada, K. Kawasaki, E. Teramoto, Spatial segregation of interacting species, Journal of Theoretical Biology 79 (1979) 83–99.

[169] H. Malchow, Dissipative pattern formation in ternary nonlinear reaction-electro-diffusion systems with concentration-dependent diffusivities, Journal of Theoretical Biology 135 (1988) 371–381.

[170] J. Jorné, Negative ionic cross diffusion coefficients in electrolytic solutions, Journal of Theoretical Biology 55 (1975) 529–532.

[171] H. Malchow, Spatial pattern formation in compartmental reaction-electrodiffusion systems with concentration-dependent diffusivities, Memoirs of the Faculty of Science, Kyoto University (Series of Biology) 13 (2) (1988) 71–82.

[172] H. Malchow, Flux-induced instabilities in ionic and population-dynamical interaction systems, Zeitschrift für Physikalische Chemie 204 (1998) 95–107.

[173] J. D. Murray, Mathematical biology, Vol. 19 of Biomathematics Texts, Springer, Berlin, 1989.

[174] E. E. Holmes, M. A. Lewis, J. E. Banks, R. R. Veit, Partial differential equations in ecology: Spatial interactions and population dynamics, Ecology 75 (1994) 17–29.

[175] N. Shigesada, K. Kawasaki, Biological invasions: Theory and practice, Oxford University Press, Oxford, 1997.

[176] R. S. Cantrell, C. Cosner, Spatial ecology via reaction-diffusion equations, Wiley Series in Mathematical and Computational Ecology, Wiley, Chichester, 2003.

[177] J. D. Murray, Mathematical biology. II. Spatial models and biomedical applications, Vol. 18 of Interdisciplinary Applied Mathematics, Springer, Berlin, 2003.

[178] A. M. Turing, On the chemical basis of morphogenesis, Philosophical Transactions of the Royal Society of London B 237 (1952) 37–72.

[179] M. L. Rosenzweig, R. H. MacArthur, Graphical representation and stability conditions of predator-prey interactions, The American Naturalist 97 (1963) 209–223.

[180] H. Malchow, Nonequilibrium structures in plankton dynamics, Ecological Modelling 75/76 (1994) 123–134.

[181] H. Malchow, B. Radtke, M. Kallache, A. B. Medvinsky, D. A. Tikhonov, S. V. Petrovskii, Spatio-temporal pattern formation in coupled models of plankton dynamics and fish school motion, Nonlinear Analysis: Real World Applications 1 (2000) 53–67.

[182] A. B. Medvinsky, S. V. Petrovskii, I. A. Tikhonova, H. Malchow, B.-L. Li, Spatiotemporal complexity of plankton and fish dynamics, SIAM Review 44 (3) (2002) 311–370.

[183] A. B. Rovinsky, M. Menzinger, Chemical instability induced by a differential flow, Physical Review Letters 69 (1992) 1193–1196.

[184] H. Malchow, Flow- and locomotion-induced pattern formation in nonlinear population dynamics, Ecological Modelling 82 (1995) 257–264.

[185] H. Malchow, Nonlinear plankton dynamics and pattern formation in an ecohydrodynamic model system, Journal of Marine Systems 7 (2-4) (1996) 193–202.

[186] A. B. Rovinsky, H. Adiwidjaja, V. Z. Yakhnin, M. Menzinger, Patchiness and enhancement of productivity in plankton ecosystems due to differential advection of predator and prey, Oikos 78 (1997) 101–106.

[187] C. A. Klausmeier, Regular and irregular patterns in semiarid vegetation, Science 284 (1999) 1826–1828.

[188] H. Malchow, Motional instabilities in predator-prey systems, Journal of Theoretical Biology 204 (2000) 639–647.

[189] J. G. Skellam, Random dispersal in theoretical populations, Biometrika 38 (1951) 196–218.

[190] H. Kierstead, L. B. Slobodkin, The size of water masses containing plankton blooms, Journal of Marine Research XII (1) (1953) 141–147.

[191] R. Luther, Räumliche Ausbreitung chemischer Reaktionen, Zeitschrift für Elektrochemie 12 (1906) 596–600.

[192] R. A. Fisher, The wave of advance of advantageous genes, Annals of Eugenics 7 (1937) 355–369.

[193] A. Kolmogorov, I. Petrovskii, N. Piskunov, Étude de l'equation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique, Bulletin de l'Université de Moscou, Série Internationale, Section A 1 (1937) 1–25.

[194] D. C. Speirs, W. S. C. Gurney, Population persistence in rivers and estuaries, Ecology 82 (2001) 1219–1237.

[195] W. C. Allee, Animal Aggregations: A Study in General Sociology, University of Chicago Press, Chicago, 1931.

[196] W. C. Allee, A. E. Emerson, O. Park, T. Park, K. P. Schmidt, Principles of Animal Ecology, Saunders, Philadelphia, 1949.

[197] B. Dennis, Allee effects: population growth, critical density, and the chance of extinction, Natural Resource Modeling 3 (1989) 481–538.

[198] F. Courchamp, T. Clutton-Brock, B. Grenfell, Inverse density dependence and the Allee effect, Trends in Ecology & Evolution 14 (1999) 405–410.

[199] M. Gyllenberg, J. Hemminki, T. Tammaru, Allee effects can both conserve and create spatial heterogeneity in population densities, Theoretical Population Biology 56 (1999) 231–242.

[200] P. A. Stephens, W. J. Sutherland, R. P. Freckleton, What is the Allee effect?, Oikos 87 (1999) 185–190.

[201] P. A. Stephens, W. J. Sutherland, Consequences of the Allee effect for behaviour, ecology and conservation, Trends in Ecology & Evolution 14 (10) (1999) 401–405.

[202] H. R. Thieme, Mathematics in Population Biology, Princeton University Press, Princeton NJ, 2003.

[203] F. Schlögl, Chemical reaction models for nonequilibrium phase transitions, Zeitschrift für Physik 253 (1972) 147–161.

[204] A. Nitzan, P. Ortoleva, J. Ross, Nucleation in systems with multiple stationary states, Faraday Symposia of the Chemical Society 9 (1974) 241–253.

[205] W. Ebeling, L. Schimansky-Geier, Nonequilibrium phase transitions and nucleation in reaction systems, in: Proceedings of the 6th International Conference on Thermodynamics, Merseburg, 1980, pp. 95–100.

[206] H. Malchow, L. Schimansky-Geier, Noise and diffusion in bistable nonequilibrium systems, Vol. 5 of Teubner-Texte zur Physik, Teubner-Verlag, Leipzig, 1985.

[207] M. A. Lewis, P. Kareiva, Allee dynamics and the spread of invading organisms, Theoretical Population Biology 43 (1993) 141–158.

[208] S. V. Petrovskii, Approximate determination of the magnitude of the critical size in the problem of the evolution of an ecological impact, Journal of Engineering Physics and Thermophysics 66 (1994) 346–352.

[209] F. M. Hilker, Spatiotemporal patterns in models of biological invasion and epidemic spread, Logos Verlag, Berlin, 2005.

[210] F. M. Hilker, M. Langlais, S. V. Petrovskii, H. Malchow, A diffusive SI model with Allee effect and application to FIV, Mathematical Biosciences 206 (2007) 61–80.

[211] D. Ludwig, D. D. Jones, C. S. Holling, Qualitative analysis of insect outbreak systems: the spruce budworm and forest, Journal of Animal Ecology 47 (1978) 315–332.

[212] C. Wissel, Theoretische Ökologie. Eine Einführung, Springer, Berlin, 1989.

[213] R. J. Field, M. Burger (Eds.), Oscillations and traveling waves in chemical systems, Wiley, New York, 1985.

[214] M. C. Boerlijst, M. Lamers, P. Hogeweg, Evolutionary consequences of spiral waves in a host-parasitoid system, Proceedings of the Royal Society of London B 253 (1993) 15–18.

[215] G. Gerisch, Cell aggregation and differentiation in *Dictyostelium*, in: A. A. Moscona, A. Monroy (Eds.), Current Topics in Developmental Biology, Vol. 3, Academic Press, New York, 1968, pp. 157–197.

[216] E. F. Keller, L. A. Segel, Initiation of slime mold aggregation viewed as an instability, Journal of Theoretical Biology 26 (1970) 399–415.

[217] G. Gerisch, Periodische Signale steuern die Musterbildung in Zellverbänden, Naturwissenschaften 58 (1971) 430–438.

[218] L. A. Segel, B. Stoeckly, Instability of a layer of chemotactic cells, attractant and degrading enzyme, Journal of Theoretical Biology 37 (1972) 561–585.

[219] L. A. Segel, A theoretical study of receptor mechanisms in bacterial chemotaxis, SIAM Journal on Applied Mathematics 32 (1977) 653–665.

[220] P. C. Newel, Attraction and adhesion in the slime mold Dictyostelium, in: J. E. Smith (Ed.), Fungal differentiation. A contemporary synthesis, Vol. 43 of Mycology Series, Marcel Dekker, New York, 1983, pp. 43–71.

[221] W. Alt, G. Hoffmann (Eds.), Biological motion, Vol. 89 of Lecture Notes in Biomathematics, Springer, Berlin, 1990.

[222] F. Siegert, C. J. Weijer, Analysis of optical density wave propagation and cell movement in the cellular slime mould *Dictyostelium discoideum*, Physica D 49 (1991) 224–232.

[223] O. Steinbock, H. Hashimoto, S. C. Müller, Quantitative analysis of periodic chemotaxis in aggregation patterns of *Dictyostelium discoideum*, Physica D 49 (1991) 233–239.

[224] B. N. Vasiev, P. Hogeweg, A. V. Panfilov, Simulation of *Dictyostelium discoideum* aggregation via reaction-diffusion model, Physical Review Letters 73 (1994) 3173–3176.

[225] G. Ivanitskii, A. B. Medvinskii, M. A. Tsyganov, From the dynamics of population autowaves generated by living cells to neuroinformatics, Physics – Uspekhi 37 (1994) 961–989.

[226] T. Höfer, J. A. Sherratt, P. K. Maini, Cellular pattern formation during Dictyostelium aggregation, Physica D 85 (1995) 425–444.

[227] E. F. Keller, L. A. Segel, Model for chemotaxis, Journal of Theoretical Biology 30 (1971) 225–234.

[228] E. F. Keller, L. A. Segel, Traveling bands of chemotactic bacteria: A theoretical analysis, Journal of Theoretical Biology 30 (1971) 235–248.

[229] S. Ikegami, I. Imai, J. Kato, H. Ohtake, Chemotaxis toward inorganic phosphate in the red tide alga *Chattonella antiqua*, Journal of Plankton Research 17 (1995) 1587–1591.

[230] J. A. Shapiro, C. Hsu, *Escherichia coli* k-12 cell-cell interactions seen by time-lapse video, Journal of Bacteriology 171 (1989) 5963–5974.

[231] J. A. Shapiro, D. Trubatch, Sequential events in bacterial colony morphogenesis, Physica D 49 (1991) 214–223.

[232] K. Kawasaki, A. Mochizuki, N. Shigesada, A mathematical model of pattern formation in a bacterial colony (in Japanese), Control & Measurement 34 (1995) 811–816.

[233] K. Kawasaki, A. Mochizuki, M. Matsushita, T. Umeda, N. Shigesada, Modeling spatio-temporal patterns generated by *Bacillus subtilis*, Journal of Theoretical Biology 188 (1997) 177–185.

[234] J. A. Sherratt, M. A. Lewis, A. Fowler, Ecological chaos in the wake of invasion, Proceedings of the National Academy of Sciences of the United States of America 92 (1995) 2524–2528.

[235] J. A. Sherratt, B. T. Eagan, M. A. Lewis, Oscillations and chaos behind predator-prey invasion: mathematical artifact or ecological reality?, Philosophical Transactions of the Royal Society of London B 352 (1997) 21–38.

[236] S. V. Petrovskii, H. Malchow, Spatio-temporal chaos in an ecological community as a response to unfavourable environmental changes, Advances in Complex Systems 4 (2 & 3) (2001) 227–249.

[237] S. V. Petrovskii, H. Malchow, F. M. Hilker, E. Venturino, Patterns of patchy spread in deterministic and stochastic models of biological invasion and biological control, Biological Invasions 7 (2005) 771–793.

[238] J. A. Fuhrman, Marine viruses and their biogeochemical and ecological effects, Nature 399 (1999) 541–548.

[239] C. A. Suttle, Viruses in the sea, Nature 437 (2005) 356–361.

[240] C. A. Suttle, Do viruses control the oceans? Ocean life infections., Monthly Magazine of the American Museum of Natural History (February 1999).

[241] C. A. Suttle, A. M. Chan, M. T. Cottrell, Infection of phytoplankton by viruses and reduction of primary productivity, Nature 347 (1990) 467–469.

[242] S. Jacquet, M. Heldal, D. Iglesias-Rodriguez, A. Larsen, W. Wilson, G. Bratbak, Flow cytometric analysis of an *Emiliana huxleyi* bloom terminated by viral infection, Aquatic Microbial Ecology 27 (2002) 111–124.

[243] M. D. Gastrich, J. A. Leigh-Bell, C. J. Gobler, O. R. Anderson, S. W. Wilhelm, M. Bryan, Viruses as potential regulators of regional brown tide blooms caused by the alga, *Aureococcus anophagefferens*, Estuaries 27 (1) (2004) 112–119.

[244] E. Beltrami, T. O. Carroll, Modelling the role of viral disease in recurrent phytoplankton blooms, Journal of Mathematical Biology 32 (1994) 857–863.

[245] F. M. Hilker, H. Malchow, Strange periodic attractors in a prey-predator system with infected prey, Mathematical Population Studies 13 (3) (2006) 119–134.

[246] F. M. Hilker, H. Malchow, M. Langlais, S. V. Petrovskii, Oscillations and waves in a virally infected plankton system. Part II: Transition from lysogeny to lysis, Ecological Complexity 3 (2006) 200–208.

[247] A. Nold, Heterogeneity in disease-transmission modeling, Mathematical Biosciences 52 (1980) 227–240.

[248] H. W. Hethcote, The mathematics of infectious diseases, SIAM Review 42 (4) (2000) 599–653.

[249] H. McCallum, N. Barlow, J. Hone, How should pathogen transmission be modelled?, Trends in Ecology & Evolution 16 (6) (2001) 295–300.

[250] J. Chattopadhyay, S. Pal, Viral infection on phytoplankton-zooplankton system – a mathematical model, Ecological Modelling 151 (2002) 15–28.

[251] J. Chattopadhyay, R. R. Sarkar, G. Ghosal, Removal of infected prey prevent limit cycle oscillations in an infected prey-predator system – a mathematical study, Ecological Modelling 156 (2002) 113–121.

[252] J. Chattopadhyay, R. R. Sarkar, S. Pal, Dynamics of nutrient-phytoplankton interaction in the presence of viral infection, BioSystems 68 (2003) 5–17.

[253] A. M. Edwards, M. A. Bees, Generic dynamics of a simple plankton population model with a non-integer exponent of closure, Chaos, Solitons & Fractals 12 (2001) 289–300.

[254] J. Chattopadhyay, R. R. Sarkar, S. Mandal, Toxin-producing plankton may act as a biological control for planktonic blooms – field study and mathematical modelling, Journal of Theoretical Biology 215 (3) (2002) 333–344.

[255] B. K. Singh, J. Chattopadhyay, S. Sinha, The role of virus infection in a simple phytoplankton-zooplankton system, Journal of Theoretical Biology 231 (2004) 153–166.

[256] R. R. Sarkar, H. Malchow, Nutrients and toxin producing phytoplankton control algal blooms – a spatiotemporal study in a noisy environment, Journal of Biosciences 30 (5) (2005) 749–760.

[257] M. Scheffer, S. Rinaldi, Y. A. Kuznetsov, Effects of fish on plankton dynamics: a theoretical analysis, Canadian Journal of Fisheries and Aquatic Sciences 57 (6) (2000) 1208–1219.

[258] E. Beretta, Y. Kuang, Modeling and analysis of a marine bacteriophage infection, Mathematical Biosciences 149 (1998) 57–76.

[259] I. Siekmann, H. Malchow, E. Venturino, Predation may defeat spatial spread of infection, Journal of Biological Dynamics (2007), in press.

[260] G. E. Hutchinson, Introduction to population ecology, Yale University Press, New Haven, 1978.

Horst Malchow
Institute of Environmental Systems Research
Department of Mathematics and Computer Science
Barbarastr. 12
D-49069 Osnabrück
Germany
e-mail: malchow@uos.de

Frank M. Hilker
Centre for Mathematical Biology
Mathematical and Statistical Sciences
University of Alberta
632 Central Academic Building
Edmonton Alberta T6G 2G1
Canada
e-mail: fhilker@math.ualberta.ca

Ivo Siekmann
Institute of Environmental Systems Research
Department of Mathematics and Computer Science
Barbarastr. 12
D-49069 Osnabrück
Germany
e-mail: isiekman@uos.de

Sergei V. Petrovskii
Department of Mathematics
University of Leicester
Leicester LE1 7RH
United Kingdom
e-mail: sp237@le.ac.uk

Alexander B. Medvinsky
Institute for Theoretical & Experimental Biophysics
Russian Academy of Sciences
Pushchino, Moscow Region
142290 Russia
e-mail: medvinsky@iteb.ru

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 27–40

Toward a General Theory of Ecosystem Stability: Plankton–Nutrient Interaction as a Paradigm

Andrei Korobeinikov and Sergei V. Petrovskii

Abstract. Identification of conditions of ecosystem stability and stable populations coexistence is a problem of highest importance in mathematical ecology. It is usually studied under specific assumptions made regarding the functional form of nonlinear feedbacks. Apparently, such an approach is lacking generality. In this paper, we consider a chemostat-type model of the phytoplankton-nutrient interaction, which can be regarded as a simple ecosystem model, in a general case. The plankton growth/nutrient uptake rate is described by an unspecified function of two variables (i.e., of the nutrient concentration N and the plankton density P) and the plankton mortality is an arbitrary function of P. We provide a rigorous mathematical consideration of the global properties of this system and derive the conditions that ensure existence and uniqueness of a globally asymptotically stable equilibrium state. Interestingly, these conditions correspond to much weaker constraints on the plankton growth rate properties than monotonicity and non-convexity that are usually assumed. We also identify a parameter that allows us to distinguish between existence and non-existence of the steady stable plankton-abundant state.

Mathematics Subject Classification (2000). Primary 92D25, Secondary 34D23.

Keywords. Plankton dynamics, Global stability, Direct Lyapunov methods, Nonlinear interaction, Lyapunov function.

1. Introduction

The issue of ecosystem stability has been a challenging problem for biologists and mathematicians for nearly a century. It had long been observed that population

AK is supported by Japan Society for Promotion of Science (JSPS) through Project 17540099.

size of ecological species can either remain approximately constant or experience fluctuations of considerable amplitude [3,25]. While stable species coexistence apparently corresponds to a self-sustained ecosystem functioning, theoretical considerations proved that population fluctuations/oscillations typically arise as a result of loss of stability of a corresponding steady state [12,23,28]. In the course of the system dynamics, the negative changes that caused the stability loss tend to increase the oscillation amplitude so that the oscillating species become prone to extinction due to the impact of stochastic factors [19]. Therefore, to identify the conditions of stable ecosystem functioning is a problem of highest theoretical and practical importance.

Mathematical consideration of population stability and population oscillations resulted in the seminal works by Lotka [22] and Volterra [35], which also marked appearance of mathematical ecology as a science. Further progress in understanding these issues has been made by May [24], Hofbauer and Sigmund [15], Takeuchi [33] and Bazykin [2]. In particular, it was shown that, in order to make a model biologically realistic, saturation in grazing/predation must be taken into account. In its turn, the effect of saturation, which results in a non-convex shape of the corresponding trophic function(s), increases the parameter range of the steady state stability.

A certain drawback of the previous studies is that in most cases they were essentially based on specific assumptions regarding the functional form of nonlinear feedbacks (e.g., bilinear in the Volterra model); see [2] for a comprehensive review. For a while this drawback has not been regarded as significant because of a widely spread intuitive expectation that a particular choice of parameterization is not important as far as the principal properties of the corresponding functions (such as monotonicity, convexity/concavity, etc.) remain the same. However, in a recent paper by Gross et al. [13], it was shown that this is not so and that a small perturbation of functional responses (changing only a sign of higher derivatives) can change the system stability dramatically. Therefore, of special value are the mathematical studies of ecosystem stability which are not based on specific choice of function(s).

In this paper, we address this issue using a conceptual model of marine ecosystem with the functional responses in a general form. Our choice of marine ecosystem as a paradigm is not accidental. Mathematical models of marine ecosystems have been attracting considerable attention over the last three decades. Marine ecosystems are among the most endangered in the world, especially in the coastal regions where anthropogenic impact is usually very high. On the other hand, mathematical modelling provides a convenient and effective research tool, especially for marine ecology where regular experimental study is usually very expensive and replicated experiments are often not possible at all.

In particular, phytoplankton plays a very important role in the dynamics of marine ecosystems. Apparently, it lies at the basis of the whole trophic chain and thus determines the ocean primary production. Also, phytoplankton can greatly affect water quality through “blooms” of certain toxic species [14]. Finally, there are indications that phytoplankton may contribute to climate changes at a global scale [7].

In its turn, phytoplankton abundance essentially depends on availability of nutrients. For that reason, a lot of attention has been paid to the properties of plankton-nutrient system. A generic mathematical model that describes the phytoplankton-nutrient interaction is given by the following equations:

$$\dot{N}(t) = \eta - h(N)P - bN, \tag{1.1}$$

$$\dot{P}(t) = \epsilon h(N)P - (\mu + c)P, \tag{1.2}$$

where $N(t)$ and $P(t)$ are the densities of nutrient and phytoplankton, respectively, at time t, η is the nutrient input rate (e.g., due to upwelling or river discharge), b and c are the washout rates for nutrient and phytoplankton respectively, μ is the phytoplankton mortality and ϵ is the nutrient consumption efficiency. Function $h(N)$ takes into account nonlinear effects in the nutrient uptake, e.g., saturation; an example is given by the Michaelis–Menten kinetics.

The system (1.1–1.2) as well as some of its generalizations have been studied in much detail [5, 6, 27, 29]. Surprisingly, however, practically all the work has been restricted to the case when the right-hand side of Eq. (1.2) is linear with respect to the phytoplankton density P (but see [10]). Meanwhile, over the last years there has been growing understanding that the phytoplankton growth rate should not be necessarily proportional to its density but can be affected by a variety of density-dependent processes; in particular, it can arise as a result of phytoplankton self-shading [4]. Thus, in a more general and biologically relevant case, nutrient consumption should be described as $h(N)g(P)$, rather than $h(N)P$, where $g(P)$ is a certain nonlinear function. In a still more general case, nutrient consumption may be given by a non-factorable function of N and P.

Also, the assumption that the plankton mortality rate is linear with respect to the plankton density P, cf. the last term in the right-hand side of Eq. (1.2), is rather restrictive and, in fact, does not always agree with experimental data. An increase in the population density normally leads to a decrease in the population multiplication rate due to effects of direct and indirect competition so that for a sufficiently large density (usually referred to as the population carrying capacity) the rate turns to zero [26]. Another reason for changing the linear term $-(\mu+c)P$ to a certain nonlinear function, say $c(P)$, is that, to be ecologically relevant, Eq. (1.2) should take into account the plankton grazing by its consumers that are not present in the model explicitly. The corresponding term in the equation is called a closure term and is essentially nonlinear [9, 32].

It should also be mentioned here that, mathematically, the model of phytoplankton-nutrient interaction is equivalent to a chemostat model which has been considered in much detail by several authors [11,16,21,30,31]. In particular, incapability of the models with density-independent feeding rate to adequately describe data on phytoplankton growth has been recognized [20]. However, models with nonlinear (density-dependent) feeding rate has not yet been considered and, correspondingly, the question of what can be the system's global properties in this case remains largely open.

In our paper, we provide a rigorous mathematical consideration of this problem. Specifically, we consider the global stability of the system and existence/stability of the steady states for a model of phytoplankton-nutrient interaction which is similar to (1.1–1.2) but where the nutrient uptake rate and the plankton mortality rate are described by unspecified functions, i.e., $\omega(N,P)$ and $c(P)$, respectively. For this rather general case, by means of constructing the Lyapunov function we derive sufficient conditions ensuring existence of a unique nontrivial steady state.

2. Model

We consider the following model of phytoplankton-nutrient interaction:

$$\dot{N}(t) \;=\; \eta - \omega(N,P) - bN, \qquad \dot{P}(t) \;=\; \epsilon\omega(N,P) - c(P) \tag{2.1}$$

where all variables and parameters are defined above.

In order to be biologically realistic, the nutrient uptake rate $\omega(N,P)$ and the mortality/washout rate $c(P)$ must be nonnegative for all values of their arguments and vanish if either N or P vanishes, i.e.,

$$\omega(N,P) \geq 0,\; c(P) \geq 0 \text{ for all } P,N > 0, \quad \omega(0,P) = \omega(N,0) = 0,\; c(0) = 0. \tag{2.2}$$

The hypotheses that are often made at this stage are that of monotonicity and convexity/concavity of functions $\omega(N,P)$ and $c(P)$. On the contrary, in order to keep the model as general as possible, in our analysis we do not impose that kind of restriction on $\omega(N,P)$ and $c(P)$. However, we do assume that functions $\omega(N,P)$ and $c(P)$ are continuous and differentiable for all $N,P \geq 0$. It is also natural to require that $\partial c(0)/\partial P > 0$ holds; otherwise for a very low phytoplankton density the phytoplankton life span tends to infinity. This later condition rules out such functions as $c = \mu P^n$.

It is readily seen that the non-negative quadrant of the NP plane is an invariant set of the system, and that, provided that $\omega(N,0) = 0$, the system (2.1) has a plankton-free equilibrium state $Q_0 = (N_0, P_0)$ where $N_0 = \eta/b$ and $P_0 = 0$. Apart from this plankton-free state, the system can have other positive "plankton-abundant" equilibrium states; the coordinate of these equilibrium states, if they exist, satisfy the equalities

$$\omega(N,P) + bN = \eta, \qquad \omega(N,P) = Bc(P) \tag{2.3}$$

where the notation $B = 1/\epsilon$ is introduced for convenience.

An issue of primary importance, which we are going to address with all mathematical rigor, is the conditions of existence and stability of the steady state(s) of the system. In particular, the question is how a change in these global properties can be quantified for unspecified functions $\omega(N,P)$ and $c(P)$. As we will show below, the properties of the system (2.1) depend crucially on the following value:

$$R_0 = \epsilon \, \frac{\partial\omega(N_0,0)}{\partial P} \bigg/ \frac{\partial c(0)}{\partial P} \, . \tag{2.4}$$

In the simplest case of bilinear nutrient uptake rate $\omega = \alpha PN$ and linear function $c(P) = \mu P$, Eq. (2.4) turns to $R_0 = \epsilon\alpha\eta/b\mu$, which coincides with the standard definition of the basic reproduction number in epidemiology [8, 34].

Now, we proceed to analysis of the global properties of the model (2.1).

3. Properties of the model

The following theorems address global properties of the system (2.1) such as existence and stability of equilibrium states.

3.1. Existence of positive equilibrium states

Theorem 3.1. *Let (i) the function $c(P)$ grow monotonically and (ii) $\omega(N,P)$ be monotonically growing with respect to N for all $P > 0$, and let*

$$\textit{(iii)} \lim_{P\to 0} \frac{\omega(N_0,P)}{\omega(N,P)} > 1 \quad \textit{for all } N \in (0, N_0).$$

Then, if $R_0 > 1$, there exist positive equilibrium states.

Proof. At a stationary state of the system, the equalities $bN = \eta - Bc(P)$ and $Bc(P) = \omega(N,P)$ hold. These equalities define, respectively, a negatively sloped line q_1 and a curve q_2 on the NP plane (Fig. 1). The equality $Bc(P) = \omega(N,P)$ defines also a function $N = f(P)$. If $\omega(N,P)$ is monotonically growing with respect to N, then the function $f(P)$ is defined and continuous for all $P > 0$. It is obvious (see Fig. 1) that if $N_* = f(0) \le N_0 = \eta/b$, then there is at least one point of intersection of the lines q_1 and q_2. The function $\omega(N,P)$ grows monotonically with respect to both N, and hence $N_0/N_* > 1$ if

$$1 < \lim_{P\to 0} \frac{\omega(N_0,P)}{\omega(N_*,P)} = \lim_{P\to 0} \frac{\omega(N_0,P)}{Bc(P)} = \epsilon \frac{\partial\omega(N_0,0)}{\partial P} \bigg/ \frac{\partial c(0)}{\partial P} = R_0 \, .$$

Thus, under assumptions (i–iii) of the theorem, $R_0 > 1$ is a sufficient condition to ensure existence of a steady plankton-abundant state. □

3.2. Stability and uniqueness of the positive equilibrium state

We assume now that the system has a positive equilibrium state $Q^* = (N^*, P^*)$ such that the equalities (2.3) hold. The properties of this equilibrium state are given by the following theorem.

Theorem 3.2. *Let*

$$\omega(N,P^*) < \omega(N^*,P^*) \quad \textit{for} \quad N < N^*, \ \textit{and}$$
$$\omega(N,P^*) > \omega(N^*,P^*) \quad \textit{for} \quad N > N^*, \tag{3.1}$$

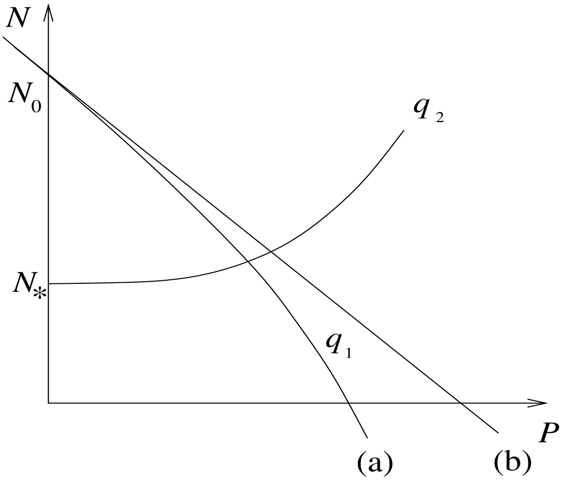


FIGURE 1. The lines q_1 and q_2 where the line (a) is for a function $c(P)$ with the rate of growth faster than linear, and the line (b) is for $c(P) = \mu P$. Note that, while the line q_1 always has a negative slope, the line q_2 can be ascending or descending. Importantly, however, even in the latter case, the line q_2 cannot cross the horizontal axis.

and let

$$\begin{aligned} &c(P)/c(P^*) \le \omega(N,P)/\omega(N,P^*) < 1 \quad \textit{for} \quad P < P^*, \ \textit{and} \\ &1 < \omega(N,P)/\omega(N,P^*) \le c(P)/c(P^*) \quad \textit{for} \quad P > P^* \end{aligned} \tag{3.2}$$

*hold for all $N > 0$. Then, if the system (*2.1*) has a positive equilibrium state $Q^* = (N^*, P^*)$, this positive equilibrium state is unique and globally asymptotically stable.*

Proof. In order to address the issue of stability, we consider a function

$$V(N,P) = N - \int_a^N \frac{\omega(N^*,P^*)}{\omega(x,P^*)}dx + B\left(P - \int_a^P \frac{c(P^*)}{c(x)}dx\right), \tag{3.3}$$

where a is a small unspecified parameter which is introduced here to avoid dealing with an improper integral; further we will direct a to zero.

This function is defined and continuous for all $N, P > a$. The function satisfies

$$\frac{\partial V}{\partial N} = 1 - \frac{\omega(N^*,P^*)}{\omega(N,P^*)}, \qquad \frac{\partial V}{\partial P} = B\left(1 - \frac{c(P^*)}{c(P)}\right), \tag{3.4}$$

and hence, by the theorem's hypotheses, $Q^* = (N^*, P^*)$ is the only stationary point of the function. Furthermore, since, by the theorem's hypotheses, $\omega(N,P)$ and $c(P)$ increase at Q^*, the point Q^* is the global minimum. Consequently, the function $V(N,P)$ is a Lyapunov function.

In the case of the system (2.1), using (2.3), the Lyapunov function (3.3) satisfies

$$\begin{aligned}
\frac{dV(N,P)}{dt} &= \eta - \omega(N,P) - bN - \eta\frac{\omega(N^*,P^*)}{\omega(N,P^*)} + \frac{\omega(N^*,P^*)}{\omega(N,P^*)}\omega(N,P) \\
&\quad + bN\frac{\omega(N^*,P^*)}{\omega(N,P^*)} + \omega(N,P) - Bc(P) \\
&\quad - \frac{c(P^*)}{c(P)}\omega(N,P) + Bc(P^*) \\
&= bN^*\left(1 - \frac{N}{N^*} - \frac{\omega(N^*,P^*)}{\omega(N,P^*)} + \frac{N}{N^*}\frac{\omega(N^*,P^*)}{\omega(N,P^*)}\right) \\
&\quad + \omega(N^*,P^*)\left(1 - \frac{\omega(N^*,P^*)}{\omega(N,P^*)} + \frac{\omega(N,P)}{\omega(N,P^*)}\right) \\
&\quad + \omega(N^*,P^*)\left(1 - \frac{c(P)}{c(P^*)} - \frac{c(P^*)}{c(P)}\frac{\omega(N,P)}{\omega(N^*,P^*)}\right) \\
&= bN^*\left(1 - \frac{N}{N^*}\right)\left(1 - \frac{\omega(N^*,P^*)}{\omega(N,P^*)}\right) \\
&\quad + \omega(N^*,P^*) \\
&\quad \times \left(3 - \frac{\omega(N^*,P^*)}{\omega(N,P^*)} - \frac{c(P^*)}{c(P)}\frac{\omega(N,P)}{\omega(N^*,P^*)} - \frac{c(P)}{c(P^*)}\frac{\omega(N,P^*)}{\omega(N,P)}\right) \\
&\quad + \omega(N^*,P^*)\left(\frac{c(P)}{c(P^*)} - \frac{\omega(N,P)}{\omega(N,P^*)}\right)\left(\frac{\omega(N,P^*)}{\omega(N,P)} - 1\right).
\end{aligned}$$

It is easy to see that $dV/dt \leq 0$ for all $N, P > 0$. Indeed,

$$\left(1 - \frac{N}{N^*}\right)\left(1 - \frac{\omega(N^*,P^*)}{\omega(N,P^*)}\right) \leq 0$$

and

$$\left(\frac{c(P)}{c(P^*)} - \frac{\omega(N,P)}{\omega(N,P^*)}\right)\left(\frac{\omega(N,P^*)}{\omega(N,P)} - 1\right) \leq 0$$

hold by the theorem's hypotheses. Furthermore,

$$\frac{\omega(N^*,P^*)}{\omega(N,P^*)} + \frac{c(P^*)}{c(P)}\frac{\omega(N,P)}{\omega(N^*,P^*)} + \frac{c(P)}{c(P^*)}\frac{\omega(N,P^*)}{\omega(N,P)} \geq 3$$

for $N, P > 0$, because the arithmetic mean is greater than or equal to the geometric mean (which is equal to 1 in this case).

We assume now that apart from the equilibrium Q^*, the system has another positive equilibrium state $Q_1 = (N_1, P_1)$. Then $\omega(N_1, P_1) + bN_1 = \eta$ and $Bc(P_1) = \omega(N_1, P_1)$ hold. The derivative of a Lyapunov function is equal to zero at any equilibrium state, and therefore $\frac{dV}{dt} = 0$ at Q_1. Thus, N_1 and P_1 must satisfy the equalities

$$\left(1-\frac{N_1}{N^*}\right)\left(1-\frac{\omega(N^*,P^*)}{\omega(N_1,P^*)}\right) = 0, \tag{3.5}$$

$$3-\frac{\omega(N^*,P^*)}{\omega(N_1,P^*)}-\frac{c(P^*)}{c(P_1)}\frac{\omega(N_1,P_1)}{\omega(N^*,P^*)}-\frac{c(P_1)}{c(P^*)}\frac{\omega(N_1,P^*)}{\omega(N_1,P_1)} = 0, \tag{3.6}$$

$$\left(\frac{c(P_1)}{c(P^*)}-\frac{\omega(N_1,P_1)}{\omega(N_1,P^*)}\right)\left(\frac{\omega(N^*,P^*)}{\omega(N^*,P_1)}-1\right) = 0. \tag{3.7}$$

By (3.1), the equality (3.5) holds only when $N_1 = N^*$. Then $c(P_1) = c(P^*)$ is necessary to satisfy (3.6). By (3.2), this means $P_1 = P^*$, and therefore Q^* is the only positive fixed point of the system, and $\frac{dV}{dt} = 0$ holds at the point Q^* only. Thus, by virtue of the Lyapunov–La Salle principle [1, 18], the point Q^* is asymptotically stable for all $P, N \geq a$. The parameter a may be made as small as required, and therefore the endemic equilibrium Q^* is globally asymptotically stable in the positive quadrant $\mathbf{R}^2_+$. □

3.3. Stability of the plankton-free equilibrium state for $R_0 \leq 1$

We have already mentioned that the plankton-free equilibrium $Q_0 = (\eta/b, 0)$ exists provided by $\omega(N,0) = 0$. It follows from theorems 3.1 and 3.2 that this equilibrium state is unstable (in fact, it is a saddle) when hypotheses of these theorems hold. The following theorem addresses the case when the hypothesis of Theorem 3.1 does not hold, namely when $R_0 \leq 1$.

Theorem 3.3. *Let*

$$\lim_{P\to 0}\frac{\omega(N_0,P)}{\omega(N,P)} > 1 \textit{ for } N < N_0, \textit{ and } \lim_{P\to 0}\frac{\omega(N_0,P)}{\omega(N,P)} \leq 1 \textit{ for } N \geq N_0, \tag{3.8}$$

and

$$\frac{\omega(N,P)}{c(P)} \leq \frac{\partial\omega(N,0)}{\partial P}\bigg/\frac{\partial c(0)}{\partial P} \quad \textit{for all } N, P > 0. \tag{3.9}$$

Then, if $R_0 \leq 1$, there is no positive equilibrium state, and the plankton-free equilibrium state Q_0 is globally asymptotically stable.

Proof. We consider a Lyapunov function

$$U(N,P) = N - \int_a^N \lim_{P\to 0}\frac{\omega(N_0,P)}{\omega(x,P)}dx + BP.$$

For a function $\omega(N,P)$ which is continuous with respect to both arguments the limit $\lim_{P\to 0}\frac{\omega(N_0,P)}{\omega(N,P)}$ is a well-defined and finite function of N, and hence this Lyapunov function is defined for all $N \geq a, P \geq 0$.

It is easy to verify that the point Q_0 is the global minimum of this function, and that this function is a Lyapunov function indeed. In the case of the system (2.1), the Lyapunov function satisfies

$$\begin{aligned}\frac{dU(N,P)}{dt} &= \eta - \omega(N,P) - bN - \eta \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} + \omega(N,P) \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} \\ &\quad + bN \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} + \omega(N,P) - Bc(P) \\ &= \eta - \eta \frac{N}{N_0} - \eta \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} \\ &\quad + \omega(N,P) \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} + \eta \frac{N}{N_0} \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} - Bc(P) \\ &= \eta \left(1 - \frac{N}{N_0}\right) \left(1 - \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)}\right) \\ &\quad + Bc(P) \left(\frac{\omega(N,P)}{Bc(P)} \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} - 1\right)\end{aligned}$$

(where we denote $\lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)} = \lim_{P\to 0} \frac{\omega(N_0,P)}{\omega(N,P)}$). By the theorem's hypothesis,

$$\left(1 - \frac{N}{N_0}\right) \left(1 - \lim \frac{\omega(N_0,P_0)}{\omega(N,P_0)}\right) \le 0 \quad \text{for all} \quad N > 0,$$

and

$$\begin{aligned}\frac{\omega(N,P)}{Bc(P)} \lim_{P\to 0} \frac{\omega(N_0,P)}{\omega(N,P)} &= \frac{\omega(N,P)}{Bc(P)} \left[\frac{\partial \omega(N_0,P_0)}{\partial P} \Big/ \frac{\partial \omega(N,P_0)}{\partial P}\right] \\ &= R_0 \frac{\omega(N,P)}{c(P)} \left[\frac{\partial c(0)}{\partial P} \Big/ \frac{\partial \omega(N,0)}{\partial P}\right] \le R_0.\end{aligned}$$

Therefore, $R_0 \le 1$ ensures that $dU(N,P)/dt \le 0$ for all $N, 0 > 0$, and hence by the asymptotic stability theorem [1, 18] the equilibrium state Q_0 is globally asymptotically stable in this case. □

Note that, if the hypothesis (3.2) does not hold for all $P > 0$ but holds for $N = N^*$ on some interval (P_1, P_2) such that $P \in (P_1, P_2)$, then Theorem 3.2 cannot ensure global stability of the system. However, in this case, by the Lyapunov–La Salle principle, the equilibrium state is locally asymptotically stable. It also follows from Theorem 3.2 that violation of the hypothesis in the vicinity of the equilibrium state is necessary for loss of stability.

It is readily seen that the theorem hypotheses hold if $\omega(\cdot, P)$ is a concave function and $c(P)$ is a convex function, see Fig. 2. However, these properties are not necessary: for instance, the functions shown in Fig. 3 satisfy Theorem 3.2 hypotheses and hence ensure the global stability of the positive equilibrium state.

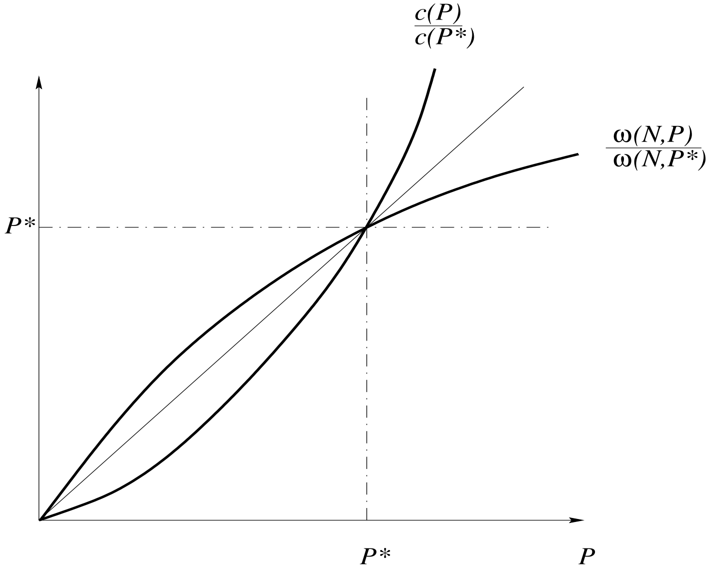


FIGURE 2. A function $\omega(N, P)$ non-convex with respect to P.

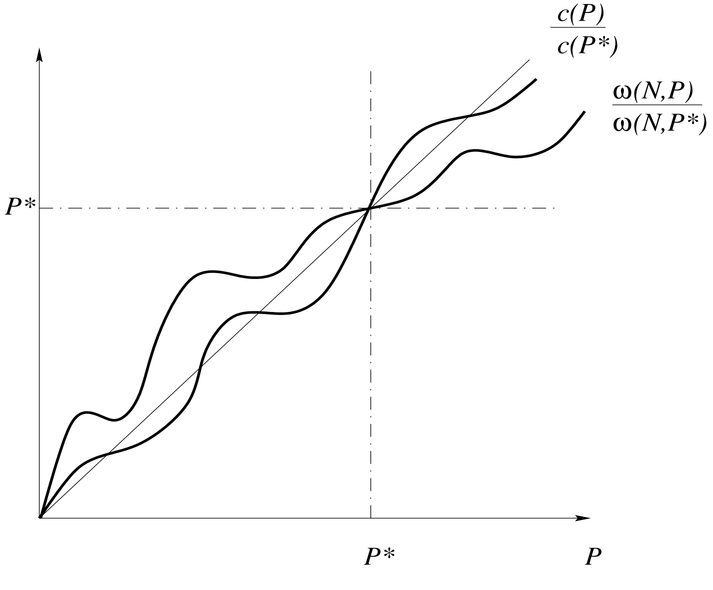


FIGURE 3. Concave/convex non-monotonic function $\omega(N, P)$ satisfying the condition (3.2).

4. Concluding remarks

In this paper, we considered a general two-species model of phytoplankton-nutrient interaction where the nutrient uptake rate is described by a function of two variables, $\omega(N, P)$, and the phytoplankton mortality/washout rate is given by an unspecified function $c(P)$; both these functions can be nonlinear with respect to their arguments. We showed that the global properties of the system depend on the properties of $\omega(N, P)$ and $c(P)$ through parameter R_0, cf. (2.4), so that a plankton-abundant steady state does not exist if $R_0 < 1$. In case $R_0 > 1$, for existence and stability of a unique plankton-abundant steady state it is only necessary that the nutrient uptake rate and the mortality/washout rate are described by functions that are increasing "in average", cf. conditions (3.1) and (3.2) and also see Fig. 3. Therefore, relations (3.1–3.2) along with $R_0 > 1$ describe the conditions of safe self-sustainable functioning of the plankton system.

In order to prove the above properties, we used the approach based on the direct Lyapunov method (3.3) which made the proof rather simple and elegant. It should be mentioned here that this approach has been proved to be an effective mathematical tool for studying various problems of mathematical biology [16, 17, 21]. In particular, the choice of the Lyapunov function is not a bottleneck: it is readily seen that, apart from (3.3), we could use either

$$V(N,P) = N - \int_a^N \frac{\omega(N^*,P^*)}{\omega(x,P^*)}dx + B\left(P - \int_a^P \frac{\omega(N^*,P^*)}{\omega(N^*,x)}dx\right)$$

or

$$V(N,P) = N - \int_a^N \frac{\omega(N^*,P^*)}{\omega(x,P^*)}dx + B\left(P - P^*\ln P\right),$$

with essentially the same outcome.

Note that Theorems 3.1 to 3.3 are proved under somewhat different hypotheses regarding the properties of functions $\omega(N, P)$ and $c(P)$. That was done in order to keep these hypotheses as non-restrictive as possible. However, it seems interesting and also useful for potential applications to understand whether the theorem assumptions can be somehow unified. It is not difficult to see that for a monotonically growing and convex function $c(P)$, and for a non-decreasing with respect to both arguments function $\omega(N, P)$, which is concave with respect to P, the hypotheses of all theorems hold automatically. Therefore, the following corollary takes place:

Corollary 4.1. *Let the nutrient uptake rate $\omega(N, P)$ be a non-decreasing function with respect to both arguments and non-convex with respect to P, and let the mortality/washout rate $c(P)$ be a monotonically growing and non-concave function. Then the system properties depend on the parameter R_0:*

(i) if $R_0 > 1$, then there is a unique and globally asymptotically stable positive (plankton-abundant) equilibrium state Q^*;

(ii) if $R_0 \leq 1$, then there is no positive equilibrium state Q^*, and the plankton-free equilibrium state Q_0 is globally asymptotically stable.

Note that concavity of $c(P)$ and convexity of $\omega(N, P)$, as well as monotonicity of both of these functions, are the properties that have immediate biological interpretation [26]. However, we want to emphasize that the actual conditions of the system stability are much weaker. In terms of ecological applications, it means that the limits of ecosystem stability may be significantly wider than they are usually thought to be.

Surprisingly, the global properties of the model do not depend on the details of the function $\omega(N, \cdot)$, e.g., whether it is concave or convex; monotonicity of $\omega(N, \cdot)$ is sufficient to ensure the global stability and the uniqueness of the equilibrium. This is a rather counter-intuitive result because different stability of equilibrium states in a resource-consumer system is often associated with a different shape of the function describing the consumer response, cf. Holling type II and type III . Furthermore, the monotonicity in a strict sense is not necessary either: in fact, for global stability it is only necessary that $\omega(N, \cdot) > \omega(N^*, \cdot)$ for all $N > N^*$, and $\omega(N, \cdot) < \omega(N^*, \cdot)$ for all $N < N^*$.

Our mathematical results seem to have important biological implications. It has long been a controversial issue how the system properties depend on a given parameterization of functional responses. It is indeed an important problem because, in a more applied study, population dynamics models are often used by means of numerical simulations and that, of course, implies a specific choice of functions. In particular, it has been shown recently [13] that, in some cases, systems with only small distinctions in the shape of the growth/uptake function exhibit essentially different stability. In contrast, we have shown that, in the chemostat-type model of the phytoplankton-nutrient interaction, existence and stability of the steady state is robust to the details of the uptake rate dependence on N and P, provided biologically reasonable properties of monotonicity and non-convexity are held.

An important inference can also be made regarding the type of mathematical model which is appropriate for studying the properties of plankton systems. In some recent studies, there has been a tendency to consider models which take into account the density-dependent higher order plankton mortality but neglect the linear one [9, 32]. That was partially based on a heuristic argument that nonlinear terms are likely to be more important to determine the properties of the system dynamics. However, it is immediately seen that the parameter R_0, which value is crucial for stability of the plankton-nutrient system (which is at the basis of any marine trophic chain), turns to infinity when the linear mortality vanishes, c.f. the lines below Eqs. (2.2). It indicates that the linear mortality is an important factor and can hardly be neglected without changing the system global properties significantly.

References

[1] E. A. Barbashin, *Introduction to the theory of stability.* Wolters-Noordhoff, 1970.

[2] A. D. Bazykin, *Nonlinear Dynamics of Interacting Populations.* World Scientific, 1998.

[3] A. A. Berryman, *Population Systems: A General Introduction.* Plenum Press, 1981.

[4] K. Boushaba, M. Pascual, *Dynamics of the 'echo' effect in a phytoplankton system with nitrogen fixation.* Bull. Math. Biol. **67** (2005), 487–507.

[5] G. Bromström, H. Drange, *On the mathematical formulation and parameter estimation of the Norwegian Sea plankton system.* Sarsia **85** (2000), 211–225.

[6] S. Busenberg, S. K. Kumar, P. Austin, G. Wake, *The dynamics of a model of a plankton-nutrient interaction.* Bull. Math. Biol. **52** (1990), 677–696.

[7] R. J. Charlson, J. E. Lovelock, M. O. Andreae, S. G. Warren, *Ocean phytoplankton, atmospheric sulphur, cloud albedo and climate.* Nature **326** (1987), 655–661.

[8] O. Diekmann, J. A. P. Heesterbek, J. A. J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations.* J. Math. Biol. **28** (1990), 365.

[9] A. M. Edwards, A. Yool, *The role of higher predation in plankton population models.* J. Plankton Res. **22** (2000), 1085–1112.

[10] G. T. Evans, J. S. Parslow, *A model of annual plankton cycles.* Biolog. Oceanogr. **3** (1985), 327–347.

[11] G. Fu, W. Ma, S. Ruan, *Qulitative analysis of a chemostat model with inhibitory exponential substrate uptake.* Chaos, Solitons and Fractals **23** (2005), 873–886.

[12] M. E. Gilpin, *Enriched predator-prey systems: theoretical stability.* Science **177** (1972), 902–904.

[13] T. Gross, W. Ebenhoh, U. Feudel, *Enrichment and foodchain stability: the impact of different forms of predator-prey interaction.* J. Theor. Biol. **227** (2004), 349–358.

[14] G. M. Hallegraeff, *A review of harmful algae blooms and the apparent global increase.* Phycologia **32** (1993), 79–99.

[15] J. Hofbauer, K. Sigmund, *The Theory of Evolution and Dynamical Systems.* Cambridge University Press, 1988.

[16] S. B. Hsu, *Limiting behavior for competing species.* SIAM J. Appl. Math. **34** (1978), 760–765.

[17] A. Korobeinikov, *Lyapunov functions and global stability for SIR and SIRS epidemiological models with non-linear transmission.* Bull. Math. Biol., **68** (2006), 615–626.

[18] J. La Salle, S. Lefschetz, *Stability by Liapunov's Direct Method.* Academic Press, 1961.

[19] R. Lande, *Risks of population extinction from demographic and environmental stochasticity and random catastrophes.* Amer. Nat. **142** (1993), 911–922.

[20] V. Lemesle, J.L. Gouze, *A biochemically based structured model for phytoplankton growth in the chemostat.* Ecological Coplexity **2** (2005), 21–33.

[21] B. Li, *Global asymptotic behavior of the chemostat: general response functions and different removal rates.* SIAM J. Appl. Math. **59** (1998), 411–422.

[22] A. J. Lotka, *Elements of Physical Biology.* Williams and Wilkins, 1925.

[23] R. M. May, *Limit cycles in predator-prey communities.* Science **177** (1972), 900–902.

[24] R. M. May, *Stability and Complexity in Model Ecosystems.* Princeton University Press, 1974.

[25] W. W. Murdoch, E. McCauley, *Three distinct types of dynamic behaviour shown by a single planktonic system.* Nature **316** (1985), 628–630.

[26] J. D. Murray, *Mathematical Biology.* Springer, 1989.

[27] O. Pardo, *Global stability for a phytoplankton-nutrient system.* J. Biol. Systems **8** (2000), 195–209.

[28] M. L. Rosenzweig, *Paradox of enrichment: destabilization of exploitation ecosystem in ecological time.* Science **171** (1971), 385–387.

[29] S. Ruan, *Persistence and coexistence in zooplankon-phytoplankton-nutrient models with instantaneous nutrient recycling.* J. Math. Biol. **31** (1993), 633–654.

[30] S. Ruan, X. Z. He, *Global stability in chemostat-type competition models with nutrient recycling.* SIAM J. Appl. Math. **58** (1998), 170–192.

[31] H. L. Smith, P. Waltman, *The Theory of the Chemostat: Dynamics of Microbial Competition.* Cambridge University Press, 1995.

[32] J. A. Steel, E. W. Henderson, *The role of predation in plankton models.* J. Plankton Res. **14** (1992), 157–172.

[33] Y. Takeuchi, *Global Dynamical Properties of Lotka-Volterra Systems.* World Scientific, 1996.

[34] P. van den Driessche, J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission.* Math. Biosci. **180** (2002), 29–48.

[35] V. Volterra, *Fluctuations in the abundance of a species considered mathematically.* Nature **118** (1926), 558–560.

Andrei Korobeinikov
Laboratory of Nonlinear Science and Computation
Research Institute for Electronic Science
Hokkaido University
Sapporo 060–0812
Japan
e-mail: `andrei@nsc.es.hokudai.ac.jp`

Sergei V. Petrovskii
Department of Mathematics
University of Leicester
Leicester, LE1 7RH
U.K.
e-mail: `sp237@le.ac.uk`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 41–63
© 2008 Birkhäuser Verlag Basel/Switzerland

Nutrient, Non-toxic Phytoplankton, Toxic Phytoplankton and Zooplankton Interaction in an Open Marine System

Nandadulal Bairagi, Samaresh Pal, Samrat Chatterjee and Joydev Chattopadhyay

Abstract. In this paper we propose a mathematical model for the interaction of nutrient, non-toxic phytoplankton, toxic phytoplankton and their predator zooplankton population in an open marine system. For a realistic representation of the open marine plankton ecosystem, we have incorporated various natural phenomena such as spatial flow, nutrient recycling, toxin effects, interspecies competition and grazing at a higher level. Nutrient–phytoplankton–zooplankton interactions are observed to be very complex and situation specific. Different exciting results, ranging from stable situation to cyclic blooms or monospecies bloom, may occur under different favourable conditions, which may give some insights for predictive management.

Mathematics Subject Classification (2000). 92D25, 92D40.

Keywords. Nutrient, non-toxic phytoplankton, toxic phytoplankton, zooplankton, spatial flow, nutrient recycling, oscillations, bloom.

1. Introduction

Phytoplankton are very small, usually single-celled organisms, chiefly diatoms, that photosynthesize and occupy the first trophic level in the marine food chain. Phytoplankton do a huge service for our Earth. Apart from food for marine life, they produce oxygen and also absorb half of the carbon dioxide that may contribute to global warming [10]. An algal bloom is characterized by a dramatic sharp increase in algae population numbers, up to several orders of magnitude [3]. Some blooms appear regularly every year (e.g., the classic Spring blooms), while others occur in an erratic fashion and may be sporadic both in time and space [18]. The dynamics

Work of S.Pal is supported by: UGC MRP NO. F.PSW-063/05-06(ERO).

of the rapid increase or decrease of plankton populations is therefore an important subject for marine plankton ecology.

Several studies have shown that a number of phytoplankton species have the ability to produce toxic substances that stun, kill or repel potential grazers [6, 9, 16, 17, 19–21, 37]. There has been a global increase in toxic or otherwise harmful plankton blooms in the last two decades [1, 15, 33], and considerable scientific attention has been paid to harmful algal blooms (HABs) in recent years [4, 35]. HABs are sometimes called red tides; they have adverse effects on human health, commercial fisheries, subsistence fisheries, recreational fisheries, tourism and coastal recreation, ecosystem and environment [2]. A broad classification of HAB species distinguishes two groups – viz. the toxin producers, which can contaminate seafood or kill fish, and high-biomass producers, which can cause anoxia and indiscriminate mortalities of marine life after reaching dense concentrations. Some HAB species have characteristics of both groups. Researchers have attempted to explain bloom phenomena in different ways. One group of researchers favours a "bottom-up" approach where, in recognition of the importance of nutrient to the growth of algae, the availability of nutrient is supposed to be one of the main regulatory factors for algal growth [8, 13, 36]. Another group of researchers believes that bloom is controlled by its grazers rather than nutrient [11, 27, 39, 40] in a "top-down" approach, and others [3, 31] have associated blooms with virus abundance. However, Chattopadhyay *et al.* [7] and Pal *et al.* [26] observed that toxin-producing phytoplankton may be responsible.

Harmful phytoplankton certainly play an important role. Reduction of the grazing pressure of the zooplankton due to the release of toxic substances by phytoplankton is evidently an important factor in this context [22]. Herbivore (zooplankton) grazing plays a crucial role in the initial stages of a red tide outbreak [41] and it has been shown that toxicity may be a strong mediator of the zooplankton feeding rate in both field [24] and laboratory studies [20, 25]. Thus toxic substances have a significant influence on the phytoplankton–zooplankton interactions, and play a most important role in the growth of the zooplankton population. In most previous modelling work, the non-toxic phytoplankton (NTP) and toxin producing phytoplankton (TPP) were considered as a single population. However, the TPP should be treated separately due to their distinctly different effects on their grazer, for a better appreciation of the phytoplankton–zooplankton interaction.

Nutrient is supposed to be one of most important factors that triggers bloom phenomena, and has been studied extensively by many researchers [5, 12, 14, 18, 28–30, 37, 42]. However, most of these nutrient-phytoplankton (NP) models that succeeded in modelling algae blooms assumed a closed system. Spatial flows of nutrients and organisms in an ecosystem are very important from an ecological point of view, so such flows should be considered in the model, to investigate their consequences for the community and population ecology.

In order to understand the ecosystem functioning better, we need to understand which factors are responsible for the initiation and rapid multiplication of algal numbers, what determines phytoplankton species composition and succession during blooms, how toxic substances released by the TPP influence the bloom dynamics, how spatial flows and recycling processes influence the ecosystem dynamics, and the interplay between them. With all that in mind, in this paper we propose a suitable mathematical model (involving the nutrient, NTP, TPP and their predator zooplankton) to investigate how the nutrient, spatial flows and toxin affect the functioning of the marine ecosystems. The organization of the paper is as follows: section 2 deals with the model formulation; our mathematical and numerical work is presented in sections 3 and 4, respectively; and a brief final discussion is presented in section 5.

2. The mathematical model

Let $N(t)$, $P_1(t)$, $P_2(t)$ and $Z(t)$ be the respective concentrations of nutrient, non-toxic phytoplankton (NTP), toxic phytoplankton (TPP) and zooplankton population at time t. Let N^0 be the constant input of nutrient concentration and D the constant dilution rate (its inverse $1/D$ has the physical dimension of a time and represents the average time that nutrient spends in the system [34]) and let D_1, D_2 and D_3 be the washout rates of the NTP, TPP and zooplankton population, respectively. Let α_1 and α_2 denote the respective nutrient uptake rates for the NTP and TPP, and $\theta_1(<\alpha_1)$ and $\theta_2(<\alpha_2)$ the corresponding conversion rates. The Michaelis–Menten uptake dynamics in general may provide a realistic modelling of the nutrient uptake dynamics, but here we simply assume that the nutrient uptake dynamics follow a linear mass action law. Let μ_1 and μ_2 be the respective mortality rates of the NTP and TPP, and let μ_3 be a parameter defining the total death rate of the zooplankton population — i.e., natural mortality plus possible grazing by higher trophic levels. Let η_1, η_2 and η_3 $(\eta_i < \mu_i, i = 1, 2, 3)$ be the nutrient recycle rates after the death of NTP, TPP and zooplankton population, respectively. For simplicity, we deliberately ignore the detailed dynamics of organic matter decomposition and nutrient recycling, which are encapsulated in a single parameter. The two phytoplankton populations compete for the same limiting resources, including nutrient and light. Let e_1 and e_2 represent the strength of the interspecies competition (e.g., e_1 is the amount by which one unit of species P_2 decreases the per capita growth rates of species P_1). Finally, let β_1 and β_2 be the respective maximal zooplankton ingestion rates of the NTP and TPP, and $\gamma_1(<\beta_1)$ the maximal zooplankton conversion rate. Liberation of toxic substances by the TPP causes substantial zooplankton mortality, and therefore reduces the growth rate of the zooplankton, and we let γ_2 denote the consequent rate at which growth rate of the zooplankton is reduced by the toxin substances.

We consequently formulate the following mathematical model:

$$\dot{\mathbf{X}} = \mathbf{F}(\mathbf{X}), \tag{2.1}$$

where $\mathbf{X} = (N,\ P_1,\ P_2,\ Z)^T \in R^4$ and $\mathbf{F}(\mathbf{X}) = [F_1(\mathbf{X}), F_2(\mathbf{X}), F_3(\mathbf{X}), F_4(\mathbf{X})]^T$, with $\mathbf{F} : C_+^4 \to R^4$ and $\mathbf{F} \in C^\infty(R^4)$; or in component form

$$\begin{aligned}
\frac{dN}{dt} &= D(N^0 - N) - \alpha_1 P_1 N - \alpha_2 P_2 N + \eta_1 P_1 + \eta_2 P_2 + \eta_3 Z \\
&\equiv F_1(N,\ P_1,\ P_2,\ Z), \\
\frac{dP_1}{dt} &= \theta_1 P_1 N - \beta_1 P_1 Z - e_1 P_1 P_2 - \mu_1 P_1 - D_1 P_1 \equiv F_2(N,\ P_1,\ P_2,\ Z), \\
\frac{dP_2}{dt} &= \theta_2 P_2 N - \beta_2 P_2 Z - e_2 P_1 P_2 - \mu_2 P_2 - D_2 P_2 \equiv F_3(N,\ P_1,\ P_2,\ Z), \\
\frac{dZ}{dt} &= \gamma_1 P_1 Z - \gamma_2 P_2 Z - \mu_3 z - D_3 Z \equiv F_4(N,\ P_1,\ P_2,\ Z). \qquad (2.2)
\end{aligned}$$

This system of ordinary differential equations is subject to the initial conditions

$$N(0) > 0,\ P_1(0) > 0,\ P_2(0) > 0,\ Z(0) > 0. \qquad (2.3)$$

3. Mathematical results

3.1. Positive invariance

From the initial conditions $\mathbf{X}(0) = \mathbf{X}_0 \in R_+^4$, it is easy to check that $F_i(x)\mid_{\mathbf{X}_i=0} \geq 0$. By Nagumo's lemma, any solution of equation (2.1) with $\mathbf{X}_0 \in R_+^4$ is such that $\mathbf{X}(t) \in {R_+}^4$ for all $t > 0$ [23].

3.2. Boundedness of the system

All the solutions of (2.2) are ultimately bounded (cf. Appendix).

3.3. Equilibrium points and their stability properties

The system (2.2) possesses the following equilibrium points:

- The plankton free equilibrium $\mathbf{E}_0 = (N^0, 0, 0, 0)$, which always exists. In addition, if

  $$N^0 < \min\{\frac{\mu_1 + D_1}{\theta_1}, \frac{\mu_2 + D_2}{\theta_2}\},$$

  then the plankton free steady state $\mathbf{E}_0$ is asymptotically stable; otherwise it is unstable.

- The TPP and zooplankton free equilibrium $\mathbf{E}_1 = (N^{(1)}, P_1^{(1)}, 0, 0)$ with

  $$N^{(1)} = \frac{\mu_1 + D_1}{\theta_1} \text{ and } P_1^{(1)} = \frac{D[\theta_1 N^0 - (\mu_1 + D_1)]}{\alpha_1(\mu_1 + D_1) - \theta_1 \eta_1}.$$

  Since $\alpha_1 > \theta_1$ and $\mu_1 > \eta_1$ so that $(\alpha_1(\mu_1 + D_1) - \theta_1\eta_1)$ is always positive, $P_1^{(1)}$ exists if $N^0 > (\mu_1 + D_1)/\theta_1$ and hence $\mathbf{E}_1$ exists if $N^0 > (\mu_1 + D_1)/\theta_1$.

  If

  $$\max\{R_1, \frac{\mu_1 + D_1}{\theta_1}\} < N^0 < \min\{R_2, \frac{\mu_2 + D_2}{\theta_2}\},$$

where

$$R_1 = \frac{(\alpha_1\mu_1+\alpha_1 D_1-\theta_1\eta_1)[\theta_2(\mu_1+D_1)-\theta_1(\mu_2+D_2)]}{e_2 D{\theta_1}^2}+\frac{\mu_1+D_1}{\theta_1},$$

$$R_2 = \frac{(D_3+\mu_3)(\alpha_1\mu_1+\alpha_1 D_1-\theta_1\gamma_1)}{\gamma_1\theta_1 D}+\frac{D_1+\mu_1}{\theta_1},$$

then $\mathbf{E}_1$, is asymptotically stable, but unstable otherwise.

- The NTP and zooplankton free equilibrium $\mathbf{E}_2(N^{(2)},0,P_2^{(2)},0)$ with

$$N^{(2)}=\frac{\mu_2+D_2}{\theta_2} \text{ and } P_2^{(2)}=\frac{D[\theta_2 N^0-(\mu_2+D_2)]}{\alpha_2(\mu_2+D_2)-\theta_2\eta_2}.$$

Therefore $\mathbf{E}_2$ exists if $N^0>(\mu_2+D_2)/\theta_2$. If

$$\max\{R_3,\frac{\mu_2+D_2}{\theta_2}\}<N^0<\frac{\mu_1+D_1}{\theta_1},$$

where

$$R_3=\frac{(\alpha_2\mu_2+\alpha_2 D_2-\theta_2\eta_2)[\theta_1(\mu_2+D_2)-\theta_2(\mu_1+D_1)]}{e_1 D{\theta_2}^2}+\frac{\mu_2+D_2}{\theta_2},$$

then $\mathbf{E}_2$ is asymptotically stable, but unstable otherwise.

- The zooplankton free equilibrium $\mathbf{E}_3(N^{(3)},P_1^{(3)},P_2^{(3)},0)$ with

$$P_1^{(3)}=\frac{\theta_2 N^{(3)}-(\mu_2+D_2)}{e_2}, \qquad P_2^{(3)}=\frac{\theta_1 N^{(3)}-(\mu_1+D_1)}{e_1}$$

and $N^{(3)}$ given by $A{(N^{(3)})}^2+BN^{(3)}+C=0$ where

$$\begin{aligned} A &= \alpha_1 e_1\theta_2+\alpha_2 e_2\theta_1>0,\\ B &= e_1 e_2 D-(\mu_2+D_2)\alpha_1 e_1-(\mu_1+D_1)\alpha_2 e_2-\eta_1 e_1\theta_2-\eta_2 e_2\theta_1,\\ C &= -e_1 e_2 D N^0+\eta_1 e_1(\mu_2+D_2)+\eta_2 e_2(\mu_1+D_1),\end{aligned}$$

which exists if

$$N^{(3)}>\max\{\frac{\mu_1+D_1}{\theta_1},\frac{\mu_2+D_2}{\theta_2}\}.$$

Note that if $C<0$ the above quadratic in $N^{(3)}$ has a unique positive real root. If

$$R_4=\frac{\eta_1 e_1(\mu_2+D_2)+\eta_2 e_2(\mu_1+D_1)}{e_1 e_2 D}<N^0,$$

then $\mathbf{E}_3$ is always an unstable saddle point.

- The toxic phytoplankton free equilibrium $\mathbf{E}_4=(N^{(4)},P_1^{(4)},0,Z^{(4)})$ with

$$N^{(4)}=\frac{DN^0\beta_1\gamma_1-\eta_3\gamma_1(\mu_1+D_1)+\eta_1\beta_1(\mu_3+D_3)}{\beta_1\gamma_1 D+\alpha_1\beta_1(\mu_3+D_3)-\eta_3\gamma_1\theta_1},\ P_1^{(4)}=\frac{\mu_3+D_3}{\gamma_1}$$

and

$$Z^{(4)} = \frac{DN^0\gamma_1\theta_1 + \eta_1\theta_1(\mu_3 + D_3) - (\mu_1 + D_1)\gamma_1 D - (\mu_1 + D_1)\alpha_1(\mu_3 + D_3)}{\beta_1\gamma_1 D + \alpha_1\beta_1(\mu_3 + D_3) - \eta_3\gamma_1\theta_1},$$

which exists if

$$DN^0\beta_1\gamma_1 - \eta_3\gamma_1(\mu_1 + D_1) + \eta_1\beta_1(\mu_3 + D_3) > 0$$
$$\beta_1\gamma_1 D + \alpha_1\beta_1(\mu_3 + D_3) - \eta_3\gamma_1\theta_1 > 0$$

and

$$DN^0\theta_1\gamma_1 + \eta_1\theta_1(\mu_3 + D_3) - \gamma_1 D(\mu_1 + D_1) - (\mu_1 + D_1)\alpha_1(\mu_3 + D_3) > 0.$$

The steady state $\mathbf{E}_4$ is an unstable saddle point if

$$R_5 = \frac{A_5}{B_5} > 1 \qquad \text{or} \qquad R_6 = \frac{A_6}{B_6} < 1,$$

where

$$\begin{aligned}
A_5 &= \theta_2\gamma_1[DN^0\beta_1\gamma_1 - \eta_3\gamma_1(\mu_1 + D_1) + \eta_1\beta_1(\mu_3 + D_3)],\\
B_5 &= \beta_2\gamma_1[DN^0\gamma_1\theta_1 + \eta_1\theta_1(\mu_3 + D_3) - (\mu_1 + D_1)\gamma_1 D\\
&- (\mu_1 + D_1)\alpha_1(\mu_3 + D_3)]\\
&+ e_2(\mu_3 + D_3)[D\beta_1\gamma_1 + \alpha_1\beta_1(\mu_3 + D_3) - \eta_3\gamma_1\theta_1]\\
&+ \gamma_1(\mu_2 + D_2)[D\beta_1\gamma_1 + \alpha_1\beta_1(\mu_3 + D_3) - \eta_3\gamma_1\theta_1],\\
A_6 &= [D\gamma_1 + \alpha_1(\mu_3 + D_3)](\alpha_1\beta_1 DN^0 + \eta_1\eta_3\theta_1)\\
&+ [DN^0\gamma_1\theta_1 + \eta_1\theta_1(\mu_3 + D_3)]\eta_3\gamma_1 \qquad \text{and}\\
B_6 &= [D\gamma_1 + \alpha_1(\mu_3 + D_3)](\alpha_1\eta_3(\mu_1 + D_1) + \eta_1\beta_1 D)\\
&+ \eta_3\gamma_1[(\mu_1 + D_1)\gamma_1 D + \alpha_1(\mu_1 + D_1)(\mu_3 + D_3)].
\end{aligned}$$

- The positive interior equilibrium $\mathbf{E}^* = (N^*, {P_1}^*, {P_2}^*, Z^*)$ with

$${P_2}^* = \frac{\gamma_1{P_1}^* - (\mu_3 + D_3)}{\gamma_2},\ Z^* = \frac{C_1 + {P_1}^* D_4}{\gamma_2(\theta_1\beta_2 - \theta_2\beta_1)},\ N^* = A_2/B_2,$$

where $A_2 = A_1 + {P_1}^* B_4$ and

$$\begin{aligned}
A_1 &= DN^0\gamma_2(\theta_1\beta_2 - \theta_2\beta_1) - \eta_2(\mu_3 + D_3)(\theta_1\beta_2 - \theta_2\beta_1)\\
&+ \eta_3\gamma_2\theta_2(\mu_1 + D_1) - \eta_3\gamma_2\theta_1(\mu_2 + D_2) - \eta_3 e_1\theta_2(\mu_3 + D_3),\\
B_2 &= (\theta_1\beta_2 - \theta_2\beta_1)[{P_1}^*(\alpha_1\gamma_2 + \alpha_2\gamma_1) - \alpha_2(\mu_3 + D_3) + D\gamma_2],\\
B_4 &= \eta_1\gamma_2(\theta_1\beta_2 - \theta_2\beta_1) + \eta_2\gamma_1(\theta_1\beta_2 - \theta_2\beta_1) + \eta_3(\theta_2\gamma_1 e_1 - \theta_1 e_2\gamma_2),\\
C_1 &= \gamma_2\theta_2(\mu_1 + D_1) - \gamma_2\theta_1(\mu_2 + D_2) - \theta_2 e_1(\mu_3 + D_3),\\
D_4 &= \theta_2 e_1\gamma_1 - \theta_1 e_2\gamma_2;
\end{aligned}$$

and ${P_1}^*$ is given by $\quad G_1{P_1}^{*2}+G_2{P_1}^*+G_3=0 \quad$ where

$$\begin{aligned}
G_1 &= -\beta_2 D_4(\alpha_1\gamma_2+\alpha_2\gamma_1)-e_2\gamma_2(\theta_1\beta_2-\theta_2\beta_1)(\alpha_1\gamma_2+\alpha_2\gamma_1),\\
G_2 &= \theta_2\gamma_2 B_4-\beta_2 C_1(\alpha_1\gamma_2+\alpha_2\gamma_1)+\alpha_2\beta_2 D_4(\mu_3+D_3)-\beta_2\gamma_2 DD_4\\
&+ e_2\gamma_2\alpha_2(\theta_1\beta_2-\theta_2\beta_1)(\mu_3+D_3)-e_2D{\gamma_2}^2(\theta_1\beta_2-\theta_2\beta_1)\\
&- \gamma_2(\mu_2+D_2)(\theta_1\beta_2-\theta_2\beta_1)(\alpha_1\gamma_2+\alpha_2\gamma_1),\\
G_3 &= \theta_2\gamma_2 A_1+\beta_2\alpha_2 C_1(\mu_3+D_3)-\beta_2\gamma_2 C_1 D\\
&+ \alpha_2\gamma_2(\mu_2+D_2)(\mu_3+D_3)(\theta_1\beta_2-\theta_2\beta_1)\\
&- D_1{\gamma_2}^2(\mu_2+D_2)(\theta_1\beta_2-\theta_2\beta_1).
\end{aligned}$$

This positive interior equilibrium $\mathbf{E}^*$ is feasible if $N^*, {P_1}^*,\ {P_2}^*, Z^*>0$ and we have the following conditions:

$$\max\{L_1,L_2\}<{P_1}^*<\min\{L_3,L_4\} \qquad \text{and} \qquad \frac{\beta_1}{\beta_2}<\frac{\theta_1}{\theta_2}<\frac{e_1\gamma_1}{e_2\gamma_2}\ ,$$

where

$$L_1=\frac{\mu_3+D_3}{\gamma_1},\quad L_2=\frac{-C_1}{D_4},\quad L_3=\frac{\alpha_2(\mu_3+D_3)-D\gamma_2}{\alpha_1\gamma_2+\alpha_2\gamma_1},\quad L_4=\frac{-A_1}{B_4};$$

and if $\mathbf{E}^*$ exists, it is asymptotically stable if

$$e_1\gamma_1\beta_2-e_2\beta_1\gamma_2>0 \tag{3.1}$$

$$N^*>\frac{\theta_1\eta_1\alpha_2+\theta_1\eta_2 e_2+\theta_2\eta_2\alpha_1+\theta_2\eta_1 e_1-\theta_1\alpha_1\eta_2-\theta_2\alpha_2\eta_1}{\theta_1\alpha_2 e_2+\theta_2\alpha_1 e_1} \tag{3.2}$$

and

$$\begin{aligned}
D_5 &= [-T_1(\beta_2\gamma_2{P_2}^*Z^*+e_1e_2{P_1}^*{P_2}^*-\beta_1\gamma_1{P_1}^*Z^*)\\
&- {P_1}^*{P_2}^*Z^*(e_1\gamma_1\beta_2-e_2\beta_1\gamma_2)\\
&- \theta_1{P_1}^*(\alpha_2e_2N^*{P_2}^*-\eta_2e_2{P_2}^*+\eta_3\gamma_1Z^*)\\
&+ \theta_2{P_2}^*(-\alpha_1e_1N^*{P_1}^*+\eta_1e_1{P_1}^*+\eta_3\gamma_2Z^*)]\\
&\times [\theta_1{P_1}^*[\alpha_1(DN^0+\eta_2{P_2}^*+\eta_3Z^*)-D\eta_1-\eta_1\alpha_2{P_2}^*+\alpha_2e_2N^*{P_2}^*\\
&- \eta_2e_2{P_2}^*+\eta_3\gamma_1Z^*]\\
&+ \theta_2{P_2}^*[\alpha_2(DN^0+\eta_1{P_1}^*+\eta_3Z^*)-D\eta_2-\eta_2\alpha_1{P_1}^*+\alpha_1e_1N^*{P_1}^*\\
&- \eta_1e_1{P_1}^*-\eta_3\gamma_2Z^*]+{P_1}^*{P_2}^*Z^*(e_1\gamma_1\beta_2-e_2\beta_1\gamma_2)]\\
&- {T_1}^2{P_1}^*{P_2}^*Z^*[T_1(-e_1\beta_2\gamma_1+e_2\beta_1\gamma_2)\\
&+ \theta_1(-\alpha_1\beta_2\gamma_2N^*+\eta_1\beta_2\gamma_2-\alpha_2\beta_2\gamma_1N^*+\eta_2\beta_2\gamma_1-\eta_3e_2\gamma_2)\\
&+ \theta_2(\alpha_1\beta_1\gamma_2N^*-\eta_1\beta_1\gamma_2+\alpha_2\beta_1\gamma_1N^*-\eta_2\beta_1\gamma_1+\eta_3e_1\gamma_1)]>0,
\end{aligned} \tag{3.3}$$

where

$$T_1=D+\alpha_1{P_1}^*+\alpha_2{P_2}^*>0.$$

Proof of the stability of the various equilibria is contained in the Appendix.

TABLE 1. Fixed set of Parameter Values

Parameter	Definition	Default value	Unit
N^0	Constant input of nutrient	1.58	h^{-1}
D	Dilution rate of nutrient	0.3	h^{-1}
α_1	Nutrient uptake rate of NTP	0.03	$ml.\ h^{-1}$
α_2	Nutrient uptake rate of TPP	0.022	$ml.\ h^{-1}$
θ_1	Conversion rate of NTP	0.02	$ml.\ h^{-1}$
θ_2	Conversion rate of TPP	0.02	$ml.\ h^{-1}$
μ_1	Death rate of NTP	0.006	h^{-1}
μ_2	Death rate of TPP	0.006	h^{-1}
μ_3	Death rate of zooplankton	0.005	h^{-1}
η_1	Nutrient recycling rate of NTP	0.004	$mg.\ h^{-1}$
η_2	Nutrient recycling rate of TPP	0.004	$mg.\ h^{-1}$
η_3	Nutrient recycling rate of zoopl.	0.0035	$mg.\ h^{-1}$
e_1	Competition coefficient	0.02	$ml.\ h^{-1}$
e_2	Competition coefficient	0.02	$ml.\ h^{-1}$
β_1	Predation rate of NTP	0.02	$ml.\ h^{-1}$
β_2	Predation rate of TPP	0.01	$ml.\ h^{-1}$
γ_1	Conversion rate for NTP	0.01	$ml.\ h^{-1}$
γ_2	Death rate due to consumption of TPP	0.008	$ml.\ h^{-1}$
D_1	Dilution rate of NTP	0.0004	h^{-1}
D_2	Dilution rate of TPP	0.0004	h^{-1}
D_3	Dilution rate of zooplankton	0.0003	h^{-1}

4. Numerical simulations

In this section, we numerically investigate the effect of the various parameters on the qualitative behaviour of the system using the parameter values given in Table 1 throughout, unless otherwise stated.

We first observe that the system (2.2) exhibits periodic behaviour, representing a recurring planktonic bloom (cf. Figure 1).

4.1. Effects of nutrient

If the constant input rate N^0 is varied but all other parameter values are unaltered, there are various effects on the plankton dynamics. If the nutrient input rate is decreased from 1.58 to 1.5, the system becomes stable around its interior equilibrium point, following oscillatory behaviour (cf. Figure 2a). If we further decrease the nutrient to 1.4, the system stabilizes to the equilibrium point $\mathbf{E}_4$ where the TPP is absent (cf. Figure 2b).

However, when the nutrient input is very low, the TPP and zooplankton free equilibrium $\mathbf{E}_1$ persists in a stable state (cf. Figure 3a). Further decrease in the nutrient input forces all plankton populations to extinction, and $\mathbf{E}_0$ becomes stable (cf. Figure 3b).

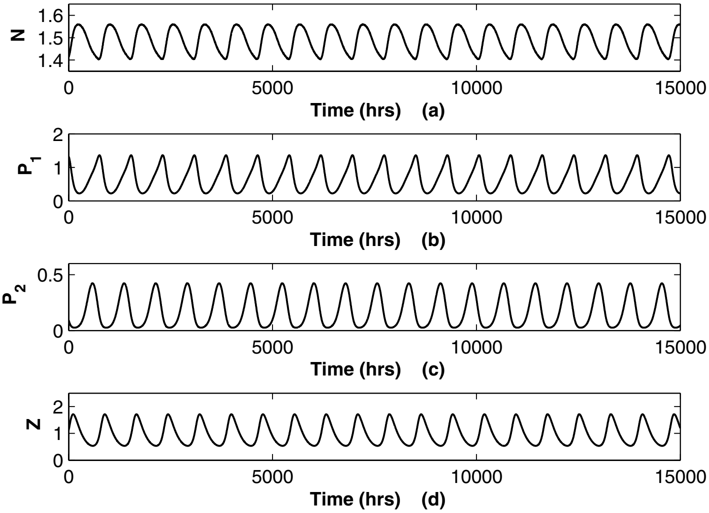


FIGURE 1. The periodic solution of the system (2.2) for the parameters given in Table 1.

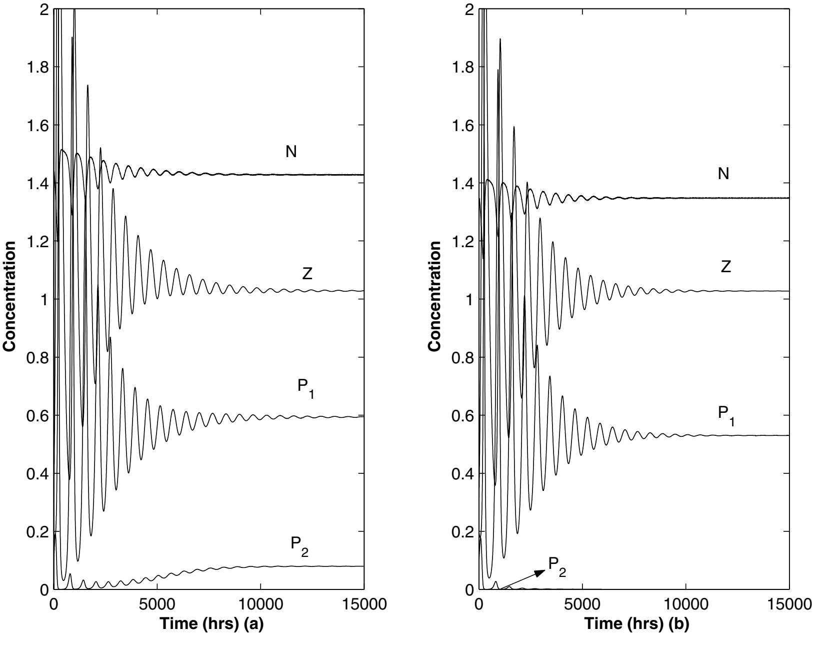


FIGURE 2. Time series solutions of the system (2.2). (a) For $N^0 = 1.5$, the system is stable around $\mathbf{E}^*$. (b) For $N^0 = 1.4$, the system is stable around $\mathbf{E}_4$.

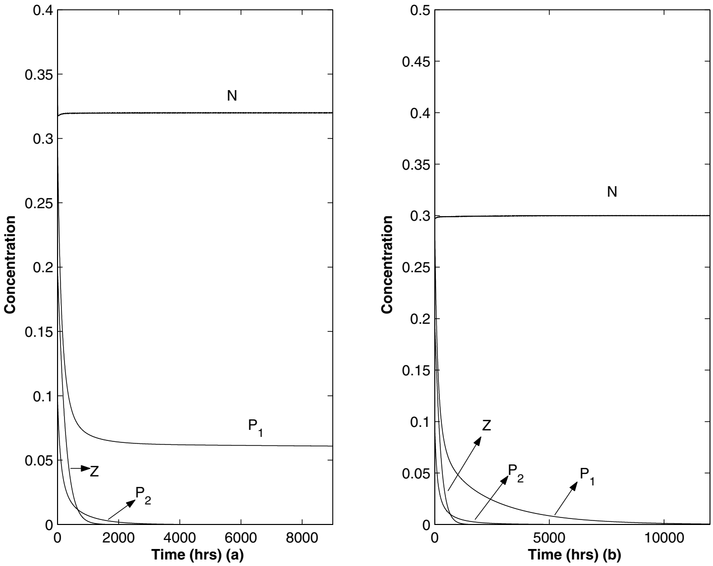


FIGURE 3. Time series solutions of the system (2.2). (a) For $N^0 = 0.321$, the system is stable around $\mathbf{E}_1$. (b) For $N^0 = 0.3$, the system is stable around $\mathbf{E}_0$.

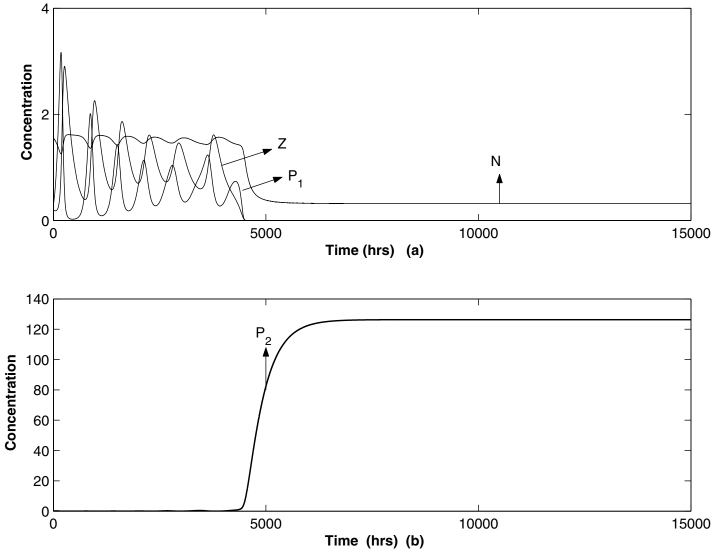


FIGURE 4. These figures show that the equilibrium point E_2 is locally asymptotically stable with high biomass of TPP for $N^0 = 1.6$. (a) Concentrations of nutrient, NTP and zooplankton. (b) Concentration of TPP.

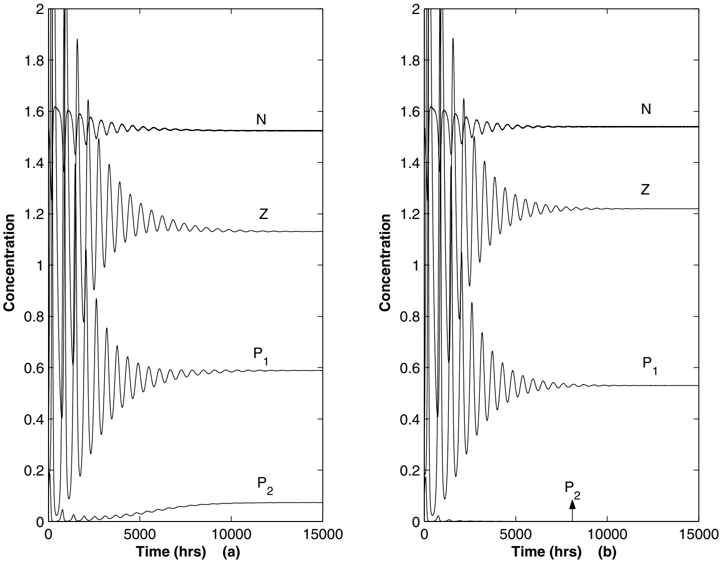


FIGURE 5. Time series solutions of the system (2.2) for $N^0 = 1.6$. (a) The stable interior equilibrium $\mathbf{E}^*$ when $\mu_2 = 0.007$. (b) The stable TPP free equilibrium when $\mu_2 = 0.008$.

Now, if we increase the nutrient input from 1.58 to 1.6, then $\mathbf{E}_2$ becomes stable with a high TPP biomass (cf. Figure 4). At this stage, if we increase the death rate of TPP from $\mu_2 = 0.006$ to $\mu_2 = 0.007$, we see that the system stabilizes to $\mathbf{E}^*$ from $\mathbf{E}_2$ (cf. Figure 5a), and any further increase in the death rate of TPP ($\mu_2 = 0.008$) forces the system to stabilize at $\mathbf{E}_4$ (cf. Figure 5b). However, if we increase μ_1, the death rate of NTP, from 0.006 to 0.008 keeping other parameters unchanged, the system reaches $\mathbf{E}_2$ with a high TPP biomass in a shorter time (cf. Figure 6).

4.2. Effects of interspecies competition

Suppose that the NTP is a stronger competitor than the TPP – i.e. $e_2 > e_1$. Assigning $e_2 = 0.0215$ (and the other parameters as in Table 1), $\mathbf{E}^*$ becomes stable from an oscillatory condition (cf. Figure 7a). If we further increase e_2 to 0.025, the stable equilibrium switches to $\mathbf{E}_4$ from $\mathbf{E}^*$ (cf. Figure 7b). However, if the TPP is a stronger competitor than the NTP (i.e., $e_1 > e_2$), the oscillatory coexistence of all the species is replaced by the stable state $\mathbf{E}_2$ with a high TPP biomass (cf. Figure 8).

4.3. Combined effects of nutrient and interspecies competition

We have already noted that the system stabilizes to $\mathbf{E}_2$ with a high TPP biomass when the nutrient input is high (viz. for $N^0 = 1.6$). If the value of e_2 is increased to 0.0215, then the system again stabilizes to $\mathbf{E}^*$ (cf. Figure 9a). For even higher

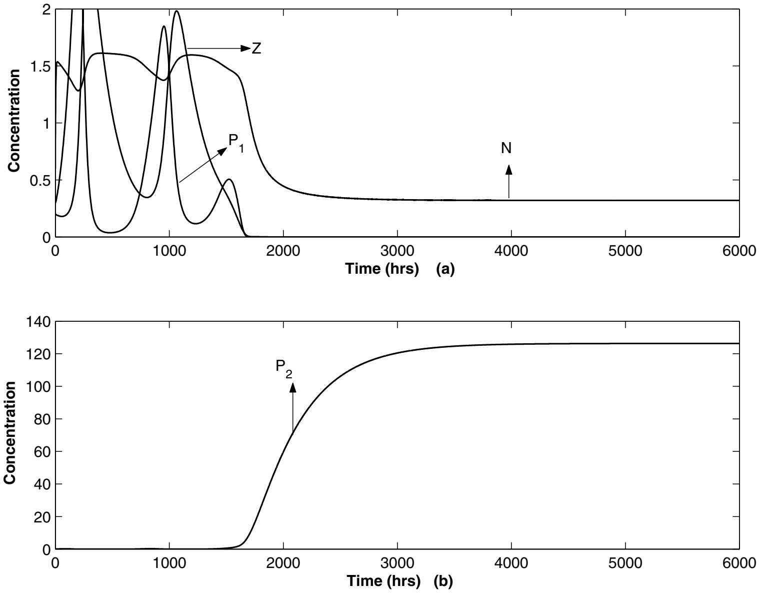


FIGURE 6. Locally asymptotically stable equilibrium point E_2 with high TPP biomass, for $N^0 = 1.6$ and $\mu_1 = 0.008$.

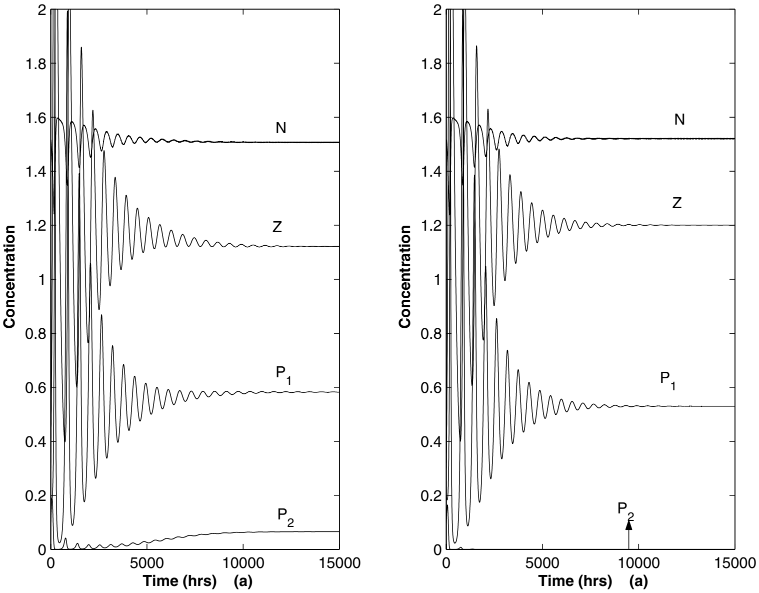


FIGURE 7. Solutions of the system (2.2) for $e_2 = 0.0215$ (cf. Figure a) and $e_2 = 0.025$ (cf. Figure b), with other parameters as given in Table 1.

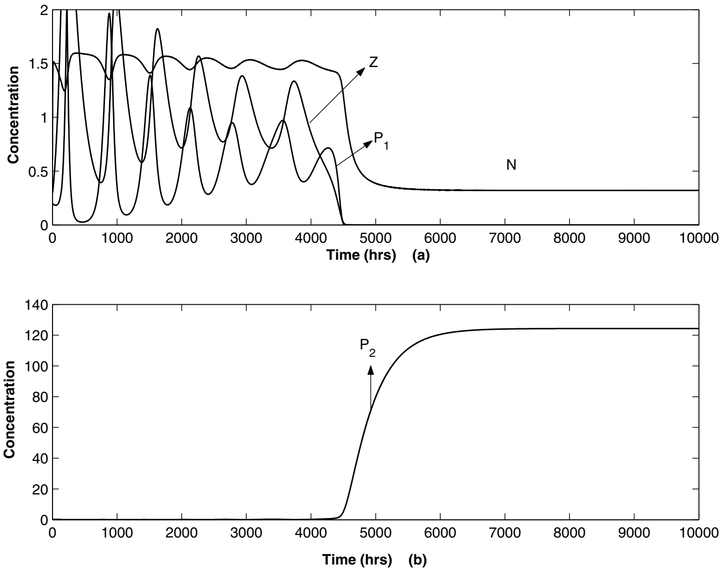


FIGURE 8. Time series solution of the system (2.2) for $e_1 = 0.022$, with other parameters as in Table 1.

e_2 ($= 0.025$), the system stabilizes at the equilibrium $\mathbf{E}_4$ (cf. Figure 9b). However, if we increase e_1 instead of e_2 from 0.02 to 0.025, then the system rapidly stabilizes at $\mathbf{E}_2$ with a high TPP biomass (cf. Figure 8).

4.4. Effects of dilution

Increased dilution rates of NTP, TPP and zooplankton dampen the oscillation and stabilise the system to $\mathbf{E}^*$ (cf. Figure 10a). If the nutrient input is increased from $N^0 = 1.58$ to $N^0 = 1.6$ with high dilution rates, the system still remains stable (cf. Figure 10b), so the system can tolerate more nutrient when the dilution rates are high. However, if we further increase N^0 from 1.6 to 1.7, then $\mathbf{E}_2$ becomes stable with a TPP high biomass (not shown), so the dilution rate can regulate monospecies bloom phenomena.

4.5. Effects of zooplankton death rate

Variation in the death rate of the zooplankton may have multiple effects. When μ_3 is increased from 0.005 to 0.0053, then $\mathbf{E}^*$ becomes asymptotically stable (cf. Figure 11a), but a further increase in μ_3 (say to $\mu_3 = 0.006$) forces the system to stabilize at $\mathbf{E}_4$ (cf. Figure 11b). However, under favourable conditions grazers at a higher level may increase the zooplankton mortality drastically, and this may cause an NTP monospecies bloom (cf. Figure 12). On the other hand, a TPP monospecies bloom may occur if the death rate of the zooplankton is decreased for some reason, and the qualitative behaviour of the system resembles that of Figure 6.

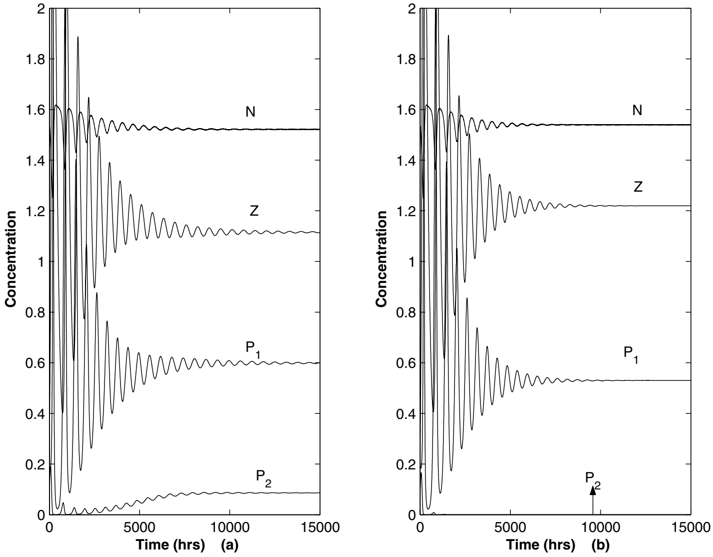


FIGURE 9. Time series solutions of the system (2.2) for $N^0 = 1.6$. (a) The interior equilibrium $\mathbf{E}^*$ is stable for $e_2 = 0.0215$. (b) The TPP free equilibrium $\mathbf{E}_4$ is stable for $e_2 = 0.025$. (The other parameters are as given in Table 1).

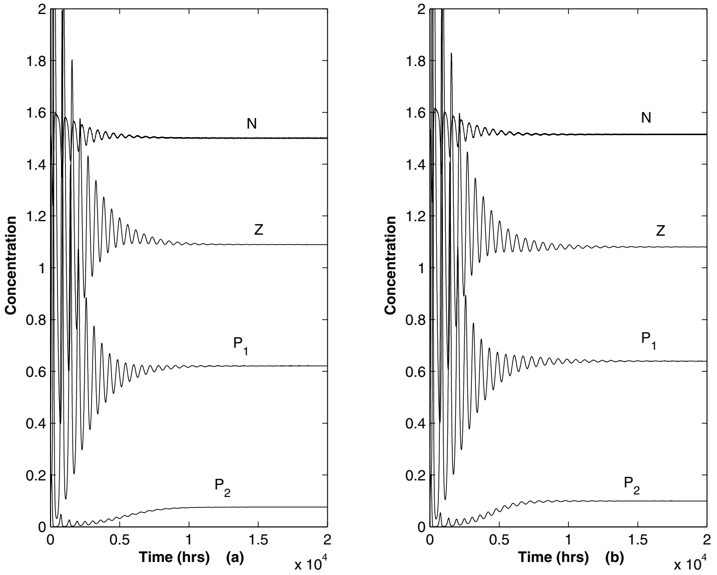


FIGURE 10. Time series solutions of the system (2.2) for $d_1 = 0.0007 = d_2$, $d_3 = 0.0006$. The interior equilibrium $\mathbf{E}^*$ is stable for $N^0 = 1.58$ as shown in (a), and remains stable for $N^0 = 1.6$ as shown in (b). (The other parameters are as given in Table 1).

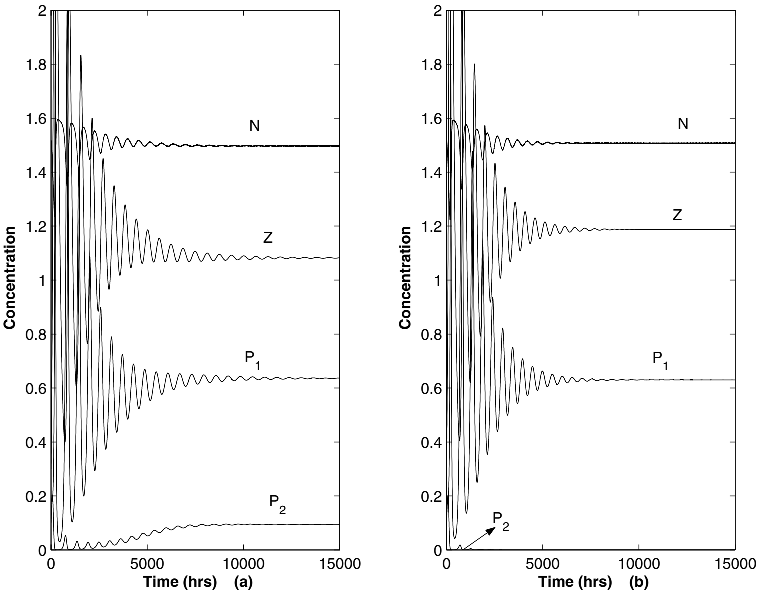


FIGURE 11. The equilibrium point $\mathbf{E}^*$ is locally asymptotically stable for $\mu_3 = 0.0053$, and the equilibrium point $\mathbf{E}_4$ is locally asymptotically stable for $\mu_3 = 0.006$

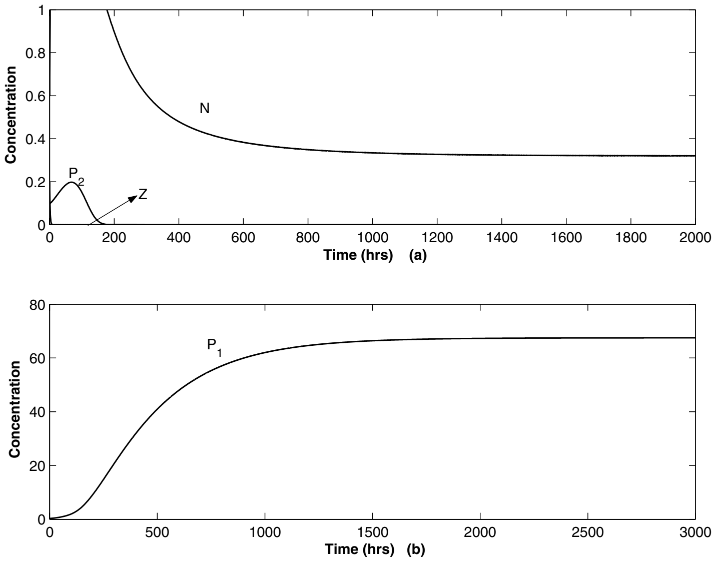


FIGURE 12. The equilibrium point $\mathbf{E}_1$ is locally asymptotically stable with a huge NTP biomass for $\mu_3 = 0.65$. (Other parameters are as given in Table 1).

5. Discussion

In this paper, we have considered a mathematical model for nutrient, phytoplankton and zooplankton populations in an open marine system. Since toxin producing phytoplankton plays an important role in the marine plankton dynamics, we have included growth equations of NTP and TPP explicitly. We have also incorporated the effects of toxin, nutrient recycling and spatial flows of nutrient and organisms in the model equations. Our analytical studies and simulations show that varying the parameters (viz. nutrient input concentration, dilution rates, interspecies competition etc.) can produce different outcomes, ranging from stable equilibria to cyclic or monospecies blooms.

Additional complications may arise from the effects of organisms at higher trophic levels – e.g., the zooplankton grazer. Grazers at higher trophic levels may increase the death rate of the zooplankton, and this intensified grazing may affect the qualitative and quantitative behaviour of the planktonic ecosystem. For example, Smayda and Villarea [32] concluded that the 1985 brown tide in Narragansett Bay may have been triggered during a period of reduced grazing. The monitoring and study of zooplankton grazing and food-web interactions should therefore be coupled with a phytoplankton-zooplankton monitoring programme focused on planktonic blooms.

Interspecies competition between the NTP and TPP may be an important factor in plankton ecosystem dynamics. The behaviour of the system may be different when the TPP becomes a stronger competitor than the NTP. The dynamics become more complicated when the nutrient effects interact with the effects of the interspecies competition.

Our analysis indicates that the nutrient-phytoplankton-zooplankton interactions are very complex and situation-specific. The nutrient controlled bloom may occur in certain favourable conditions. Top-down effects such as predation by higher trophic levels may also trigger blooms under other suitable conditions. Other mechanisms considered in this model for better biological realism (such as dilution rate, interspecies competition etc.) may also change planktonic dynamics significantly. Mathematical models may be used as a guide for predictive management, and to reach an improved understanding of bloom dynamics we have to devote more effort to define the individual as well as the combined role of nutrient, toxin, diffusion, grazer at higher trophic levels etc. To achieve this goal, collaboration between phytoplanktologists, zooplanktologists and mathematical biologists is to be encouraged.

Acknowledgment

The authors are grateful to Professor Ezio Venturino for his helpful comments on the draft of the manuscript.

References

[1] Anderson, D. M. *Toxic algae blooms and red tides: a global perspective. In: Red Tides: Biology, Environmental Science and Toxicology (T. Okaichi, D. M. Anderson and T. Nemoto, eds).* Elsevier, New York. U.S.A., (1989), 11–21.

[2] Anderson, D. M., Kaoru, Y., White, A. W. *Estimated Annual Economic Impacts from Harmful Algal Blooms (HABs) in the United States. Sea Grant Woods Hole.* (2000).

[3] Beltrami, E. and Carroll, T. O. *Modelling the role of viral disease in recurrent phytoplankton blooms.* J. Math. Biol. **32** (1994), 857–863.

[4] Blaxter, J. H. S and Southward, A. J. *Advances in Marine Biology.* Academic Press, London. U.K., (1997).

[5] Busenberg, S., Kishore, K. S., Austin, P. and Wake, G. *The dynamics of a model of a plankton–nutrient interaction.* J. Math. Biol. **52** (1990), 677–696.

[6] Carlsson, P., Graneli, E., Finenko, G. and Maestrini, S. Y. *Copepod grazing on a phytoplankton community containing the toxic dinoflagellate Dinophysis acuminata.* J. Plankton Res. **17** (1995), 1925–1938.

[7] Chattopadhyay J., Sarkar R. R., Mandal S. *Toxin producing plankton may act as a biological control for planktonic blooms-field study and mathematical modeling.* J. Theor. Biol. **215** (2002), 333–344.

[8] DeAngelis, D. L. *Dynamics of nutrient cycling and food webs.* Chapman & Hall, London, (1992).

[9] De Mott, W. R. and Moxter, F. *Foraging on cyanobacteria by copepods: responses to chemical defenses and resource abundance.* Ecology **72** (1991), 1820–1834.

[10] Duinker, J. and Wefer, G. *Das CO_2-Problem und die Rolle des Ozeans.* Naturwissenschahten **81** (1994), 237–242.

[11] Edwards, A. M. and Brindley, J. *Zooplankton mortality and the dynamical behaviour of plankton population-models.* Bull. Math. Biol. **61** (1999), 303–339.

[12] Evans, G. T. and Parslow, J. S. *A model of annual plankton cycles.* Biol. Oceanogr. **3** (1985), 327–427.

[13] Evans, G. T. *A framework for discussing seasonal succession and coexistence of phytoplankton species.* Limnol. Oceanogr. **33** (1988), 1027–1036.

[14] Frost, B. W. *Grazing control of phytoplankton stock in the open sub-arctic Pacific Ocean: A model assessing the role of mesozooplankton, particularly the large calanoid copepod neocalanus.* Mar. Ecol. Ser. **39** (1987), 49–68.

[15] Hallegraeff G. M. *A review of harmful algal blooms and the apparent global increase.* Phycologia **32** (1993), 79–99.

[16] Hansen, P. J. *The red tide dinoflagellate Alexandrium tamarense: Effects on behaviour and growth of a tintinnid ciliate.* Mar. Ecol. Prog. Ser. **53** (1989), 105–116.

[17] Hansen, P. J. *Growth and grazing response of a ciliate feeding on the red tide dinoflagellate Gyrodinium aureolum in monoculture and in mixture with a non-toxic alga.* Mar. Ecol. Prog. Ser. **121** (1995), 65–72.

[18] Huppert, A., Blasius, B. and Stone, L. *Bottom-Up Excitable Models of Phytoplankton Blooms.* Bull. Math. Biol. **66** (2004), 865–878.

[19] Ives, J. D. *The relationship between Gonyaulax tamarensis cell toxin levels and copepod ingestion rates.* In Toxic dinoflagellates: Proc. 3rd Int. conf. Elsevier. (1985), 413–418.

[20] Ives, J. D. *Possible mechanism underlying copepod grazing responses to levels of toxicity in red tide dinoflagellates.* J. Exp. Mar. Biol. Ecol. **112** (1987), 131–145.

[21] Kamiyama, T. and Arima, S. *Lethal effect of the dinoflagellate Heterocapsa circularisquama upon the tintinnid ciliate Favella taraikaensis.* Mar. Ecol. Prog. Ser. **160** (1997), 27–33.

[22] Kirk K., Gilbert J. *Variations in herbivore response to chemical defences: zooplankton foraging on toxic cyanobacteria.* Ecology **73** (1992), 2208.

[23] Nagumo, N. *Uber die Lage der Integralkurven gewonlicher Differantialgleichungen.* Proc. Phys. Math. Soc. Japan **24** (1942), 551.

[24] Nielsen T. G. et al. *Effects of Chrysochromulina polylepis subsurface bloom on the plankton community.* Mar. Ecol. Prog. Ser. **62** (1990), 21–35.

[25] Nejstgaard, J. C. and Solberg, P. T. *Repression of copepod feeding and fecundity by the toxic haptophyte Prymnesium patelliferum.* Sarsia **81** (1996), 339–344.

[26] Pal, S., Chatterjee, S., Chattopadhyay J. *Role of toxin and nutrient for the occurrence and termination of plankton bloom – Results drawn from field observations and a mathematical model.* Biosystems. **90** (2007), 87–100.

[27] Pitchford, J. W. and Brindley, J. *Iron limitation, grazing pressure and oceanic high nutrient and low chlorophyll (HNLC) regions.* J. Plank. Res. **21** (1999), 525–547.

[28] Pardo, O. *Global stability for a phytoplankton-nutrient system.* J. Biol. Syst. **8** (2000), 195–209.

[29] Ruan, S. *Persistence and coexistence in zooplankton-phytoplankton-nutrient models with instantaneous nutrient recycling.* J. Math. Biol. **31** (1993), 633–654.

[30] Ruan, S. *Oscillations in Plankton Models with Nutrient Recycling.* J. Theor. Biol. **208** (2001), 15–26.

[31] Sarno, Z. A. D. and Forlani, G. *Seasonal dynamics in the abundance of Micromonus pusilla (Prasinophyceae) and its viruses in the Gulf of Naples (Mediterranean Sea).* J. Plankton. Res. **21** (1999), 2143–2159.

[32] Smayda, T. J. and Villarea, T. A. *The 1985 "brown-tide" and open phytoplankton niche in Narragansett Bay during summer PIn E. M. Cosper et al. [eds.], novel phytoplankton blooms. causes and impacts of recurrent brown tides and other unusual blooms.* Springer, (1989), 159–187.

[33] Smayda, T. J. *Novel and nuisance phytoplankton blooms in the sea: evidence for a global epidemic. In: Toxic Marine Phytoplankton (E. Graneli, B. Sundstr øm, L. Edler and D.M. Anderson, eds.)*, Elsevier, New York. U.S.A., (1990), 29–40.

[34] Smith, H. L. *Competitive coexistence in an oscillating chemostat.* SIAM J. Appl. Math. **40** (1981), 498–522.

[35] Stoermer, E.F. and Smol, J.P. *In: The Diatoms Cambridge University Press*, Cambridge. U.K., (1999).

[36] Stone, L. and Berman, T. *Positive feedback in aquatic ecosystems: the case of microbial loop.* Bull. Math. Biol. **55** (1993), 919–936.

[37] Sykes, P. F. and Huntley, M. E. *Acute physiological reactions of Calanus pacificus: to selected dinoflagellates: Direct observations.* Mar. Biol. **94** (1987), 19–24.

[38] Taylor, A. J. *Characteristic properties of model for the vertical distribution of phytoplankton under stratification.* Ecol. Model. **40** (1988), 175-199.

[39] Truscott, J. E. and Brindley, J. *Ocean plankton populations as excitable media.* Bull. Math. Biol. **56** (1994), 981–998.

[40] Truscott, J. E. and Brindley, J. *Equilibria, stability and excitability in a general class of plankton population-models.* Philos. Trans. R. Soc. Lond. A **347** (1994), 703–718.

[41] Uye, S. *Impact of copepod grazing on the red tide flagellate Chattonella antiqua.* Mar. Biol. **92** (1986), 35.

[42] Wroblewski, J. S., Sarmiento, J. L. and Flierl, G. R. *An ocean basin scale model of plankton dynamics in the North Atlantic, 1, Solutions for the climatological oceanographic condition in May.* Global Biogeochem. Cycles **2** (1988), 199–218.

APPENDIX

Proof of boundedness of the system (2.2)

Let us define $w = N + P_1 + P_2 + Z$. Taking its time derivative along the solutions of (2.2) we have

$$\frac{dw}{dt} \le -D_0(N + P_1 + P_2 + Z) + DN^0,$$

with $D_0 = \min\{D, D_1, D_2, D_3\}$ — i.e., we can find a constant $m > 0$ such that

$$\frac{dw}{dt} + D_0 w \le m. \tag{A.1}$$

Gronwall's inequality then gives

$$w(N(t), P_1(t), P_2(t), Z(t)) \le \frac{m}{D_0}(1 - e^{-D_0 t}) + w(N(0), P_1(0), P_2(0), Z(0))e^{-D_0 t},$$

whence for $t \ge 0$

$$(N + P_1 + P_2 + Z)(t) \le max\left[(N(0) + P_1(0) + P_2(0) + Z(0)), \frac{m}{D_0}\right],$$

and for $t \to \infty$

$$(N + P_1 + P_2 + Z)(t) \le \frac{m}{D_0}.$$

Hence all of the biologically meaningful solutions of (2.2) are eventually confined in the region

$$B = \{(N, P_1, P_2, Z) \in R^4{}_+ : N + P_1 + P_2 + Z = \frac{m}{D_0} + \epsilon, \forall \epsilon > 0\}.$$

Stability analysis of the system (2.2) at boundary equilibria

Let $\overline{E} = (\overline{N}, \overline{P_1}, \overline{P_2}, \overline{Z})$ be an arbitrary equilibrium. Then the variational matrix $\overline{V} \equiv V(\overline{E})$ at $\overline{E}$ is given by

$$\overline{V} = \begin{pmatrix} -D-\alpha_1\overline{P_1}-\alpha_2\overline{P_2} & -\alpha_1\overline{N}+\eta_1 & -\alpha_2\overline{N}+\eta_2 & \eta_3 \\ \theta_1\overline{P_1} & \begin{array}{c}\theta_1\overline{N}-\beta_1\overline{Z}-e_1\overline{P_2}\\ -(\mu_1+D_1)\end{array} & -e_1\overline{P_1} & -\beta_1\overline{P_1} \\ \theta_2\overline{P_2} & -e_2\overline{P_2} & \begin{array}{c}\theta_2\overline{N}-\beta_2\overline{Z}-e_2\overline{P_1}\\ -(\mu_2+D_2)\end{array} & -\beta_2\overline{P_2} \\ 0 & \gamma_1\overline{Z} & -\gamma_2\overline{Z} & \begin{array}{c}\gamma_1\overline{P_1}-\gamma_2\overline{P_2}\\ -(\mu_3+D_3)\end{array} \end{pmatrix}.$$

The four eigenvalues of the variational matrix $V_0 \equiv V(E_0)$ are $\lambda_1 = -D < 0$, $\lambda_2 = \theta_1 N^0 - (D_1+\mu_1)$, $\lambda_3 = \theta_2 N^0 - (D_2+\mu_2)$ and $\lambda_4 = -(\mu_3+D_3) < 0$. Clearly this steady state is asymptotically stable if and only if

$$N^0 < \min\{\frac{\mu_1+D_1}{\theta_1}, \frac{\mu_2+D_2}{\theta_2}\}.$$

However, if

$$N^0 > \min\{\frac{\mu_1+D_1}{\theta_1}, \frac{\mu_2+D_2}{\theta_2}\},$$

the plankton free steady state becomes unstable, a saddle, and there exists a TPP and zooplankton free steady state E_1 if

$$\frac{D_1+\mu_1}{\theta_1} < N_0 < \frac{D_2+\mu_2}{\theta_2}, \tag{A.2}$$

or a NTP and zooplankton free steady state E_2 if

$$\frac{D_2+\mu_2}{\theta_2} < N_0 < \frac{D_1+\mu_1}{\theta_1}. \tag{A.3}$$

The eigenvalues of the variational matrix V_1 are ${\lambda_1}^{(1)}$, ${\lambda_2}^{(1)}$, which are the roots of the equation $\lambda^2 + \lambda(D + \alpha_1 P_1) + D(N^0\theta_1 - \mu_1 - D_1) = 0$, together with ${\lambda_3}^{(1)} = \gamma_1 P_1 - (D_3 + \mu_3)$ and ${\lambda_4}^{(1)} = \theta_2 N - e_2 P_1 - (D_2 + \mu_2)$. Clearly ${\lambda_1}^{(1)}$ and ${\lambda_2}^{(1)}$ have negative real parts, since $N^0\theta_1 - \mu_1 - D_1 > 0$ for the existence of E_1. Now ${\lambda_3}^{(1)} < 0$ if

$$N^0 < R_2 \tag{A.4}$$

where

$$R_2 = \frac{(D_3+\mu_3)(\alpha_1\mu_1+\alpha_1 D_1-\theta_1\gamma_1)}{\gamma_1\theta_1 D} + \frac{D_1+\mu_1}{\theta_1},$$

and ${\lambda_4}^{(1)} < 0$ if

$$R_1 < N^0 \tag{A.5}$$

where

$$R_1 = \frac{(\alpha_1\mu_1+\alpha_1 D_1-\theta_1\eta_1)[\theta_2(\mu_1+D_1)-\theta_1(\mu_2+D_2)]}{e_2 D{\theta_1}^2} + \frac{\mu_1+D_1}{\theta_1}.$$

Combining (A.2), (A.4) & (A.5) we observe that E_1 is asymptotically stable if

$$\max\{R_1, \frac{\mu_1 + D_1}{\theta_1}\} < N^0 < \min\{R_2, \frac{\mu_2 + D_2}{\theta_2}\} \tag{A.6}$$

where $R_1,\ R_2$ are as above.

Further, the eigenvalues of the variational matrix V_2 around the equilibrium E_2 of the system (2.2) are the roots ${\lambda_1}^{(2)}$, ${\lambda_2}^{(2)}$ of the equation

$$\lambda^2 + \lambda(D + \alpha_2 P_2) + D(N^0\theta_2 - \mu_2 - D_2) = 0,$$

${\lambda_3}^{(2)} = -\gamma_2 P_2 - (D_3 + \mu_3) < 0$ and ${\lambda_4}^{(2)} = \theta_1 N - e_1 P_2 - (D_1 + \mu_1)$. Clearly ${\lambda_1}^{(2)}$ and ${\lambda_2}^{(2)}$ have negative real parts, since $N^0\theta_2 - \mu_2 - D_2 > 0$ for the existence of E_2. Now if ${\lambda_4}^{(2)} < 0$ if

$$R_3 < N^0 \tag{A.7}$$

where

$$R_3 = \frac{(\alpha_2\mu_2 + \alpha_2 D_2 - \theta_2\eta_2)[\theta_1(\mu_2 + D_2) - \theta_2(\mu_1 + D_1)]}{e_1 D{\theta_2}^2} + \frac{\mu_2 + D_2}{\theta_2}.$$

Combining (A.3) and (A.7), we conclude that E_2 is asymptotically stable if

$$\max\{R_3, \frac{\mu_2 + D_2}{\theta_2}\} < N^0 < \frac{\mu_1 + D_1}{\theta_1}\,. \tag{A.8}$$

One eigenvalue of the variational matrix V_3 is ${\lambda_1}^{(3)} = \gamma_1 P_1 - \gamma_2 P_2 - (\mu_3 + D_3)$, and the other eigenvalues are given by $\lambda^3 + {Q_1}^{(3)}\lambda^2 + {Q_2}^{(3)}\lambda + {Q_3}^{(3)} = 0$ where

$$\begin{aligned}
{Q_1}^{(3)} &= D + \alpha_1 P_1 + \alpha_2 P_2, \\
{Q_2}^{(3)} &= -e_1 e_2 P_1 P_2 + \alpha_1 N\theta_1 P_1 - \eta_1\theta_1 P_1 + \alpha_2 N\theta_2 P_2 - \eta_2\theta_2 P_2, \\
{Q_3}^{(3)} &= P_1 P_2[\eta_1 e_1\theta_2 + \eta_2 e_2\theta_1 - (D + \alpha_1 P_1 + \alpha_2 P_2)e_1 e_2 - \alpha_1 N e_1\theta_2 - \alpha_2 N e_2\theta_1].
\end{aligned}$$

Now ${Q_3}^{(3)} = P_1 P_2[C - AN^{(2)}]/N < 0$ since $C < 0$ — i.e., $N^0 > R_3$ so E_3 is an unstable saddle point.

The eigenvalues of V_4 are ${\lambda_1}^{(4)} = \theta_2 N - \beta_2 Z - e_2 P_1 - (\mu_2 + D_2)$, and the roots of $\lambda^3 + {Q_1}^{(4)}\lambda^2 + {Q_2}^{(4)}\lambda + {Q_3}^{(4)} = 0$ where

$$\begin{aligned}
{Q_1}^{(4)} &= D + \alpha_1 P_1, \\
{Q_2}^{(4)} &= \beta_1\gamma_1 P_1 Z + \alpha_1 N\theta_1 P_1 - \eta_1\theta_1 P_1, \\
{Q_3}^{(4)} &= (D + \alpha_1 P_1)\beta_1\gamma_1 P_1 Z - \eta_3\theta_1\gamma_1 P_1 Z.
\end{aligned}$$

Now ${\lambda_1}^{(4)} < 0$ if $R_4 < 1$ and ${Q_3}^{(4)} = [D\beta_1\gamma_1 + \alpha_1\beta_1(\mu_3 + D_3) - \eta_3\gamma_1\theta_1]P_1 Z > 0$ since $N > 0$. Also ${Q_1}^{(4)}{Q_2}^{(4)} - {Q_3}^{(4)} = \theta_1 P_1[(D + \alpha_1 P_1)(\alpha_1 N - \eta_1) + \eta_3\gamma_1 Z] > 0$ if $R_5 > 1$. Thus all the conditions of the Routh–Hurwitz criterion on the roots are satisfied, and the system is stable at the equilibrium E_4. On the other hand, if $R_4 > 1$ and $R_5 < 1$, then E_4 is an unstable saddle point.

Stability analysis of the interior equilibrium of (2.2)

The variational matrix of system (2.2) around the positive equilibrium $E^* = (N^*, {P_1}^*, {P_2}^*, Z^*)$ is

$$V^* = \begin{pmatrix} -D - \alpha_1 P_1^* - \alpha_2 P_2^* & -\alpha_1 N^* + \eta_1 & -\alpha_2 N^* + \eta_2 & \eta_3 \\ \theta_1 P_1^* & 0 & -e_1 P_1^* & -\beta_1 P_1^* \\ \theta_2 P_2^* & -e_2 P_2^* & 0 & -\beta_2 P_2^* \\ 0 & \gamma_1 Z^* & -\gamma_2 Z & 0 \end{pmatrix}.$$

The characteristic equation of the Jacobian at E^* is

$$\lambda^4 + T_1\lambda^3 + T_2\lambda^2 + T_3\lambda + T_4 = 0,$$

where the coefficients are

$$\begin{aligned} T_1 &= D + \alpha_1 {P_1}^* + \alpha_2 {P_2}^*, \\ T_2 &= -\beta_2\gamma_2 {P_2}^* Z^* - e_1 {P_1}^* e_2 {P_2}^* + \beta_1 {P_1}^* \gamma_1 Z^* - \theta_1 {P_1}^*(-\alpha_1 N^* + \eta_1) \\ &\quad +\theta_2 {P_2}^*(\alpha_2 N^* - \eta_2), \\ T_3 &= -T_1(\beta_2\gamma_2 {P_2}^* Z^* + e_1 e_2 {P_1}^* {P_2}^* - \beta_1\gamma_1 {P_1}^* Z^*) \\ &\quad -(e_1\gamma_1\beta_2 - e_2\beta_1\gamma_2){P_1}^*{P_2}^*Z^* - \theta_1 {P_1}^*(\alpha_2 e_2 N^* {P_2}^* - \eta_2 e_2 {P_2}^* + \eta_3\gamma_1 Z^*) \\ &\quad +\theta_2 {P_2}^*(\eta_1 e_1 {P_1}^* - \alpha_1 e_1 N^* {P_1}^* + \eta_3\gamma_2 Z^*), \\ T_4 &= {P_1}^*{P_2}^*Z^*[T_1(-e_1\beta_2\gamma_1 + e_2\beta_1\gamma_2) + \theta_1(-\alpha_1\beta_2\gamma_2 N^* + \eta_1\beta_2\gamma_2 - \alpha_2\beta_2\gamma_1 N^* \\ &\quad +\eta_2\beta_2\gamma_1 - \eta_3 e_2\gamma_2) + \theta_2(\alpha_1\beta_1\gamma_2 N^* - \eta_1\beta_1\gamma_2 + \alpha_2\beta_1\gamma_1 N^* \\ &\quad -\eta_2\beta_1\gamma_1 + \eta_3 e_1\gamma_1)]. \end{aligned}$$

If (3.1) and (3.2) are satisfied, it follows that $T_1 > 0$ and $T_1T_2 - T_3 > 0$; while if (3.3) is satisfied then $D_5 = T_3(T_1T_2 - T_3) - {T_1}^2 T_4 > 0$. Thus by the Routh–Hurwitz criterion, E^* is locally asymptotically stable.

Nandadulal Bairagi
Centre for Mathematical Biology and Ecology,
Department of Mathematics, Jadavpur University,
Kolkata - 700032, India
e-mail: nbairagi@math.jdvu.ac.in

Samaresh Pal
Department of Mathematics
Ramakrishna Mission Vivekananda Centenary College
Rahara, Kolkata 700118, India
e-mail: samaresp@yahoo.co.in

Samrat Chatterjee
Dipartimento di Matematica, Universita' di Torino
via Carlo Alberto 10, 10123 Torino, Italia
e-mail: samrat_ct@rediffmail.com

Joydev Chattopadhyay
Agricultural and Ecological Research Unit,
Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India
e-mail: joydev@isical.ac.in

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 65–80

Stability and Optimal Harvesting in a Stage Structure Predator-Prey Switching Strategy

Qamar J. A. Khan, Lakdere Benkherouf and Nejib Smaoui

Abstract. A predator-prey interaction is considered, where the prey has a stage structure — i.e., two life stages, immature and mature. The predator consumes both the immature and mature prey, and the prey is more prone to the predator at higher prey population densities. Both local and global stability of the system equilibria are discussed. With harvesting of the mature prey, there are threshold conditions for a sustainable yield.

Mathematics Subject Classification (2000). 90B05.

Keywords. Stage structure, Local stability, Global stability, Optimal harvesting, Switching.

1. Introduction

It is now recognised that a predator may prefer to eat prey species according to age, size, weight, number, etc. For a prey species of small size, with little or no defence capability against its predator, the predator catches a member of that species proportional to its abundance. The predator feeds preferentially on the most numerous species, which is consequently over-represented in the predator's diet. It is also likely that a predator will consume other species more when a given prey species becomes relatively less abundant, a behaviour known as predator switching. Many examples may be cited where a predator prefers to prey on species that are most abundant at any particular time, and switching is a normal feature of predator behaviour [1–3]. Mathematical models involving one predator and two prey species have generally been studied, in which the predator feeds more intensively on the more abundant species [4–13].

Results in this paper were presented at the International Conference on Mathematical Modelling and Computation held at the University of Brunei Darussalam during 5–8 June 2006, in conjunction with the 20th anniversary celebration of the foundation of the university.
This work was supported by SQU Grant IG/SCI/DOMS/03/05.

Almost all animals have an immature and mature age structure, and in particular mammals and some amphibians. There are two types of stage-dependent predation in predator-prey models — viz., where the predators eat only adults (e.g., where insects are preyed upon only in the adult stage [14]), and well documented cases where the predators consume only immature prey [15, 16].

Several mathematical models have been proposed to account for immature and mature stage structure [17–22]. In this paper, we consider the case where a second species is a predator of both the mature and immature stages – where, instead of choosing individuals at random, the predator catches a member of the immature or mature prey populations proportional to their abundance. Thus the predator feeds preferentially on the most numerous stage, implying a switch from the immature to the mature population or vice versa. As in reference [21], we also consider harvesting of the mature prey population, which is appropriate from both an economic and biological viewpoint for renewable resource management [23–26]. We obtain conditions for the local and global stability of the system equilibria, and a threshold for harvesting at a sustainable yield.

The mathematical model is presented in Section 2. Section 3 is concerned with equilibrium and stability, Section 4 with asymptotic stability, and optimal harvesting is discussed in Section 5. A summary of our numerical results is given in Section 6, together with a final discussion.

2. The mathematical model

The prey-predator model with a simple multiplicative effect, where the prey species has an immature and mature stage structure, is of the form:

$$\begin{aligned}\frac{dx_1}{dt} &= \alpha x_2 - kx_1 - \beta x_1 - \eta x_1^2 - \frac{bx_1^2 y}{x_1 + x_2},\\ \frac{dx_2}{dt} &= \beta x_1 - kx_2 - \frac{bx_2^2 y}{x_1 + x_2} - qex_2,\\ \frac{dy}{dt} &= \left(\frac{bx_1^2}{x_1 + x_2} + \frac{bx_2^2}{x_1 + x_2} - d\right) y,\end{aligned} \tag{2.1}$$

where

x_i is the population of the immature and mature prey species of stage i,
y is the population of the predator species,
α is the per capita birth rate of the mature prey species,
β is the maturation rate from the immature to the mature stage,
k is the per capita death rate of both prey species,
η is the proportionality of self-interaction of the immature population,
d is the per capita death rate of the predator, and
b is the prey-predator response rate towards each prey species.

We assume all the parameters in the models are positive, and that $x_i(0) > 0$, $i = 1, 2,\quad y(0) > 0$.

Let $h = qex_2$ be the harvesting yield, where q is the catchability coefficient and e is the harvesting effort, and let us consider optimal harvesting of the mature population. In order to reduce the number of parameters, we introduce $\tau = b\,t$ and write

$$\frac{\alpha}{b} = \alpha_1, \frac{\eta}{b} = \eta_1, \frac{k}{b} = k_1, \frac{h}{b} = H, \frac{\beta}{b} = \beta_1, \frac{d}{b} = d_1, \text{and } \frac{e}{b} = e_1,$$

so the system of equations (2.1) becomes

$$\begin{aligned} \frac{dx_1}{d\tau} &= \alpha_1 x_2 - k_1 x_1 - \beta_1 x_1 - \eta_1 x_1^2 - \frac{x_1^2 y}{x_1 + x_2}, \\ \frac{dx_2}{d\tau} &= \beta_1 x_1 - k_1 x_2 - \frac{x_2^2 y}{x_1 + x_2} - qe_1 x_2, \\ \frac{dy}{d\tau} &= \left(\frac{x_1^2}{x_1 + x_2} + \frac{x_2^2}{x_1 + x_2} - d_1 \right) y. \end{aligned} \tag{2.2}$$

It is also convenient to write $H = qe_1 x_2$, for our discussion later.

3. Steady states and stability analysis

The steady states of the system (2.2) of course correspond to equating the derivatives on the left-hand sides to zero, and solving the resulting algebraic equations we identify three possible steady states as follows:

(i) $\bar{E}_0 = (0, 0, 0)$, where all populations are extinct, which always exists;

(ii) $$\bar{E}_1 = (\hat{x}_1, \hat{x}_2, 0) = \left(\frac{\alpha_1 - (k_1 + \beta_1)\bar{x}}{\eta_1 \bar{x}}, \frac{\alpha_1 - (k_1 + \beta_1)\bar{x}}{\eta_1 \bar{x}^2}, 0 \right),$$

$$\text{where} \quad \bar{x} = \frac{\hat{x}_1}{\hat{x}_2} = \frac{k_1 + qe_1}{\beta_1} \text{ and } \frac{\hat{x}_1^2 + \hat{x}_2^2}{\hat{x}_1 + \hat{x}_2} < d_1,$$

in which the prey populations have reached equilibrium levels but the predator population has died out; and

(iii) $$\begin{aligned} \bar{E}_2 &= (\hat{x}_1, \hat{x}_2, \hat{y}) \\ &= \left(\frac{d_1(\bar{x}+1)\bar{x}}{(1+\bar{x}^2)}, \frac{d_1(1+\bar{x})}{(1+\bar{x}^2)}, \frac{(1+\bar{x})}{\bar{x}} (\frac{\alpha_1}{\bar{x}} - k_1 - \beta_1 - \eta_1 \hat{x}_1) \right), \end{aligned}$$

or equivalently

$$\bar{E}_2 = \left(\frac{d_1(\bar{x}+1)\bar{x}}{1+\bar{x}^2}, \frac{d_1(\bar{x}+1)}{1+\bar{x}^2}, (1+\bar{x})(\beta_1 \bar{x} - k_1 - qe_1) \right), \tag{3.1}$$

where both prey and predator exist. Here $\bar{x} = \dfrac{\hat{x}_1}{\hat{x}_2}$ is a real positive root of the equation

$$\begin{aligned} \beta_1 \bar{x}^5 + \bar{x}^4 \left(-(k_1 + qe_1)\right) + \bar{x}^3((k_1 + \beta_1) + \beta_1 + \eta_1 d_1) \\ + \bar{x}^2 \left(-k_1 - qe_1 + \eta_1 d_1 - \alpha_1\right) + \bar{x}\left(k_1 + \beta_1\right) - \alpha_1 = 0. \end{aligned} \tag{3.2}$$

For equilibrium values $(\hat{x}_1, \hat{x}_2, \hat{y})$ to be positive, a positive real root of (3.2) must be bounded as

$$\frac{k_1 + qe_1}{\beta_1} < \bar{x} < \frac{\alpha_1}{k_1 + \beta_1 + \eta_1 \hat{x}_1}. \tag{3.3}$$

Lemma 1. *Polynomial (3.2) has only one positive root if* $-k_1 - qe_1 + \eta_1 d_1 - \alpha_1 > 0$.

For the proof see Appendix A.

3.1. Stability

3.1.1. Stability analysis of the equilibrium $\bar{\mathbf{E}}_0 = (0, 0, 0)$. We proceed in the usual manner, by considering small disturbances from the steady state and linearising the resulting equations. The stability matrix and characteristic equation are found to be

$$\begin{bmatrix} -(k_1 + \beta_1) - \lambda & \alpha_1 & 0 \\ \beta_1 & -(k_1 + \beta_1) - \lambda & 0 \\ 0 & 0 & -d_1 - \lambda \end{bmatrix},$$

and

$$(\lambda + d_1)\left[\lambda^2 + \lambda((k_1 + \beta_1) + (k_1 + qe_1)) + (k_1 + \beta_1)(k_1 + qe_1) - \alpha_1\beta_1\right] = 0.$$

The eigenvalues for the equilibrium point $\bar{E}_0 = (0, 0, 0)$ are $\lambda_1 = -d_1$ and

$$\lambda_{2,3} = \frac{1}{2}\left[-(2k_1 + \beta_1 + qe_1) \pm \sqrt{\Delta_1}\right],$$

where $\Delta_1 = (2k_1 + \beta_1 + qe_1)^2 - 4\left((k_1 + \beta_1)(k_1 + qe_1) - \alpha_1\beta_1\right)$. It can also be shown that all these eigenvalues are negative if

$$k_1 + qe_1 > \alpha_1.$$

Theorem 2. *If $k_1 + qe_1 > \alpha_1$, the equilibrium E_0 is locally asymptotically stable if and only if $k_1 + qe_1 > 0$, and otherwise it is unstable.*

Next we proceed to global stability analysis of all equilibria mentioned above, by applying the Lyapunov indirect method to the system of differential equations (2.2). Consider the Lyapunov function

$$L = x_1 + x_2 + y,$$

which leads to $\dfrac{dL}{d\tau} = \dfrac{dx_1}{d\tau} + \dfrac{dx_2}{d\tau} + \dfrac{dy}{d\tau} = x_2(\alpha_1 - qe_1 - k_1) - (k_1x_1 + \eta_1 x_1^2 + d_1 y),$

so that

$$\frac{dL}{d\tau} < 0 \text{ for all } x_1, x_2, y > 0 \text{ if } \alpha_1 < k_1 + qe_1.$$

The next theorem is then immediate.

Theorem 3. *The zero solution of (2.2) is globally asymptotic stable (GAS) if and only if $\alpha_1 < k_1 + q\ e_1$.*

3.1.2. Stability analysis of equilibrium $\bar{\mathbf{E}}_1 = (\hat{x}_1, \hat{x}_2, 0)$. The stability matrix of the equilibrium $\bar{E}_1$ is

$$\begin{bmatrix} -k_1 - \beta_1 - 2\eta_1\hat{x}_1 - \lambda & \alpha_1 & \dfrac{-\hat{x}_1^2}{\hat{x}_1 + \hat{x}_2} \\ \beta_1 & -k_1 - qe_1 - \lambda & \dfrac{-\hat{x}_2^2}{\hat{x}_1 + \hat{x}_2} \\ 0 & 0 & \xi - \lambda \end{bmatrix} \tag{3.4}$$

where $\quad \xi = \dfrac{\hat{x}_1^2 + \hat{x}_2^2}{\hat{x}_1 + \hat{x}_2} - d_1.$

The characteristic equation is

$$(\lambda - \xi)\left(\lambda^2 + \lambda(2k_1 + qe_1 + \beta_1 + 2\eta_1\hat{x}_1) + (k_1 + qe_1)(k_1 + \beta_1 + 2\eta_1\hat{x}_1) - \alpha_1\beta_1\right) = 0.$$

The three eigenvalues of the stability matrix (3.4) are $\lambda_1 = \xi$ and

$$\lambda_{2,3} = \frac{1}{2}\left[-(2k_1 + qe_1 + \beta_1 + 2\eta_1\hat{x}_1) \pm \sqrt{\Delta_2}\right],$$

where $\Delta_2 = (2k_1 + qe_1 + \beta_1 + 2\eta\hat{x}_1)^2 - 4((k_1 + qe_1)(k_1 + \beta_1 + 2\eta_1\hat{x}_1) - \alpha_1\beta_1)$.

When $k_1 + qe_1 > \alpha_1$, this produces two real negative eigenvalues — and consequently all three eigenvalues are negative, so the equilibrium $\bar{E}_1$ is asymptotically stable:

Theorem 4. *If $k_1 + qe_1 > \alpha_1$, then $\bar{E}_1$ is asymptotically stable, but otherwise it is unstable.*

3.1.3. Stability analysis of equilibrium $\bar{E}_2 = (\hat{x}_1, \hat{x}_2, y)$. The stability matrix of the equilibrium $\bar{E}_2$ is

$$\begin{bmatrix} L - \lambda & B & \dfrac{-\hat{x}_1\bar{x}}{1 + \bar{x}} \\ A & -A\bar{x} - \lambda & \dfrac{-\hat{x}_2}{1 + \bar{x}} \\ C + D\ \hat{x}_1 & C + D\ \hat{x}_2 & -\lambda \end{bmatrix}, \tag{3.5}$$

where

$$\begin{aligned} B &= \alpha_1 + \frac{\hat{x}_1^2\ \hat{y}}{(\hat{x}_1 + \hat{x}_2)^2}, \\ L &= -\frac{B}{\bar{x}} - \eta_1\hat{x}_1, \\ A &= \beta_1 + \frac{\hat{x}_2^2\ \hat{y}}{(\hat{x}_1 + \hat{x}_2)^2}, \\ C &= \frac{-\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)}{(\hat{x}_1 + \hat{x}_2)^2}, \\ D &= \frac{2\hat{y}}{\hat{x}_1 + \hat{x}_2}. \end{aligned} \tag{3.6}$$

The characteristic equation associated with the positive equilibrium $\bar{E}_2$ is

$$\lambda^3 + \lambda^2 (A\bar{x} - L) - \lambda \left(L\ A\ \bar{x} - \frac{\hat{x}_2}{1+\bar{x}} (C + D\ \hat{x}_2) + AB - \frac{\hat{x}_1\bar{x}}{1+\bar{x}} (C + D\ \hat{x}_1) \right)$$
$$- \frac{L\ \hat{x}_2}{(1+\bar{x})} (C + D\ \hat{x}_2) + \frac{B\ \hat{x}_2}{1+\bar{x}} (C + D\ \hat{x}_1) + \frac{AC\ \hat{x}_1\bar{x}}{1+\bar{x}} + \frac{AC\ \hat{x}_1\bar{x}^2}{1+\bar{x}}$$
$$+ \frac{A\ D\ \hat{x}_1\bar{x}}{1+\bar{x}} (\hat{x}_2 + \hat{x}_1\bar{x}) = 0.$$

The above equation can be written in the form

$$\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0, \tag{3.7}$$

where

$$\begin{aligned}
a_1 &= A\bar{x} - L, \\
a_2 &= \frac{\hat{x}_2}{1+\bar{x}} (C + D\hat{x}_2) + \frac{\hat{x}_1\bar{x}}{1+\bar{x}} (C + D\hat{x}_1) - LA\bar{x} - AB, \\
a_3 &= \frac{B\hat{x}_2}{1+\bar{x}} (C + D\hat{x}_1) + \frac{AD\hat{x}_1\bar{x}}{1+\bar{x}} (\hat{x}_2 + \hat{x}_1\bar{x}) + \frac{AC\hat{x}_1\bar{x}}{1+\bar{x}} \\
&- \frac{L\ \hat{x}_2}{(1+\bar{x})} (C + D\hat{x}_2) + \frac{AC\bar{x}_1\bar{x}^2}{1+\bar{x}}.
\end{aligned}$$

The Routh–Hurwitz stability criteria for this third-order system are:

$$\text{(a)}\ \ a_1 > 0,\ a_3 > 0 \qquad \text{and} \qquad \text{(b)}\ \ a_1 a_2 > a_3.$$

Hence, the equilibrium $\bar{E}_2 = (\hat{x}_1, \hat{x}_2, \hat{y})$ will be locally stable to small perturbations provided $\bar{x} > 1$ and unstable for $\bar{x} \leq 0$. The details of the analysis are given in Appendix B. We summarise these results in the following theorem.

Theorem 5. *If $\bar{x} > 1$, then $\bar{E}_2 = (\hat{x}_1, \hat{x}_2, \hat{y})$ is locally asymptotically stable.*

4. Asymptotic stability of interior equilibrium

Theorem 6. *A positive interior equilibrium point of system (2.2) is asymptotically stable provided the ratio of the young and adult prey species at any time has the same value as the positive root of equation (3.2).*

Proof. We make use of the general Lyapunov function

$$v(x_1, x_2, y) = \sum_{i=1}^{2} \left[(x_i - \hat{x}_i) - \hat{x}_i \ln\left(\frac{x_i}{\hat{x}_i}\right) \right] + (y - \hat{y}) - \hat{y} \ln\left(\frac{y}{\hat{y}}\right). \tag{4.1}$$

On calculating the time derivative of equation (4.1) along the solutions of equation (2.2), we have

$$\begin{aligned}\frac{dv}{d\tau} &= (x_1-\hat{x}_1)\left[\alpha_1\frac{x_2}{x_1}-\beta_1-k_1-\eta_1 x_1-\frac{x_1 y}{x_1+x_2}\right]\\ &+(x_2-\hat{x}_2)\left[\beta_1\frac{x_1}{x_2}-k_1-\frac{x_2 y}{x_1+x_2}-q\ e_1\right]\\ &+(y-\hat{y})\left[\frac{x_1^2}{x_1+x_2}+\frac{x_2^2}{x_1+x_2}-\frac{\hat{x}_1^2}{\hat{x}_1+\hat{x}_2}-\frac{\hat{x}_2^2}{\hat{x}_1+\hat{x}_2}\right]. \qquad (4.2)\end{aligned}$$

At equilibrium

$$\begin{aligned}k_1+\beta_1 &= \frac{\alpha_1\hat{x}_2}{\hat{x}_1}-\eta_1\hat{x}_1-\frac{\hat{x}_1\hat{y}}{\hat{x}_1+\hat{x}_2},\\ q_1e_1\hat{x}_2 &= \beta_1\hat{x}_1-k_1\hat{x}_2-\frac{\hat{x}_2^2\hat{y}}{\hat{x}_1+\hat{x}_2},\end{aligned} \qquad (4.3)$$

so that

$$\begin{aligned}\frac{dv}{d\tau} &= -\eta_1(x_1-\hat{x}_1)^2+\frac{(\hat{x}_1x_2-x_1\hat{x}_2)}{(x_1+x_2)(\hat{x}_1+\hat{x}_2)}\left[\hat{y}(x_1-x_2)-y(\hat{x}_1-\hat{x}_2)\right]\\ &+\ \alpha_1(x_1-\hat{x}_1)\left(\frac{(\hat{x}_1x_2-x_1\hat{x}_2)}{x_1\hat{x}_1}\right)-\beta_1(x_2-\hat{x}_2)\left(\frac{(\hat{x}_1x_2-x_1\hat{x}_2)}{x_2\hat{x}_2}\right). \qquad (4.4)\end{aligned}$$

If $\dfrac{\hat{x}_1}{\hat{x}_2}=\dfrac{x_1}{x_2}$ (i.e., the ratio of young and adult prey species is constant and the same as the ratio at equilibrium value), then

$$\frac{dv}{d\tau}=-\eta(x_1-\hat{x}_1)^2<0.$$

Hence $\bar{E}_2=(\hat{x}_1,\hat{x}_2,\hat{y})$ is a basin of attraction for $\dfrac{\hat{x}_1}{\hat{x}_2}=\dfrac{x_1}{x_2}$, for all $\tau\geq 0$. □

5. Optimal harvesting

From both an economic and biological point of view for renewable resource management, it is more appropriate for the mature prey species to be consumed. It is also desirable to have a unique positive stable equilibrium. If the ratio of the mature and the immature populations is constant and the immature population is higher than the mature, then the unique positive equilibrium of system (2.2) is asymptotically stable. In this section, we consider harvesting the mature population and study the maximum sustainable yield under the system (2.2).

System (2.2) has a positive equilibrium $\bar{E}_2$ if and only if inequality (3.3) holds, hence

$$e_1<\frac{\beta_1\bar{x}-k_1}{q}. \qquad (5.1)$$

Thus the maximum value of the harvesting effort is given by equation (5.2),

viz., $e_1 = e_1^* = \dfrac{\beta_1 \bar{x} - k_1}{q}$, that is $e_1 \in [0, e_1^*)$.

Letting $x_2 = \hat{x}_2$, the harvesting yield of the system (2.2) is

$$H\left(e_1\right) = q e_1 \hat{x}_2 = d_1 q \left[\frac{e_1\left(1+\bar{x}\right)}{\left(\bar{x}^2+1\right)}\right]. \tag{5.2}$$

Finding the derivative of $H\ (e_1)$, we get

$$\frac{dH}{de_1} = d_1 q \left[\frac{\left\{(1+\bar{x}) + \dfrac{d\bar{x}}{de_1} e_1\right\}\left(\bar{x}^2+1\right) - 2\bar{x}\left(\bar{x}+1\right) e_1 \dfrac{d\bar{x}}{de_1}}{\left(\bar{x}^2+1\right)^2}\right],$$

such that $\dfrac{dH}{de_1} > 0$ if $e_1 < \dfrac{(1+\bar{x})\left(\bar{x}^2+1\right)}{\dfrac{d\bar{x}}{de_1}\left(\bar{x}^2+2\bar{x}-1\right)}$, (5.3)

where $\bar{x} > 1$.

Consequently, we have the following:

(i) if $e_1 > e_1^*$ and the inequality in (5.4) holds, then the maximum sustainable yield is

$$\max H = q\ e_1^* \hat{x}_2 = \frac{d_1\left(\beta_1 \bar{x} - k\right)\left(1+\bar{x}\right)}{\left(\bar{x}^2+1\right)}; \quad \text{and} \tag{5.4}$$

(ii) the solution of $\dfrac{dH}{de_1} = 0$ is $e_1 = \bar{e}_1 = \dfrac{(1+\bar{x})\left(\bar{x}^2+1\right)}{\dfrac{d\bar{x}}{de_1}\left(\bar{x}^2+2\bar{x}-1\right)}$,

giving the point of maximum effort.

The maximum yield depends on the values of $\bar{e}_1$ and e^* such that:

(a) if $\bar{e}_1 > e_1^*$ and $dH/de_1 > 0$, the maximum yield is given by equation (5.5); and
(b) if $\bar{e}_1 \in [0, e_1^*)$ and inequality (5.4) is not satisfied. the corresponding maximum yield is

$$\max H = q \bar{e}_1 \hat{x}_2 = \frac{q d_1 \left(1+\bar{x}\right)^2}{\dfrac{d\bar{x}}{de_1}\left(\bar{x}^2+2\bar{x}-1\right)}. \tag{5.5}$$

Theorem 7.
(i) If $e_1 < \dfrac{(1+\bar{x})\left(\bar{x}^2+1\right)}{\dfrac{d\bar{x}}{de_1}\left(\bar{x}^2+2\bar{x}-1\right)}$, *and* $\bar{e}_1 > e_1^*$, *then the maximum sustainable yield for system (2.2) is given by (5.5).*

(ii) If $e_1 \geq \dfrac{(1+\bar{x})\left(\bar{x}^2+1\right)}{\dfrac{d\bar{x}}{de_1}\left(\bar{x}^2+2\bar{x}-1\right)}$ *and* $\bar{e}_1 \in [0, e_1^*)$, *then the maximum sustainable yield for system (2.2) is given by (5.6).*

6. Numerical results and discussion

The numerical solution of system (2.1) using a fourth-order Runge–Kutta algorithm for different sets of parameter values yielded an interior equilibrium point. Figure 1 presents the phase plane y vs. x_2 when $\alpha = 1$, $k = 0.25$, $\beta = 0.3$, $d = 0.15$, $\eta = 1$, $b = 0.4$, $q = 0.02$, and $e = 1$. Three equilibrium points $\bar{E}_0 = (0, 0, 0)$, $\bar{E}_1 = (0.561111, 0.623456, 0)$, and $\bar{E}_2 = (0.380488, 0.369346, 0.198195)$ were found. We observe that $\bar{E}_0$ and $\bar{E}_1$ are unstable and $\bar{E}_2$ is asymptotically stable. The results are in accordance with the theoretical results presented in section 3.

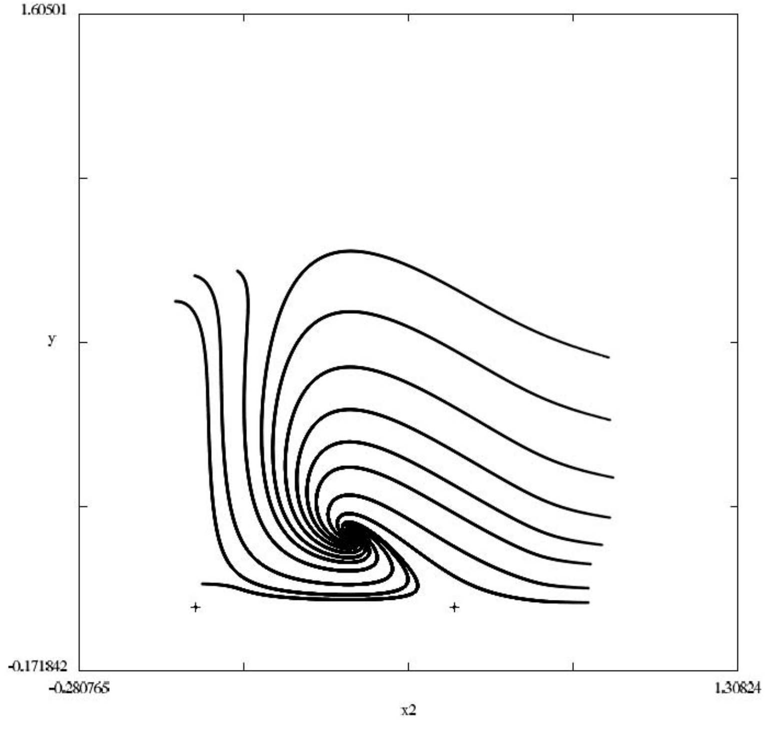


FIGURE 1. The phase plane y vs. x_2 of system (2.1) for the values of $\alpha = 1$, $\eta = 1$, $k = 0.25$, $\beta = 0.3$, $d = 0.15$, $b = 0.4$, $q = 0.02$ and $e = 1$.

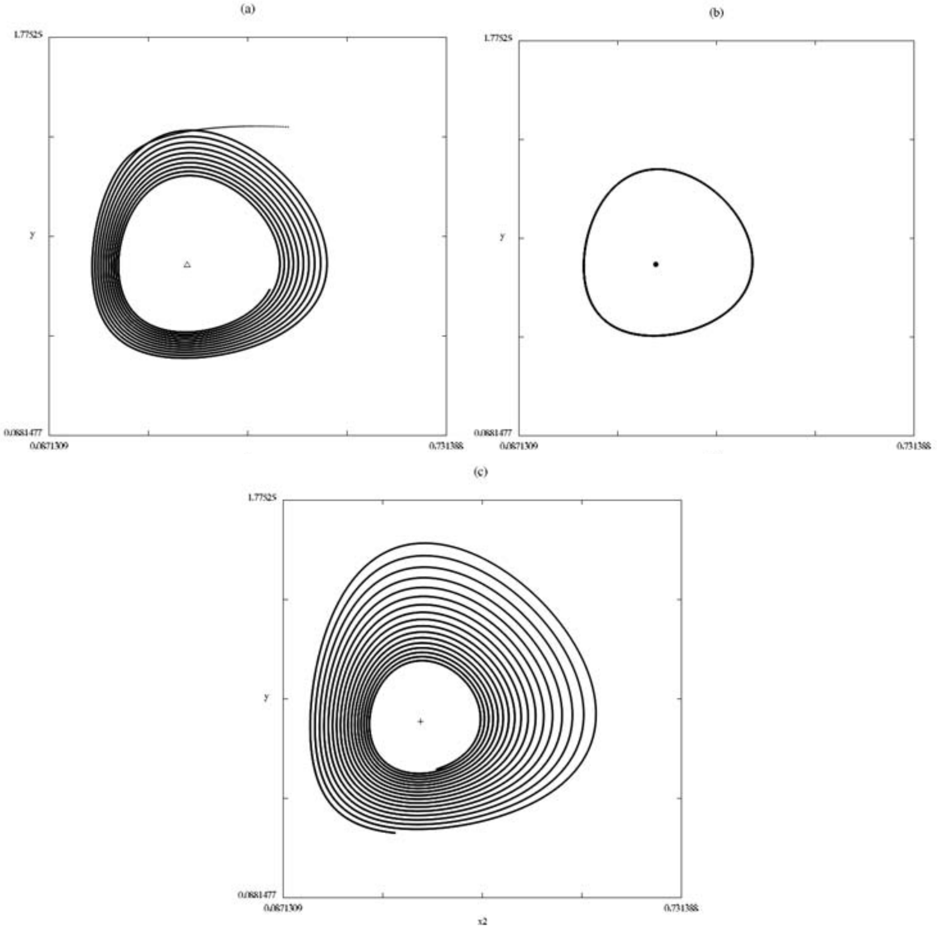


FIGURE 2. A degenerate Hopf-bifurcation at $\eta = -0.01236$, $\alpha = 1$, $k = 0.25$, $\beta = 0.3$, $d = 0.15$, $b = 0.4$, $q = 0.02$ and $e = 1$. a) The equilibrium point $\bar{E}_2$ is asymptotically stable at $protect\eta = 0$; (b) $\bar{E}_2$ is stable at $\eta = -0.01236$; (c) $\bar{E}_2$ is unstable at $\eta = -0.03$.

Varying the value of η from 1 to 0, the equilibrium point $\bar{E}_3$ is still asymptotically stable, with a very slow convergence rate (the real parts of the complex eigenvalues are still in the left half-plane) . Decreasing the value of η further, a Hopf-bifurcation occurs at $\eta = -0.01236$ (the complex eigenvalues cross the imaginary axis where the real parts of the complex eigenvalues become zero). For η less than a critical value, the real parts of the complex eigenvalues become positive — i.e., the asymptotically stable fixed point becomes unstable at values of η less than -0.01236 (Figure 2). This type of Hopf-bifurcation is called degenerate. (We remark that negative η values have no physical meaning in the current context, but in passing we wished to illustrate the presence of the Hopf-bifurcation.)

The mathematical model we have proposed consists of three nonlinear ordinary differential equations, corresponding to an immature population, a mature population, and their predator. The predator can feed on either stage of the prey, but instead of choosing individuals at random the predator catches a member of the immature or mature prey population proportional to their abundance — i.e., the predator feeds preferentially on the most numerous stage species. This behaviour is termed predator switching.

We have found conditions for the stability of the equilibria. The dynamical behaviour shows that the system is locally stable in some region of parametric space around a positive interior equilibrium, and unstable in some other region of parametric space. It was also observed that the system is asymptotically stable around the positive equilibrium if the ratio of the young and adult prey species at any time has the same value as at the equilibrium point. We studied the maximum sustainable yield of the system. The economic and biological viewpoint for renewable resource management led us to study exploitation of the mature population, when it is desirable to have a unique stable positive equilibrium in order to plan harvesting strategies and maintain sustainable development of the ecosystem. We obtained threshold harvesting conditions for the mature population, and also considered its optimal harvesting.

References

[1] C. R. Fisher-Piett, *Soc. Biolgeogr*, **92**, 47-48 (1934).

[2] J. H. Lawton, J. R. Beddington and R. Bonser, Switching in invertebrate predators. *Ecological Studies*, **9**, 144-158 (1974).

[3] W. W. Murdoch and A. Oaten, Predation and population stability. *Adv. Ecol. Res.*, **9**, 1-131 (1975).

[4] C. S. Holling, Principles of insect predation. *Ann. Rev. Entomol.* **6**, 163-182 (1961).

[5] F. Takahashi, Reproduction curve with two equilibrium points: a consideration of the fluctuation of insect population. *Res. Pop. Ecol.* **47**, 733-745 (1964).

[6] R. M. May, *Stability and Complexity in model ecosystems.* Princeton, NJ: Princeton University Press (1973).

[7] R. M. May, Some mathematical problems in biology, Providence, RI. *Am. Math. Soc.*, **4**, 11-29 (1974).

[8] W. W. Murdoch and A. Oaten, Predation and population stability *Adv. Ecol. Res.*, **9**, 1-131 (1975).

[9] J. Roughgarden and M. Feldman, Species packing and predation pressure, *Ecology*, **56**, 489-492 (1975).

[10] M. Tansky, Switching effects in prey-predator system, *J. Theor. Biol.* **70**, 263-271 (1978).

[11] Prajneshu and P. Holgate, A prey-predator model with switching effect. *J. Theor. Biol.* **125**, 61-61-66 (1987).

[12] Q. J. A. Khan, B. S. Bhatt and R. P. Jaju, Stability of a switching model with two habitats and a predator, *J. Phys. Soc. Jpn.*, **63**, 1995-2001 (1994).

[13] Q. J. A. Khan, B. S. Bhatt and R. P. Jaju, Hopf Bifurcation analysis of a predator-prey system involving switching, *J. Phys. Soc. Jpn.*, 65, 3, 864-867 (1996).

[14] M. Lloyd and H. S. Dybas, The periodical cicada problem, *Evolution*, **20**, 133-149, 466-505 (1966).

[15] E. D. LeCaven, C. Kipling, J. C. McCormack, A study of the numbers, biomass and year-class strengths of perch (perca fluviatillis L) in winteremiere, *J. Anim., Ecol.* **46**, 281-306 (1977).

[16] L. Nielsen, Effect of Walleye (Stizostedion Vitreun), Predation on Juvenile mortality and recruitment of yellow pereh (perca flavereens) in Oneida lake, New York. *Can. J. Fish. Aquat. Sci.* **37**, 11-19 (1980).

[17] H. I. Freedman, J. W.-H. So and J. Wu, A model for the growth of a population exhibiting stage structure: Cannibalism and cooperation, *J. Comp. and App. Math.*, **52**, 177-198 (1994).

[18] W. S. C. Gurney, R. M. Nisbet and J. H. Lawton, The systematic formulation of tractable single-species population models incorporating age structure, *J. Animal Ecol.* **52**, 479-495 (1983).

[19] X. Song and L. Chen, Optimal harvesting and stability for a two-species competitive system with stage structure, *Math. Biosci.*, **170**, 173-186 (2001).

[20] Q. J. A. Khan, E. V. Krishnan and M. A. Al-Lawatia, A stage structure model for the growth of a population involving switching and cooperation, *Z. Angew. Math. Mech*, **82**, 125-135 (2002).

[21] X. Zhang, L. Chen and A. U. Neumann, The stage structured predator-prey model and optimal harvesting policy, *Math. Biosci*, **168**, 201-210 (2000).

[22] W. Wang and L. Chen, A predator-prey system with stage structure for predator, *Comput. Math. Appl.* **33**, 207 (1997).

[23] C. W. Clark, *Mathematical Bioeconomics: The Optimal management of renewable resources*, 2nd Ed., Wiley, New York (1990).

[24] A. W. Leng, Optimal harvesting-coefficient control of steady-state prey-predator diffusive Volterra-Lotka systems, *Appl. Math. Optim*, **31**, 219 (1995).

[25] D. K. Bhattacharya and S. Begun, Bioeconomic equilibrium of two-species systems I, *Math. Biosci.*, **135**, 111 (1996).

[26] T. L. John, *Variational Calculus and Optimal control*, Springer, New York (1996).

Appendix A

Let

$$P(x) = b_5x^5 + b_4x^4 + b_3x^3 + b_2x^2 + b_1x + b_0,$$

then

$$\begin{aligned} P^{'}(x) &= 5b_5x^4 + 4b_4x^3 + 3b_3x^2 + 2b_2x + b_1, \\ P''(x) &= 20b_5x^3 + 12b_4x^2 + 6b_3x + 2b_2, \\ P'''(x) &= 6(10b_5x^2 + 4b_4x + b_3), \end{aligned}$$

where

$$\begin{aligned} b_0 &= -\alpha_1, \\ b_1 &= k_1 + \beta_1, \\ b_2 &= -k_1 - qe_1 + \eta_1 d_1 - \alpha_1, \\ b_3 &= k_1 + 2\beta_1 + \eta_1 d_1, \\ b_4 &= -(k_1 + qe_1), \\ b_5 &= \beta_1. \end{aligned}$$

Assume that $b_2 > 0$. The polynomial $P(x)$ is strictly positive for large x and negative for $x = 0$. Therefore $P(x)$ has at least one positive zero and consequently the system (2.2) has at least one positive equilibrium. By Descarte's Rule $P(x)$ has either three positive zeros or one positive zero. Suppose that $P(x)$ has three positive zeros and note that $P'(x) > 0$ for large x and $x = 0$. Again by Descarte's rule $P'(x)$ will have two positive zeros. This leads after some simple algebra for $P''(x)$ to have one positive root. But close examination of $P''(x)$ using Descarte's Rule shows that this cannot happen as $P''(x)$ has either two zeros or no zero. Therefore $P(x)$ must have a single positive zero. This completes the proof.

Appendix B

To show that the non-zero equilibrium is locally stable to small perturbations, we need to show the Routh–Hurwitz conditions are satisfied:

(a) $a_1 > 0$, $a_3 > 0$;

(b) $a_1a_2 > a_3$.

Clearly

$$a_1 = A\bar{x} - L = \beta_1\bar{x} + \frac{\hat{x}_2^2\hat{y}}{(\hat{x}_1 + \hat{x}_2)^2} + \frac{\alpha_1}{\bar{x}} + \frac{\hat{x}_1^2\hat{y}}{\bar{x}\,(\hat{x}_1 + \hat{x}_2)^2} + \eta_1\hat{x}_1 > 0. \qquad \text{(B.1)}$$

To show $a_3 > 0$. Write

$$\begin{aligned} P &= \frac{C\hat{x}_2}{\bar{x}} + \frac{D\hat{x}_2^2}{\bar{x}} + C\hat{x}_2 + D\hat{x}_1\hat{x}_2, \\ Q &= C\hat{x}_1\bar{x} + D\hat{x}_1\hat{x}_2\bar{x} + C\hat{x}_1\bar{x}^2 + D\hat{x}_1\bar{x}^2, \\ S &= C\hat{x}_1\hat{x}_2 + D\hat{x}_1\hat{x}_2\bar{x} + C\hat{x}_1\bar{x}^2 + D\hat{x}_1^2\bar{x}^2, \end{aligned} \tag{B.2}$$

so that

$$a_3 = \frac{1}{(1+\bar{x})}\left[BP + AQ + \eta S\right]. \tag{B.3}$$

We have

$$\begin{aligned} P = {} & \frac{-\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)\ \hat{x}_2^2}{\left(\hat{x}_1 + \hat{x}_2\right)^2 \hat{x}_1} + \frac{2\hat{y}\ \hat{x}_2^3}{\left(\hat{x}_1 + \hat{x}_2\right)\hat{x}_1} - \frac{\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)\hat{x}_2}{\left(\hat{x}_1 + \hat{x}_2\right)^2} \\ & + \frac{2\hat{y}\ \hat{x}_1\hat{x}_2}{\hat{x}_1 + \hat{x}_2} = \frac{\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)\ \hat{x}_2}{\hat{x}_1\left(\hat{x}_1 + \hat{x}_2\right)} > 0, \end{aligned} \tag{B.4}$$

$$\begin{aligned} Q = {} & C\ \hat{x}_1\bar{x} + D\ \hat{x}_1\hat{x}_2\bar{x} + C\ \hat{x}_1\bar{x}^2 + D\ \hat{x}_1^2\bar{x}^2 \\ = {} & \frac{-\hat{y}\left(\hat{x}_1^2 + \hat{x}_2^2\right)\hat{\hat{x}}_1\bar{x}}{\left(\hat{x}_1 + \hat{x}_2\right)^2} + \frac{2\hat{y}\hat{x}_1\hat{x}_2\bar{x}}{\left(\hat{x}_1 + \hat{x}_2\right)} - \frac{\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)\ \hat{x}_1\bar{x}^2}{\left(\hat{x}_1 + \hat{x}_2\right)^2} \\ & + \frac{2\hat{y}\ \hat{x}_1^2\bar{x}^2}{\left(\hat{x}_1 + \hat{x}_2\right)} = \frac{\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)\hat{x}_1^2}{\hat{x}_2^2\left(\hat{x}_1 + \hat{x}_2\right)} > 0, \end{aligned} \tag{B.5}$$

$$S = C\ \hat{x}_1\hat{x}_2 + D\hat{x}_1\hat{x}_2^2 = \frac{\hat{y}\ \hat{x}_1\hat{x}_2\left(\hat{x}_2^2 + 2\hat{x}_1\hat{x}_2 - \hat{x}_1^2\right)}{\left(\hat{x}_1 + \hat{x}_2\right)^2}. \tag{B.6}$$

Now, from (B.3), (B.4), (B.5) and (B.6),

$$\begin{aligned} a_3 = \frac{\hat{y}\ \hat{x}_2}{\left(\hat{x}_1 + \hat{x}_2\right)^2}\bigg\{ & \frac{\alpha_1\hat{x}_2\left(\hat{x}_1^2 + \hat{x}_2^2\right)}{\hat{x}_1} + \frac{\hat{x}_1\hat{x}_2\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)}{\left(\hat{x}_1 + \hat{x}_2\right)^2} \\ & + \frac{\beta_1\hat{x}_1^2\left(\hat{x}_1^2 + \hat{x}_2^2\right)}{\hat{x}_2^2} + \frac{\hat{x}_1^2\hat{y}\ \left(\hat{x}_1^2 + \hat{x}_2^2\right)}{\left(\hat{x}_1 + \hat{x}_2\right)^2} \\ & + \frac{\eta_1\hat{x}_1\hat{x}_2}{\left(\hat{x}_1 + \hat{x}_2\right)}\left(\hat{x}_2^2 + 2\hat{x}_1\hat{x}_2 - \hat{x}_1^2\right)\bigg\}. \end{aligned} \tag{B.7}$$

Recall that

$$\hat{y} = \left[\frac{1+\bar{x}}{\bar{x}}\left(\frac{\alpha_1}{\bar{x}} - \kappa_1 - \beta_1 - \eta_1\hat{x}_1\right)\right] > 0.$$

So

$$\alpha_1\hat{x}_2 - \eta_1\hat{x}_1^2 > 0. \tag{B.8}$$

Thus, $a_3 > 0$. This completes the proof of (a).

(b) We must show that $a_1 a_2 > a_3$. It follows from (3.8) that

$$\begin{aligned}(A\bar{x} - L)\left[-LA\bar{x} + \frac{\hat{x}_2}{(1+\bar{x})}(C + D\ \hat{x}_2) - AB + \frac{\hat{x}_1\bar{x}\ C}{1+\bar{x}} + \frac{D\ \bar{x}\ \hat{x}_1^2}{1+\bar{x}}\right] \\ > -\frac{L\ \hat{x}_2 C}{1+\bar{x}} - \frac{L\ D\ x_2^2}{1+\bar{x}} + \frac{B\ \hat{x}_2}{1+x}(C + D\hat{x}_1) + \frac{AC\ \hat{x}_1\bar{x}}{1+\bar{x}} + \frac{A\ D\ \hat{x}_1\hat{x}_2\bar{x}}{1+\bar{x}} \\ + \frac{AC\ \hat{x}_1\bar{x}^2}{1+\bar{x}} + \frac{A\ D\ \hat{x}_1^2\hat{x}^2}{1+\bar{x}}. \qquad \text{(B.9)}\end{aligned}$$

After some simplifications (B.9) can be expressed as

$$\begin{aligned}&-\frac{AC\ \bar{x}\ \ (\hat{x}_1 - \hat{x}_2)}{1+\bar{x}} + \frac{BC\ \ (\hat{x}_1 - \hat{x}_2)}{1+\bar{x}} - \frac{A\ D\ \hat{x}_2\bar{x}}{1+\bar{x}}(\hat{x}_1 - \hat{x}_2) \\ &+ \frac{B\ D\ \hat{x}_1}{1+\bar{x}}(\hat{x}_1 - \hat{x}_2) + \eta_1\left[A_1^2\hat{x}_1\bar{x}^2 + \eta_1\hat{x}_1^2 A\bar{x} + \frac{C\ \hat{x}_1^2\bar{x}}{1+\bar{x}} + \frac{D\ \bar{x}\ \bar{x}_1^3}{1+\bar{x}} + A\ B\ \hat{x}_1\right] > 0. \\ &\qquad \text{(B.10)}\end{aligned}$$

All the terms inside square bracket of inequality (B.10) are positive except $\dfrac{C\ \hat{x}_1^2\ \bar{x}}{1+\bar{x}}$, but

$$\begin{aligned}\frac{C\hat{x}_1^2\bar{x}}{1+\bar{x}} + \frac{D\bar{x}\hat{x}_1^3}{1+\bar{x}} &= \frac{\hat{x}_1^2\bar{x}}{1+\bar{x}}(C + D\hat{x}_1) \\ &= \frac{\hat{x}_1^2\bar{x}\bar{y}}{(1+\bar{x})(\hat{x}_1+\hat{x}_2)^2}\left[(\hat{x}_1 - \hat{x}_2)(\hat{x}_1 + \hat{x}_2) + 2\hat{x}_1\hat{x}_2\right] > 0. \qquad \text{(B.11)}\end{aligned}$$

If $\hat{x}_1 > \hat{x}_2$, then

$$\begin{aligned}&\frac{AC\ \bar{x}\ \ (\hat{x}_1 - \hat{x}_2)}{1+\bar{x}} + \frac{BC\ \ (\hat{x}_1 - \hat{x}_2)}{1+\bar{x}} - \frac{AD\ \hat{x}_2\bar{x}\ \ (\hat{x}_1 - \hat{x}_2)}{1+\bar{x}} \\ &+ \frac{BD\ \hat{x}_1(\hat{x}_1 - \hat{x}_2)}{1+\bar{x}} \\ &= \frac{(\hat{x}_1\hat{x}_2)}{(\hat{x}_1 + \hat{x}_2)}\left[C\ \ (\alpha_1\hat{x}_2 - \beta_1\hat{x}_1) + D\ \hat{x}_1\hat{x}_2(\alpha_1 - \beta_1)\right] \\ &+ \frac{\hat{x}_1\hat{x}_2\hat{y}\ \ (\hat{x}_1 - \hat{x}_2)^2}{(\hat{x}_1 + \hat{x}_2)^3}\left[C + D\ \ (\hat{x}_1 + \hat{x}_2)\right]. \qquad \text{(B.12)}\end{aligned}$$

Let

$$\begin{aligned}U &= C\ \ (\alpha_1\hat{x}_2 - \beta_1\hat{x}_1) + D\ \hat{x}_1\hat{x}_2(\alpha_1 - \beta_1), \\ V &= C + D\ \ (\hat{x}_1 + \hat{x}_2). \qquad \text{(B.13)}\end{aligned}$$

$V > 0$ because $C + D\ \hat{x}_1 > 0$ from (B.11). To show that $U > 0$ where $\alpha_1\hat{x}_2 - \beta_1\hat{x}_1 < \alpha_1\hat{x}_2 - \beta_1\hat{x}_2$, since we are assuming $\hat{x}_1 > \hat{x}_2$ and $D\ \hat{x}_1 > C$ from (B.11). So, $U > 0$. Hence, $a_1 a_2 > a_3$ if $\bar{x} > 1$. This completes the proof.

Qamar J. A. Khan
Department of Mathematics and Statistics
College of Science
Sultan Qaboos University
P.O. Box 36, P. C. 123, Al-Khod, Muscat
Sultanate of Oman
e-mail: qjalil@squ.edu.om

Lakdere Benkherouf
Department of Statistics and Operations Research
College of Science
Kuwait University, P.O.Box 5969, Safat 13060
Kuwait
e-mail: lakdereb@kuc01.kuniv.edu.kw

Nejib Smaoui
College of Science
Department of Mathematics and Computer Science
P.O.Box 5969, Safat 13060
Kuwait
e-mail: nsmaoui64@yahoo.com

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 81–94

Insecticidal Bt Crops Under Massive Bt-resistant Pest Invasion: Mathematical Simulation

Alexander B. Medvinsky, Maria M.Gonik, Yuri V. Tyutyunov, Bai-Lian Li, Alexey V. Rusakov and Horst Malchow

Abstract. There is growing public concern that pests may develop resistance to Bt toxins produced by genetically modified Bt plants. We develop and analyse a conceptual reaction-diffusion model of the agricultural ecosystem, to simulate the Bt-resistant insect massive invasion when the insect fecundity rate is limited by the number of females rather than by the mating frequency. We show by computer simulations that reproduction of Bt-resistant pests is a factor that significantly affects the Bt plant biomass under the Bt-resistant pest invasion. We demonstrate that periodical Bt plant sowing can lead to both regular and irregular oscillations in Bt plant and Bt-resistant insect biomass. The character of the oscillations (regular or irregular) is shown to be dependent on local insect fluxes, which are characterised by the diffusion number Ξ. The oscillations are irregular if $\Xi \geq 0.02$, but otherwise the oscillations of the plant and insect biomass are regular.

Mathematics Subject Classification (2000). Primary 92D40.

Keywords. Invasion, Transgenic plants, Bt-resistant pests, Mathematical modelling.

1. Introduction

Genetically modified plants, which express genes encoding insecticidal toxins from the bacterium *Bacillus thuringiensis* (Bt plants), have become an important component of modern agricultural practice in many countries [2, 12, 23, 6]. Whenever pests are exposed to pesticides, selection occurs for alleles that confer resistance

This work was partially supported by ISTC, DFG, RFBR, NSF Biocomplexity Program, and by the University of California Agricultural Experiment Station. We are thankful to the anonymous referees for helpful suggestions.

to those pesticides. The increase in resistance to Bt sprays in the field, as well as the greenhouse and laboratory selection of resistant strains of several major pests, demonstrate that the probability of appearance and expansion of Bt resistant insects can build up [27, 25, 1, 7, 5, 28, 9, 26].

Computer simulations have been used to reveal key factors and relationships, to evaluate the environmental impact of Bt-resistant pest invasions. In particular, we have developed a conceptual reaction-diffusion model that describes the main features of the interaction between genetically modified Bt crops and pests resistant to Bt toxins [18, 17, 16]. This model is based on the assumption that the effective fecundity rate of the recessive Bt-resistant pests essentially depends on the insect mating frequency. Such an interrelation between the mating frequency and the fecundity rate is typical for populations that are characterised by modest mating frequency [4], when the dependence between the insect fecundity rate and the density of the insect population is quadratic [8, 4].

In this paper, we consider the spatio-temporal dynamics of an agricultural ecosystem consisting of transgenic Bt plants, insects susceptible to Bt toxins, and adapted Bt-resistant insects carrying a recessive mutation enabling them to grow on Bt plants, with the assumption of massive Bt-resistant pest invasion. Under massive pest invasion, the insect fecundity rate is limited by the number of females rather then by the mating frequency. Consequently, under the assumption of constant 50/50 sex ratio, the insect fecundity rate is directly proportional to the density of the insect population [4]. We show that reproduction of the Bt-resistant pests is a factor that significantly affects the Bt plant biomass under the Bt-resistant pest invasion.

2. Model

Conceptual minimal models have been shown to be an appropriate tool for searching and understanding basic mechanisms underlying population dynamics [13, 30, 10, 11, 14, 20, 19, 4, 21]. We follow this approach to develop a mathematical model of the Bt plant – pest population dynamics. Let

$$I = I_{rr} + I_{rs} + I_{ss}, \tag{1}$$

where I is the total insect biomass, while I_{rr}, I_{ss}, and I_{rs} respectively denote the biomass of Bt-resistant (rr), Bt-susceptible (ss), and heterozygote (rs) insects. We assume a constant 50/50 male-female ratio for all the genotypes: rr, rs, and ss. Table 1 lists all the possible outcomes of matings between different genotypes. The probability of mating between males of genotype i and females of genotype j (i,j = rr, rs, ss) is $I_i I_j / I^2$. Similarly, the probability of mating between males of genotype j and females of genotype i is $I_j I_i / I^2$. We also assume that the effective insect fecundity η_{ij} depends on the plant biomass P. The total increase in the ss-insect biomass (cf. Table 1) can then be represented as $\frac{1}{2}(\eta_{ss}(P)/I)\left(I_{ss} + \frac{1}{2}I_{rs}\right)^2$ where the insect biomass I is defined in equation (1).

Similarly, the total increase in the rr-insect biomass is $\frac{1}{2}(\eta_{rr}(P)/I)\left(I_{rr}+\frac{1}{2}I_{rs}\right)^2$. The total increase in the rs-insect biomass resulting from the insect reproduction is given by $(\eta_{rs}(P)/I)\left(I_{ss}+\frac{1}{2}I_{rs}\right)\left(I_{rr}+\frac{1}{2}I_{rs}\right)$.

Table 1

FEMALE	MALE	OUTCOME
ss	ss	ss
rr	rr	rr
ss	rr	rs
rr	ss	rs
ss	rs	$\frac{1}{2}ss+\frac{1}{2}rs$
rs	ss	$\frac{1}{2}ss+\frac{1}{2}rs$
rs	rr	$\frac{1}{2}rs+\frac{1}{2}rr$
rr	rs	$\frac{1}{2}rs+\frac{1}{2}rr$
rs	rs	$\frac{1}{4}ss+\frac{1}{4}rr+\frac{1}{2}rs$

Taking into account all the foregoing results, we simulate the dynamics of Bt plant and insect biomass at any point X and time τ in the following mathematical model:

$$\begin{aligned}\frac{\partial P}{\partial \tau} &= rP\left(1-\frac{P}{K}\right)-\frac{C_1P}{C_2+P}\\ &-\frac{\chi_{[nT,nT+\epsilon T]}(\tau)}{2I}\left[\eta_{ss}(P)\left(I_{ss}+\frac{1}{2}I_{rs}\right)^2+2\eta_{rs}(P)\left(I_{ss}+\frac{1}{2}I_{rs}\right)\left(I_{rr}+\frac{1}{2}I_{rs}\right)\right]\\ &+\frac{\chi_{[nT,nT+\epsilon T]}(\tau)}{2I}\left[\eta_{rr}(P)\left(I_{rr}+\frac{1}{2}I_{rs}\right)^2\right], \qquad (2)\end{aligned}$$

$$\frac{\partial I_{ss}}{\partial \tau}=\frac{k_{ss}C_1P}{C_2+P}I_{ss}-\mu_{ss}I_{ss}+\frac{\chi_{[nT,nT+\epsilon T]}(\tau)}{2I}\delta_{ss}\eta_{ss}(P)\left[I_{ss}+\frac{1}{2}I_{rs}\right]^2+D\frac{\partial^2 I_{ss}}{\partial X^2}, \qquad (3)$$

$$\begin{aligned}\frac{\partial I_{rs}}{\partial \tau} &= \frac{k_{rs}C_1P}{C_2+P}I_{rs}-\mu_{rs}I_{rs}\\ &+ \frac{\chi_{[nT,nT+\epsilon T]}(\tau)}{I}\delta_{rs}\eta_{rs}(P)\left[\left(I_{ss}+\frac{1}{2}I_{rs}\right)\left(I_{rr}+\frac{1}{2}I_{rs}\right)\right]+D\frac{\partial^2 I_{rs}}{\partial X^2}, \qquad (4)\end{aligned}$$

$$\frac{\partial I_{rr}}{\partial \tau}=\frac{k_{rr}C_1P}{C_2+P}I_{rr}-\mu_{rr}I_{rr}+\frac{\chi_{[nT,nT+\epsilon T]}(\tau)}{2I}\delta_{rr}\eta_{rr}(P)\left[I_{rr}+\frac{1}{2}I_{rs}\right]^2+D\frac{\partial^2 I_{rr}}{\partial X^2}, \qquad (5)$$

where the characteristic function of the interval $[nT, nT + \epsilon T]$ is denoted and defined by

$$\chi_{[nT,nT+\epsilon T]}(t) = \begin{cases} 1, & t \in [nT, nT + \epsilon T] \\ 0, & t \notin [nT, nT + \epsilon T]. \end{cases}$$

Equations (2) – (5) describe the spatio-temporal dynamics of the plant $P(X, \tau)$, Bt-resistant insect $I_{rr}(X, \tau)$, Bt-susceptible insect $I_{ss}(X, \tau)$ and heterozygote insect $I_{rs}(X, \tau)$ biomass, respectively. The parameters r and K denote the intrinsic growth rate and the carrying capacity of the plant population; μ_{rr}, μ_{ss} and μ_{rs} are the combined mortality rates per unit mass of the rr-, ss- and rs-insects. The constants C_1 and C_2 parametrise the saturating functional responses. D is the insect diffusion coefficient. The terms in the square brackets operate only throughout the reproduction period of length ϵT. The functions $\eta_{rr}(P), \eta_{ss}(P)$ and $\eta_{rs}(P)$ are the effective fecundity rates of the rr-, ss- and rs-insects, respectively. Thus the terms involving these functions of the plant biomass P account for both the insect reproduction and maturation of the insect offspring. We allow for the fact that Bt plants selectively affect larvae [22], embodied in the form of the functions $\eta_{ss}(P)$ and $\eta_{rs}(P)$. Thus we suppose that the ss- and rs-insects are both liable to the action of Bt toxins, although not to the same extent. Consequently, their effective fecundity rates fall as soon as the Bt plant biomass P is more than a critical value P_{cr}, so that

$$\eta_{ss}(P) = \eta_s^o P exp\left(-\frac{\lambda P}{K}\right), \tag{6}$$

$$\eta_{rs}(P) = \eta_{rs}^o P\left(exp\left(-\frac{\lambda P}{K}\right) + \omega_{rs}^*\right), \tag{7}$$

where η_s^o, η_{rs}^o and $\lambda > 1$ are constants and $P_{cr} = K/\lambda$. In contrast, the effective fecundity rate $\eta_{rr}(P)$ of Bt-resistant insects does not tend to decrease with an increase in the Bt plant biomass, so we suppose that

$$\eta_{rr}(P) = \eta_r^o P, \tag{8}$$

where η_r^o is a constant. The parameters k_{ss}, k_{rs} and k_{rr} in equations (3), (4) and (5) are the respective yield coefficients of the ss-, rs- and rr-insects to the plants; and δ_{ss}, δ_{rs} and δ_{rr} are the respective yield coefficients of larvae of the ss-, rs- and rr-insects. Our model also accounts for Bt plants being periodically sown, implying that

$$P = P_0 \text{ at } \tau = nT, \text{ where } n = 0, 1, 2, \ldots. \tag{9}$$

We set $T = 150$ days, which roughly corresponds to the length of a normal growing season [24].

For later convenience, we introduce the dimensionless variables

$$p = \frac{P}{K}, \quad i_{ss} = \frac{I_{ss}}{K}, \quad i_{rs} = \frac{I_{rs}}{K}, \quad i_{rr} = \frac{I_{rr}}{K}, \quad t = \frac{A}{T}\tau, \quad x = X\sqrt{\frac{A}{DT}},$$

to render our mathematical model (1) – (9) in dimensionless form:

$$\begin{aligned}\frac{\partial p}{\partial t} &= \alpha p(1-p) - \frac{\beta p}{\gamma + p} i \\ &\quad - \frac{\chi_{[nA,nA+\epsilon A]}(t)}{i}\left[\omega_{ss}(p)\left(i_{ss} + \frac{1}{2} i_{rs}\right)^2 + \omega_{rs}(p)\left(i_{ss} + \frac{1}{2} i_{rs}\right)\left(i_{rr} + \frac{1}{2} i_{rs}\right)\right] \\ &\quad + \frac{\chi_{[nA,nA+\epsilon A]}(t)}{i}\left[\omega_{rr}(p)\left(i_{rr} + \frac{1}{2} i_{rs}\right)^2\right],\end{aligned} \tag{10}$$

$$\frac{\partial i_{ss}}{\partial t} = \frac{k_{ss}\beta p}{\gamma + p} i_{ss} - \nu_{ss} i_{ss} + \frac{\chi_{[nA,nA+\epsilon A]}(t)}{i}\delta_{ss}\omega_{ss}(p)\left[i_{ss} + \frac{1}{2} i_{rs}\right]^2 + \frac{\partial^2 i_{ss}}{\partial x^2}, \tag{11}$$

$$\begin{aligned}\frac{\partial i_{rs}}{\partial t} &= \frac{k_{rs}\beta p}{\gamma + p} i_{rs} - \nu_{rs} i_{rs} \\ &\quad + \frac{2\chi_{[nA,nA+\epsilon A]}(t)}{i}\delta_{rs}\omega_{rs}(p)\left[\left(i_{ss} + \frac{1}{2} i_{rs}\right)\left(i_{rr} + \frac{1}{2} i_{rs}\right)\right] + \frac{\partial^2 i_{rs}}{\partial x^2},\end{aligned} \tag{12}$$

$$\frac{\partial i_{rr}}{\partial t} = \frac{k_{rr}\beta p}{\gamma + p} i_{ss} - \nu_{rr} i_{rr} + \frac{\chi_{[nA,nA+\epsilon A]}(t)}{i}\delta_{rr}\omega_{rr}(p)\left[i_{rr} + \frac{1}{2} i_{rs}\right]^2 + \frac{\partial^2 i_{rr}}{\partial x^2}, \tag{13}$$

$$i = i_{ss} + i_{rs} + i_{rr}, \tag{14}$$

$$\omega_{ss}(p) = \omega_s^0 p exp(-\lambda p), \tag{15}$$

$$\omega_{rs}(p) = \omega_{rs}^0 p \left(exp(-\lambda p) + \omega_{rs}^*\right), \tag{16}$$

$$\omega_{rr}(p) = \omega_r^0 p, \tag{17}$$

$$p = p_0 \text{ at } t = nA, \quad \text{where } \; n = 0, 1, 2, \ldots. \tag{18}$$

Here we have set $\alpha = rT/A$, $\beta = C_1 T/A$, $\gamma = C_2/K$, $\nu_{ss} = \mu_{ss} T/A$, $\nu_{rs} = \mu_{rs} T/A$, $\nu_{rr} = \mu_{rr} T/A$, $\omega_s^0 = (KT/2A)\,\eta_s^0$, $\omega_{rs}^0 = (KT/A)\,\eta_{rs}^0$, $\omega_r^0 = (KT/2A)\,\eta_r^0$, and again use the characteristic function

$$\chi_{[nA,nA+\epsilon A]}(t) = \begin{cases} 1, & t \in [nA, nA + \epsilon A] \\ 0, & t \notin [nA, nA + \epsilon A]. \end{cases}$$

For our numerical work described below, we put $A = T$ so that $t = \tau$, $x = X\sqrt{1/D}$, $\alpha = r$, $\beta = C_1$, $\gamma = C_2/K$, $\nu_{ss} = \mu_{ss}$, $\nu_{rs} = \mu_{rs}$, $\nu_{rr} = \mu_{rr}$, $\omega_s^0 = (K/2)\eta_s^0$, $\omega_{rs}^0 = K\eta_{rs}^0$, and $\omega_r^0 = (K/2)\eta_r^0$.

To investigate the dynamics of the plant – insect model system, we solved equations (10) – (18) numerically by an explicit finite difference scheme. Neuman zero-flux boundary conditions were used. The initial conditions assumed at $t = 0$ were that plants and susceptible (ss) insects are homogeneously distributed in space, while heterozygote (rs) insects are concentrated in a small region at the centre of the whole domain and Bt-resistant (rr) insects are absent. The mesh step sizes chosen were $\Delta x = 1$ and $\Delta t = 0.001$, sufficiently small to ensure the accuracy of the simulations.

3. Results

The study of the after-effects of an invasion of Bt-resistant insects is a main objective of this work, and there is a key parameter that could characterise the Bt-resistant pest's survival and impact on the plant biomass — viz. the so-called growth number, defined as [18]

$$G = \frac{k\beta}{\gamma + 1} - \nu.$$

Our computer simulations show that the initially small population of heterozygote rs-insects can trigger invasion of the model Bt-resistant insects into the agricultural ecosystem, if the growth number of resistant insects (G_{rr}) is positive and greater than the growth numbers of susceptible insects (G_{rs} and G_{ss}) — i.e.,

$$G_{rr} > 0, \tag{19}$$

$$G_{rr} > G_{rs} \text{ and } G_{rr} > G_{ss}. \tag{20}$$

If conditions (19) and (20) are met, Bt-susceptible pests are displaced by Bt-resistant ones. As this takes place, the susceptible insect biomass decreases towards 0 (i.e., $i_{ss} = 0$ and $i_{rs} = 0$).

We have also shown that the parameter space $(\nu, k_{rr}\beta)$ is subdivided into two regions, characterised by different types of intrinsic plant-insect dynamics. The intrinsic dynamics imply that periodical sowing of plants is absent, and all of the terms on the right-hand sides of equations (10) – (13) are taken into account. At $\omega_r^0 = 0$, an Hopf bifurcation takes place in the $(\nu, k_{rr}\beta)$ parameter space if $G = G^H$, where

$$G^H = \frac{1}{2}(k_{rr}\beta - \nu). \tag{21}$$

Figure 1 presents the function $k_{rr}\beta = 2G^H + \nu$, which demarcates these two regions in the $(\nu, k_{rr}\beta)$ parameter space. In the lower region 2, the plant-insect intrinsic dynamics exhibits limit cycles, whereas in the upper region 1 there are stable nodes. The Hopf bifurcation occurs if the parameter $\omega_r^0 \neq 0$, so that the oscillatory behaviour of the plant-insect system can invade region 1.

Figure 2 shows the dependence of the average plant biomass

$$\langle p \rangle = \frac{1}{Lt} \int_0^t \int_0^L p(x, t^{`}) dx dt^{`}, \quad t = nA \tag{22}$$

on the duration of the insect reproduction period ϵA in each of the two regions in Figure 1, where L is the size of the whole domain. All the graphs in Figure 2 were obtained with periodic Bt plant sowing (cf. equation (18)), at the same values of the growth numbers ($G_{rr} = 0.07$, $G_{rs} = 0.0425$ and $G_{ss} = 0.0425$) and with $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.1$ (Figure 2a), $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$ (Figure 2b) and $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 7.0$ (Figure 2c). Notice that stable nodes are characteristic of the intrinsic plant and insect dynamics at $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.1$; whereas under $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$ and $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 7.0$, the intrinsic dynamics is oscillatory.

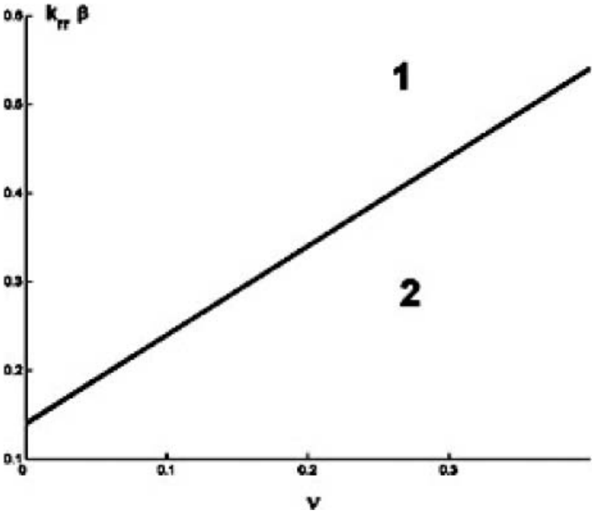


FIGURE 1. The function $k_{rr} \quad = 2 \qquad ; \qquad = 0.07$ Below the graph (region 2) one has a limit cycle, while stable nodes are located above the graph (region 1) at $\omega_r^0 = 0$; if $\omega_r^0 = 0$, limit cycles, stable focuses, and stable nodes appear in region 1. For example, under $k_{rr} \quad = 0.41$, stable nodes are at $\omega_r^0 \quad 0.45$, limit cycles are at $0.45 \quad \omega_r^0 \quad 3.5$, and stable focuses are at $\omega_r^0 \quad 3.5$. The simulations were carried out under the following set of the model parameters: $\alpha = 0.1; \quad = 4.0598; \quad _{rr} = \quad _{rs} = \quad _{ss} = 0.2; \quad _{rr} = \quad _{rs} = \quad _{ss} = 1; \quad = 100$. The parameter values correspond to observational and experimental data (see [29] and the references therein). The only exception is the values of the parameters $_{rr}$, $_{rs}$ and $_{ss}$, which in nature are less than 1; the parameter values $_{rr} = \quad _{rs} = \quad _{ss} = 1$ imply the strongest impact of pests on the plant biomass. Notice that for smaller values of these parameters the results remain ualitatively the same, however, the length of the transition period, when Bt-susceptible and Bt-resistant insects coexist, increases.

There is a distinct difference between the graphs $\langle p \rangle \, (\epsilon A)$ obtained for each of the regions 1 and 2 shown in Figure 1, characteristic of intermediate parameter values ω_s^0, ω_{rs}^0 and ω_r^0, ($\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$ in Figure 2b), rather than under small ($\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.1$ in Figure 2a) or large ($\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 7.0$ in Figure 2c) parameter values. To gain greater insight into why such a difference between the graphs occurs (characteristic of region 1 and region 2), let us consider the phase trajectories shown in Figure 3, which describe the dynamics of the spatially averaged Bt plant and Bt-resistant insect biomasses

$$p = \frac{1}{L} \int_0^L p(x,t)dx, \tag{23}$$

$$i_{rr} = \frac{1}{L} \int_0^L i_{rr}(x,t)dx, \tag{24}$$

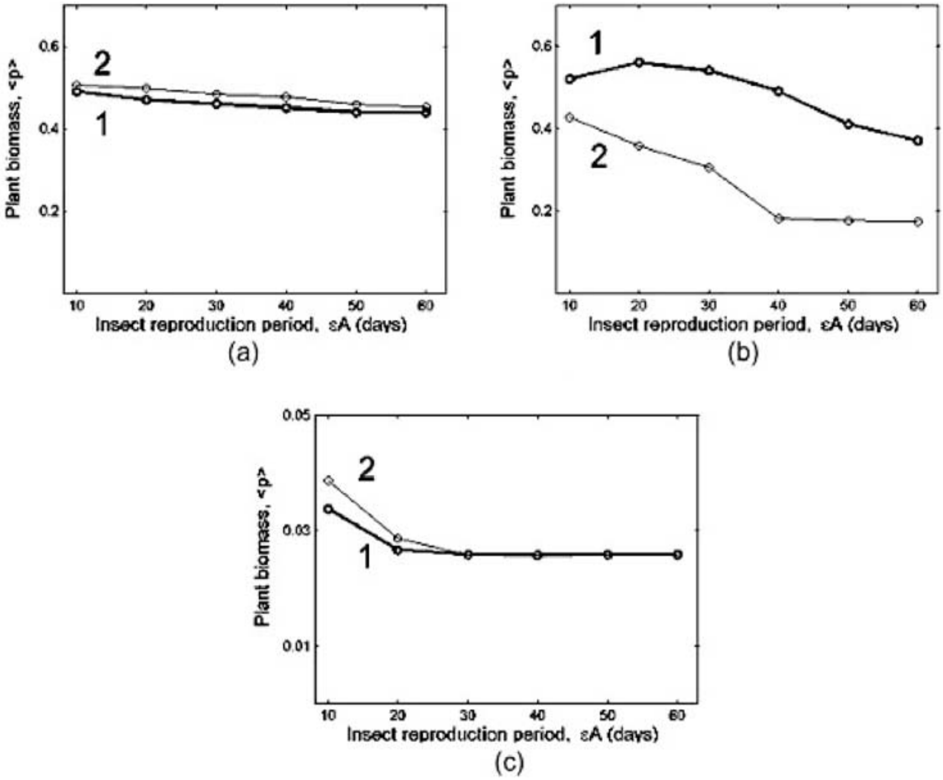


FIGURE 2. The dependencies of the plant biomass $\langle p \rangle$ on the length of the insect reproduction period at (a) $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.1$; (b) $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$;(c) $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 7.0$ $p_0 = 3 \times 10^{-1}$; the start of the growing season coincides with the beginning of the reproduction period. Here and hereafter we set $A = 150$, which roughly corresponds to the length of a normal growing season [24]. Graphs 1 represent the dependencies for a set of the model parameters from region 1 of the model parameter space shown in Figure 1: $k_{rr} = 0.1$; $k_{rs} = k_{ss} = 0.0898$; $\gamma = 0.5036$; other parameters are the same as in Figure 1. Graphs 2 represent the dependencies for a set of the model parameters from region 2 of the model parameter space shown in Figure 1: $k_{rr} = 0.07$; $k_{rs} = k_{ss} = 0.0629$; $\gamma = 0.0525$; other parameters are the same as in Figure 1.

at $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$. It is evident from Figure 3 that at $\epsilon A = 10$ (i.e., when the length of the insect reproduction period is equal to 10 days) the trajectory represents regular oscillations of the plant and insect biomass in region 1 of Figure 1, while in region 2 the trajectory is irregular. The irregular character of this trajectory implies irregularity in changes of local biomass values $p(x,t)$ and $i_{rr}(x,t)$ — cf. equations (23) and (24). In turn, such a local instability arises from fickle local

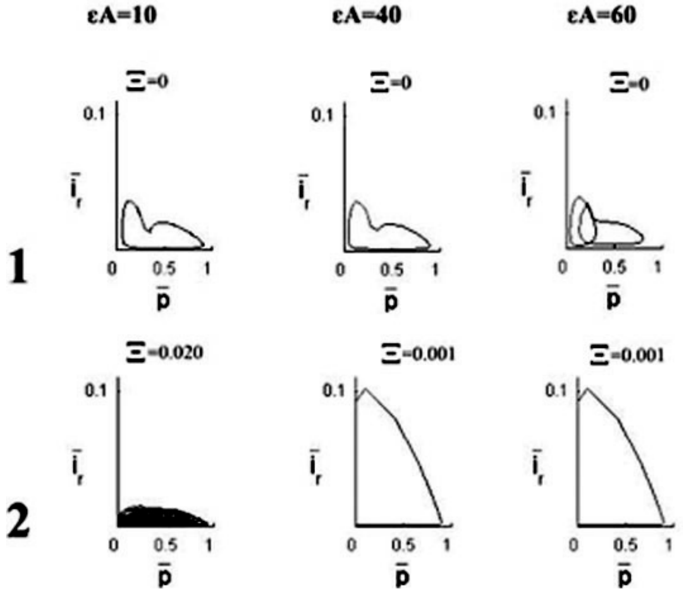


FIGURE 3. Phase portraits of the Bt plant – Bt-resistant insect dynamics in regions 1 and 2 of the parameter space shown in Figure 1 at $\epsilon A = 10; 40; 60$; $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$. The phase portraits are obtained for different values of the duration of the insect reproduction period, ϵA. The values of the diffusion number Ξ, which correspond to each of the portraits, are shown above the phase trajectories. The parameter values are the same as in Figure 2. The calculations were carried out after completion of the transition processes, when Bt-resistant and Bt-susceptible insects still coexisted.

insect diffusion fluxes, due to temporal variations of the inhomogeneous spatial plant and insect biomass distributions (not shown).

To characterise the total effect of the diffusion fluxes, we introduce another parameter, a diffusion number:

$$\Xi = \frac{1}{nA} \int_0^{nA} \int_0^L \left| \frac{\partial^2 i_{rr}(x, t^{‘})}{\partial x^2} \right| dx dt^{‘}. \tag{25}$$

Appropriate values of Ξ are displayed in Figure 3. One can see that $\Xi = 0$ for the regular plant-insect dynamics in region 1 of Figure 1 at $\epsilon A = 10$. This is due to the fact that in this region the transient spatial plant and insect patterns that form in the early stage of the invasion of the Bt resistant insects are eroded through the diffusion of the insects. Both the plant and insect biomasses are therefore finally distributed homogeneously in space (not shown). In contrast, in region 2 with $\Xi \neq 0$ (in Figure 3 at $\epsilon A = 10$), the Bt plant and Bt-resistant insect patterns do not erode and vary in time (not shown). Variations of the patterns result from the joint impact of flattening action of insect diffusion on the one hand, and growth of the insect ant plant biomass on the other.

As Figure 3 suggests, in region 2 at $\epsilon A = 10$ the maximum values of the plant biomass are distributed in a wide range between 0 and 1 along the horizontal axis. Thus on average, the plant biomass in region 2 takes smaller values than in region 1, where the maximum plant biomass is close to 1.

As the duration of the insect reproduction period ϵA increases, the plant-insect biomass oscillations arising in region 2 of the parameter space become regular (cf. Figure 2 at $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.8$), and the values of Ξ get very close to zero (cf. Figure 3). The maximum values of the plant biomass in region 2 become even larger than the maximum values of the plant biomass oscillations in region 1. However, the averaged plant biomass $\langle p \rangle$ given by equation (22) is still smaller in region 2 than in region 1 (cf. Figure 2). The reason is that in region 2 the minimum plant biomass value is less than the minimum plant biomass value in region 1. Besides, in region 2 the plant biomass takes its minimum values for a much longer time than in region 1 (cf. Figure 3), resulting in the total decrease of the $\langle p \rangle$ value (cf. Figure 2).

It is notable that the difference in the character of the plant-insect dynamics does not necessarily lead to essentially different dependencies of the averaged plant biomass $\langle p \rangle$ on the duration of the insect reproduction period ϵA. Indeed, for example at $\omega_s^0 = \omega_{rs}^0 = \omega_r^0 = 0.1$, the graphs $\langle p \rangle (\epsilon A)$ in region 1 and region 2 are very close to each other (cf. Figure 2), even though the underlying plant-insect dynamics are different (not shown). Thus in region 1 the oscillations of the Bt plant and Bt-resistant insect biomass are regular, while in region 2 the oscillations are very irregular. The corresponding values of the diffusion number are zero in region 1 and around 0.02 in region 2, at all values of the duration of the insect reproduction period ϵA.

4. Concluding remarks

We have developed a conceptual reaction-diffusion model of the Bt crop – Bt-susceptible ss- and rs-insects – Bt-resistant rr-insects system, in order to simulate massive invasion of Bt resistant insects. We have shown that the growth numbers G_{rr}, G_{rs} and G_{ss} determine whether or not Bt-resistant insects can invade a field sown with a Bt crop. The results obtained imply that thorough measurements could allow us to estimate the liability of invasion for a given transgenic crop sown in a given geographical region — viz. of the constants β and γ, which parametrise the saturating functional responses (equations 10 – 13); the parameters k_{ss}, k_{rs} and k_{rr}, which are yield coefficients of the susceptible ss- and rs-insects respectively, and resistant rr-insects to the plants (equations 11 – 13); and the insect mortality rates ν_{ss}, ν_{rs} and ν_{rr} (equations 11 – 13) to calculate the growth numbers G_{rr}, G_{rs} and G_{ss}. Such measurements would be of particular interest to check the validity of the growth numbers in agricultural practice, and for invasion risk assessments. To our knowledge, such measurements have not yet been carried out.

Since our model for the plant and insect dynamics depends on uncertain input parameters, the model output should not be interpreted as predictive of the plant-insect system behaviour, but must be interpreted within the context of possible scenarios of population dynamics. Within our conceptual approach, we investigated the dynamics of the model agricultural system in parameter space, which we have shown consists of two regions in Figure 1 with different intrinsic dynamics [16].

We have shown that values of the constants ω_s^0, ω_{rs}^0 and ω_r^0, which parametrise the effective fecundity rates of ss-, rs-, and rr-insects respectively as represented in (15) – (17), essentially affect the character of the dependence of the averaged plant biomass $\langle p \rangle$ in equation (22) on the length of the insect reproduction period, in both region 1 and region 2 of the parameter space (cf. Figure 2).

It has been shown that in the general case the dependence of fecundity rate can be represented by the function

$$F(i) = \frac{B_1 i^2}{B_2 + i}, \tag{26}$$

where i is the population density and the function B_2 represents the population density such that half of the females can be fertilised [3, 4]. If the population density is high (i.e., $i >> B_2$), then $F \sim i$ and in this case the spatio-temporal dynamics can be described with the use of our model (10) – (18). If the population density is low (i.e., $i << B_2$), then $F \sim i^2$. Proper plant-insect dynamics have been analysed in [18, 17, 16], and it is of interest to compare the results obtained from studies of both limiting cases.

(1) In [16] where $F \sim i^2$, we have found that the plant-insect dynamics resulting from an invasion of Bt-resistant insects can be non-unique throughout the insect reproduction period. In particular, a chaotic attractor and the limit cycle have been shown to coexist — so that the Bt-plant — Bt-resistant insect system manifests either chaotic or regular oscillations of plant and insect biomass, depending upon the spatial distributions of the plant and insect biomass. Our analysis of the massive invasion of Bt-resistant insects, carried out with the model (10) – (18) where $F \sim i$, showed that the intrinsic dynamics of Bt plant and Bt-resistant insect biomass can be either an equilibrium or regular oscillations with no tinge of chaoticity (the dominant Lyapunov exponent was not positive). Hence chaotic dynamics do not emerge if the density of Bt-resistant insects is high, resulting in the dependence $F \sim i$.

(2) For $F \sim i^2$, the distinction between the intrinsic dynamics in regions 1 and 2 of parameter space in Figure 1 has been shown to lead to distinctions between the dependencies of the averaged plant biomass on the duration of the insect reproduction period [15]. In contrast, for $F \sim i$ as in model (10) – (18), these dependencies can be very similar to each other — viz. when values of the constants ω_s^0, ω_{rs}^0 and ω_r^0, which parametrise the effective fecundity rates of ss-, rs-, and rr-insects respectively in equations (15) – (17), are beyond some intervening interval (cf. Figure 2).

(3) For both $F \sim i^2$ and $F \sim i$, the regular or irregular character of plant-insect oscillations under periodic Bt plant sowing essentially depends upon local insect fluxes resulting from inhomogeneous spatial insect distributions under Bt-resistant pest invasion. To characterise the insect diffusion fluxes, we introduced a parameter, the diffusion number Ξ — cf. equation (25). For $F \sim i$, the plant and insect oscillations are irregular as $\Xi \geq 0.02$, independently of the system location in parameter space. For $F \sim i^2$, the interrelation between the character of the plant-insect oscillations and the value of the diffusion number Ξ has been shown to be more complex than for $F \sim i$, and depends on the model system position in the parameter space [15].

The results obtained imply that the insect fecundity rate can essentially impact the regular or irregular character of plant-insect dynamics and Bt plant biomass, under an invasion of Bt-resistant pests.

References

[1] Alstad, D. N., Andow, D. A. (1995) Managing the evolution of insect resistance to transgenic plants. *Science* **268**, 1894-1896.

[2] Armstrong, C. L., Parker, G. B., Pershing, J. C., Brown, S. M., Sanders, P. R., Duncan, D. R., Stone, T., Dean, D. A., DeBoer, D. L., Hart, J., Howe, A. R., Morrish, F. M., Pajeau, M. E., Petersen, W. L., Reich, B. J., Rodriguez, R., Santino, C. G., Sato, S. I., Schuler, W., Sims, S. R., Stehling, S., Tarochione, L. J., Fromm, M. E. (1995) Field evaluation of European corn borer control in progeny of 173 transgenic corn evens expressing an insecticidal protein from *Bacillus thuringiensis*. *Crop Sci.* **35**, 550-557.

[3] Bazykin, A. D. (1969) A model of the dynamics of a species size, and the problem of coexistence of closely related species. *Journal of General Biology* **30**, 259-264.

[4] Bazykin, A. D. (2004) *Nonlinear Dynamics of Interacting Populations.* Moscow, Izhevsk: ICS.

[5] Frutos, R., Rang, C., Royer, M. (1999) Managing resistance to plants producing *Bacillus thuringiensis* toxins. *Crit. Rev. Biotechnol.* **19**, 227-276.

[6] Groot, A. T., Dicke, M. (2002) Insect-resistant transgenic plants in a multi-trophic context. *Plant J.* **31**, 387-406.

[7] Gould, F. (1998) Sustainability of transgenic insecticidal cultivars: integrating pest genetics and ecology. *Annu. Rev. Entomol.* **43**, 701-726.

[8] Hillier, J. G., Birch, A. N. E. (2002) A Bt-trophic mathematical model for pest adaptation to a resistant crop. *J. Theor. Biol.* **215**, 305-319.

[9] Janmaat, A. F., Meyers, J. (2003) Rapid evolution and the cost of resistance to *Bacillus thuringiensis* in greenhouse populations of cabbage loopers, *Trichoplusia ni*. *Proc. R. Soc. Lond. B Biol. Sci.* **270**, 2263-2270.

[10] Kolmogorov, A., Petrovskii, I., Piskunov, S. (1937) Étude de l'equation de la diffusion avec croissance de la quantité de mattière et son application à un problème biologique. *Bull. Univ. Moscou, Serie Int. (Section A)* **1**, 1-25.

[11] Kot, M. (2001) *Elements of Mathematical Ecology.* Cambridge: Cambridge University.

[12] Lewis, W. J., van Lenteren, J. C., Phatak, S. C., Tumlinson, J. H. (1998) A total system approach to sustainable pest management. *Proc. Natl Acad. Sci. USA* **94**, 12243-12248.

[13] Lotka, A. J. (1925) *Elements of Physical Biology.* Baltimore: Williams and Wilkins.

[14] Malchow, H., Petrovskii, S. V., Medvinsky, A. B.(2001) Pattern formation in models of plankton dynamics. *Oceanologica Acta* **24**, 479-487.

[15] Medvinsky, A. B., Gonik, M. M., Li, B.-L., Malchow, H. (2007) Beyond Bt resistance of pests in the context of population dynamics complexity. *Ecological Complexity* **4**, 201–211.

[16] Medvinsky, A. B., Gonik, M. M., Li, B.-L., Velkov, V. V., Malchow, H. (2006) Invasion of pests resistant to Bt toxin can lead to inherent non-uniqueness in genetically modified Bt-plant dynamics: mathematical modeling. *J. Theor. Biol.* **242**, 539-546.

[17] Medvinsky, A. B., Gonik, M. M., Velkov, V. V., Li, B.-L., Malchow, H. (2005) Modeling invasion of pests resistant to Bt toxins produced by genetically modified plants: recessive vs. dominant invaders. *Natural Resource Modeling* **18**, 347-362.

[18] Medvinsky, A. B., Morozov, A. Y., Velkov, V. V., Li, B.-L., Sokolov, M. S., Malchow, H. (2004) Modeling the invasion of recessive Bt-resistant insects: an impact on transgenic plants. *J. Theor. Biol.* **231**, 121-127.

[19] Medvinsky, A. B., Petrovskii, S. V., Tikhonova, I. A., Malchow, H., Li, B.-L. (2002) Spatiotemporal complexity of plankton and fish dynamics. *SIAM Review* **44**, 311-370.

[20] Medvinsky, A. B., Tikhonova, I. A., Petrovskii, S. V., Malchow, H., Venturino, E. (2001) Chaos and order in spatially structured plankton dynamics. A theoretical study. In: *Nonlinear Dynamics in the Life and Social Sciences* (ed. by W. Sulis and I. Trofimova). Amsterdam, Berlin, Oxford, Tokyo, Washington: IOS.

[21] Petrovskii, S. V., Li, B.-L. (2006) *Exactly Solvable Models of Biological Invasion.* Boca Raton: Chapman & Hall/CRC.

[22] Rajamohan, F., Lee, M. K., Dean, D. H. (1998) *Bacillus thuringiensis* insecticidal protein: molecular mode of action. *Prog. Nucl. Res. Mol. Biol.* **60**, 1-23.

[23] Scott, S.E., Wilkinson, M.J. (1998) Transgene risk is low. *Nature* **393**, 320.

[24] Storer, N. P., Peck, S. L., Gould, F., van Duyn, J. W., Kennedy, G. G. (2003) Spatial processes in the evolution of resistance in *Helicoverpa zea* (*Lepidoptera: Noctiudae*) to Bt transgenic corn and cotton in a mixed agroecosystem: a biology-rich stochastic simulation model. *J. Econ. Entomol.* **96**, 156-172.

[25] Tabashnik, B. E. (1994) Evolution of resistance to *Bacillus thuringiensis. Annu. Rev. Entomol.* **39**, 47-49.

[26] Tabashnik, B. E., Carriere, Y., Dennehy, T. J., Morin, S., Sisterson, M. S., Roush, R. T., Shelton, A. M., Zhao, J.-Z. (2003) Insect resistance to transgenic Bt crops: Lessons from the laboratory and the field. *J. Econ. Entomol.* **96**, 1031-1038.

[27] Tabashnik, B. E., Cushing, N. L., Finson, N., Johnson, M. W. (1990) Field development of resistance to *Bacillus thuringiensis* in diamondback moth (*Lepidoptera: Plutellidae*). *J. Econ. Entomol.***83**, 1671-1676.

[28] Tabashnik, B. E., Patin, A. L., Dennehy, T. J., Liu, Y. B., Carriere, Y., Sims, M. A., Antilla, L. (2000) Frequency of resistance to *Bacillus thuringiensis* in field populations of pink bollworm. *Proc. Natl Acad. Sci. USA* **97**, 12980-12984.

[29] Velkov, V. V., Medvinsky, A. B., Sokolov, M. S., Marchenko, A. I. (2005) Will transgenic plants adversely affect environment? *J. Biosciences***30**, 515-548.

[30] Volterra, V. (1926) Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558-560.

Alexander B. Medvinsky
Institute for Theoretical & Experimental Biophysics
Pushchino
Moscow Region
142290 Russia
e-mail: `medvinsky@iteb.ru`

Maria M.Gonik
Institute for Theoretical & Experimental Biophysics
Pushchino
Moscow Region
42290 Russia
e-mail: `mgmaria@yandex.ru`

Yuri V. Tyutyunov
Vorovich Research Institute of Mechanics and Applied Mathematics
South Federal University
Rostov on Don
344090 Russia
e-mail: `tyutyunov@rsu.ru`

Bai-Lian Li
Department of Botany and Plant Sciences
University of California
Riverside
CA 92521-0124, USA
e-mail: `bai-lian.li@ucr.edu`

Alexey V. Rusakov
Institute for Theoretical & Experimental Biophysics
Pushchino
Moscow Region
142290 Russia
e-mail: `rusakov_a@rambler.ru`

Horst Malchow
Institute of Environmental Systems Research
University of Osnabrueck
D-49069 Osnabrueck
Germany
e-mail: `malcow@uos.de`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 95–107
© 2008 Birkhäuser Verlag Basel/Switzerland

Reducing the Emission of Pollutants in Industrial Wastewater through the Use of Membrane Bioreactors

Mark I. Nelson, X. Dong Chen and Harvinder S. Sidhu

Abstract. Many industrial processes produce wastewater containing pollutants, the concentration of which must be reduced before the wastewater can be discharged. One way to do this is through the use of a biological species ('biomass') that consumes the pollutant ('substrate'). In a membrane bioreactor the biomass is constrained to remain within the reactor whereas the feed stream flows through it.

We investigate the behaviour of a reaction governed by Contois growth kinetics in both single and double membrane reactor configurations. The optimal performance of both configurations is determined and compared. It is found that in many cases the cascade reactor may outperform the single reactor by two orders of magnitude.

Mathematics Subject Classification (2000). Primary 92C45; Secondary 92E20.

Keywords. Bioreactors, Membrane bioreactors, Reaction engineering, Wastewater reclamation.

1. Introduction

Many industrial processes produce wastewater containing high levels of pollutants. Before the wastewater can be discharged the pollutant concentration has to be decreased. One way to achieve this is to pass the wastewater through a reactor containing biomass, which grows through consumption of the pollutant. In a membrane bioreactor, a membrane filtration process is used to to separate the effluent from the biomass. This process retains biomass within the bioreactor, increasing its

Results in this paper were presented at the International Conference on Mathematical Modelling and Computation held at the University of Brunei Darussalam during 5–8 June 2006, in conjunction with the 20th anniversary celebration of the foundation of the university.

concentration and allowing for a more efficient treatment of contaminated wastewater. This produces a higher quality effluent than is obtained using conventional reactors. Consequently, membrane reactors are increasingly being used as elements of advanced water processing schemes.

We consider a simple model for a membrane bioreactor in which the treatment process is modelled as a continuously stirred tank reactor (CSTR). However, unlike a CSTR the biomass is constrained to remain in the bioreactor whilst the pollutant ('substrate') flows through it. The biomass growth kinetics are modelled using the Contois expression [1]. Contois growth kinetics have previously been used to model the aerobic degradation of wastewater originating in the industrial treatment of black olives [2], the anaerobic treatment of dairy manure [3], the anaerobic digestion of ice-cream wastewater [4], the anaerobic treatment of textile wastewater [5] and the aerobic biodegradation of solid municipal organic waste [6]. In these papers the use of a Contois expression was validated by a comparison of model predictions with experimental data. In [4, 5] Contois growth kinetics were shown to give a superior fit to experimental data than other growth rate expressions. Contois growth kinetics have also been used to simulate the cleaning of wastewater by microorganisms [7].

We investigate the performance of a single membrane reactor by determining the conditions for the biomass to die-out and the conditions for self-sustained oscillations to be generated within the reactor. We then investigate the performance of a cascade of two reactors and compare its performance against that of a single reactor.

2. Model equations

We investigate a microbial system in which cell mass ($X_i, i = 1, 2$) grows through consumption of a substrate species (S_i). The specific growth rate, equation (5), is given by a Contois expression with variable yield coefficient, equation (6). The objective is to minimise the substrate concentration leaving the reactor (S_2). Although our model is simplistic, it is still worth analysing since it is a standard bioreactor engineering model. The dimensional and dimensionless forms of our model are stated in § 2.1 and § 2.2 respectively.

2.1. Dimensional model

The governing equations for the Contois kinetic system in a cascade of two reactors arise from a simple mass balance on substrate and biomass and are given by

Reactor 1

$$V_1 \frac{\mathrm{d}S_1}{\mathrm{d}t} = F\left(S_0 - S_1\right) - V_1 X_1 \frac{\mu\left(S_1, X_1\right)}{Y\left(S_1\right)}, \tag{1}$$

$$V_1 \frac{\mathrm{d}X_1}{\mathrm{d}t} = V_1 X_1 \mu\left(S_1, X_1\right) - V_1 d X_1, \tag{2}$$

Reactor 2

$$V_2 \frac{\mathrm{d}S_2}{\mathrm{d}t} = F\left(S_1 - S_2\right) - V_2 X_2 \frac{\mu\left(S_2, X_2\right)}{Y\left(S_2\right)}, \tag{3}$$

$$V_2 \frac{\mathrm{d}X_2}{\mathrm{d}t} = V_2 X_2 \mu\left(S_2, X_2\right) - V_2 d X_2. \tag{4}$$

Specific growth rate

$$\mu\left(S_i, X_i\right) = \frac{\mu_m S_i}{K_s X_i + S_i}. \tag{5}$$

Yield Coefficient

$$Y\left(S_i\right) = \alpha + \beta S_i, \quad (\alpha, \beta > 0). \tag{6}$$

Residence times

$$\tau_i = \frac{V_i}{F}, \tag{7}$$

$$\tau_1 + \tau_2 = \tau_t. \tag{8}$$

This is the simplest model for a membrane reactor. The terms that appear in equations (1)–(6) are defined in the nomenclature. The linear dependence of yield coefficient on substrate concentration ($\beta \neq 0$) was proposed by Essajee and Tanner [8].

2.2. Dimensionless equations

By introducing dimensionless variables for the substrate concentrations ($S^* = \beta K_s S$), the cell mass concentrations ($X^* = \beta K_s^2 X$) and time ($t^* = \mu_m t$) the system of differential equations (1–4) can be written in the dimensionless form

Reactor 1

$$\frac{\mathrm{d}S_1^*}{\mathrm{d}t^*} = \frac{1}{\tau_1^*}\left(S_0^* - S_1^*\right) - \frac{S_1^* X_1^*}{\left(S_1^* + X_1^*\right)\left(\alpha^* + S_1^*\right)}, \tag{9}$$

$$\frac{\mathrm{d}X_1^*}{\mathrm{d}t^*} = \frac{S_1^* X_1^*}{S_1^* + X_1^*} - d^* X_1^*. \tag{10}$$

Reactor 2

$$\frac{\mathrm{d}S_2^*}{\mathrm{d}t^*} = \frac{1}{\tau_2^*}\left(S_1^* - S_2^*\right) - \frac{S_2^* X_2^*}{\left(S_2^* + X_2^*\right)\left(\alpha^* + S_2^*\right)}, \tag{11}$$

$$\frac{\mathrm{d}X_2^*}{\mathrm{d}t^*} = \frac{S_2^* X_2^*}{S_2^* + X_2^*} - d^* X_2^*. \tag{12}$$

Dimensionless residence time relationship

$$\tau_1^* + \tau_2^* = \tau_t^*. \tag{13}$$

This form contains four parameters $S_0^*, \alpha^*, d^*, \tau^*$ and we take the residence time (τ^*) as the primary bifurcation parameter. The substrate concentration in the feed (S_0^*), the dimensionless yield constant (α^*) and the dimensionless death rate (d^*)

are the secondary bifurcation parameters. The values for α^* and d^* are determined by the choice of microbial system and are therefore not 'tunable' parameters.

A feature of our dimensionless scheme is that there is a one-to-one relationship between the dimensionless variables and their dimensional counterparts. Hence we often write, for example, 'the residence time', rather than 'the dimensionless residence time'. We refer to S_0^* simply as the feed concentration.

2.3. Summary of results for planar systems

Here we summarise some useful results [9] regarding the behaviour of planar systems of the form

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f\left(x, y\right), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= g\left(x, y\right). \end{aligned}$$

Much of the behaviour is determined by properties of the Jacobian matrix (J) evaluated at any steady-state (x_0, y_0);

$$J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix} \tag{14}$$

where

$$\begin{aligned} J_{11} &= f_x, \quad J_{12} = f_y, \\ J_{21} &= g_x, \quad J_{22} = g_y. \end{aligned}$$

The Jacobian matrix (14) has a zero eigenvalue when

$$J_{11}J_{22} - J_{12}J_{21} = 0.$$

The conditions for a double-zero eigenvalue are

$$J_{11}J_{22} - J_{12}J_{21} = 0 \quad \text{and} \quad H = J_{11} + J_{22} = 0.$$

The conditions for a Hopf bifurcation are

$$J_{11}J_{22} - J_{12}J_{21} > 0 \quad \text{and} \quad H = J_{11} + J_{22} = 0. \tag{15}$$

A degenerate Hopf bifurcation, at which two Hopf points annihilate each other in an unfolding diagram ($\mathrm{H2}_1$ degeneracy), occurs when the following conditions are satisfied:

$$\begin{aligned} H &= 0, \\ \frac{\mathrm{d}H}{\mathrm{d}\tau^*} &= 0. \end{aligned} \tag{16}$$

It is possible for isolated branches of periodic solutions to be formed that are not associated with a Hopf bifurcation [9], through particular degenerate Hopf bifurcations — but this possibility has not been investigated in this study.

2.4. Numerics

Steady-state diagrams were obtained using the path-following software Auto 97 [10]. In these diagrams, the standard representation is used: solid and dotted lines represent stable and unstable steady states respectively; squares are Hopf bifurcation points; and open and filled-in circles represent unstable and stable periodic orbits respectively. For a periodic orbit, the norm used is the integral over the period of the solution. We investigate the effluent concentration leaving the reactor (S^*) as a function of the residence time (τ^*).

3. Results for a single reactor

3.1. Analytical results

3.1.1. Steady-state solutions. The equations (9) and (10) have physically meaningful steady-state solutions given by

$$(S^*, X^*) = (S_0^*, 0)\,, \tag{17}$$

$$(S^*, X^*) = \left(\frac{d^*\hat{X}}{1-d^*}, \hat{X}\right), \tag{18}$$

where $\hat{X}$ is given by

$$\begin{aligned} \hat{X} &= \frac{1-d^*}{2d^*}\left[-\left(1-d^*\right)\tau^* - \left(\alpha^* - S_0^*\right) + \sqrt{a}\right], \\ a &= \left(1-d^*\right)^2 {\tau^*}^2 + 2\left(1-d^*\right)\left(\alpha^* - S_0^*\right)\tau^* + \left(\alpha^* + S_0^*\right)^2 . \end{aligned} \tag{19}$$

We refer to the steady-state solution defined by equation (17) as the ‘death solution’, because all the biomass has died. We refer to the steady-state solution of equation (18) as the ‘no-death solution’, and note that $d^* < 1$ is required for the substrate component of the no-death solution to be physically meaningful ($S^* > 0$). It can be shown that the cell-mass component of this steady-state solution ($\hat{X}$) is strictly positive.

Calculation shows that

$$\frac{\mathrm{d}\hat{X}}{\mathrm{d}\tau^*} < 0.$$

Consequently for the no-death state

$$\frac{\mathrm{d}S^*}{\mathrm{d}\tau^*} < 0,$$

so the substrate concentration is a decreasing function with respect to the residence time (τ^*) and therefore is minimised at infinite residence time.

3.1.2. Stability analysis of the 'death' solution. From an operational viewpoint, the 'death solution' must be avoided since this would result in no biomass present in the reactor to consume the substrate (pollutant). The eigenvalues of the Jacobian matrix evaluated at the 'death solution' are found to be

$$\begin{aligned}\lambda_1 &= -\frac{1}{\tau^*},\\ \lambda_2 &= 1 - d^*.\end{aligned}$$

Thus the death state is stable if

$$d^* > 1,$$

which occurs when the cell mass death-rate is greater than the maximum specific growth rate. Note that the stability of the death state is independent of the operating parameters (τ_1^*, S_0^*). Henceforth we assume that the condition $d^* < 1$ holds.

3.1.3. Hopf bifurcation on the 'no-death' state. Next we find the condition for Hopf bifurcations to occur on the 'no-death' state. Note that this steady-state is only physically meaningful for $d^* < 1$. After some algebra we obtain

$$J_{11}J_{22} - J_{12}J_{21} = \frac{d^*\left(1-d^*\right)}{\tau^*} \cdot \left\{ 1 + \frac{\alpha^*\left(1-d^*\right)^3 \tau^*}{\left[\alpha\left(1-d^*\right) + d^*\hat{X}\right]^2} \right\}.$$

Thus the determinant of the Jacobian matrix is always positive, with the assumption that $d^* < 1$, and consequently a double-zero eigenvalue degeneracy cannot occur.

Thus from § 2.3 the condition for a Hopf bifurcation to occur in the system, after some algebra, is given by

$$H\left(\tau^*\right) = a_0 + a_1\left(1-d^*\right)\tau^* + a_2\left(1-d^*\right)^2 {\tau^*}^2 + a_3\left(1-d^*\right)^3 {\tau^*}^3 = 0, \qquad (20)$$

$$\begin{aligned}a_0 &= \left(\alpha^* + S_0^*\right)^2,\\ a_1 &= 2\left(\alpha^* - S_0^*\right) + \left(\alpha^* + S_0^*\right)^2 d^*,\\ a_2 &= 1 + 2\left(\alpha^* - S_0^*\right)d^* + \left(\alpha^* + S_0^* - 1\right){d^*}^2,\\ a_3 &= d\left[1 + \left(\alpha^* - 1\right)d^*\right].\end{aligned}$$

Observe that because $0 < d < 1$ the coefficient a_3 of the cubic term is strictly positive, and since

$$H\left(0\right) = \left(\alpha + S_0^*\right)^2 > 0$$

there are at most two positive Hopf bifurcation points.

With $\alpha^* = 0.01$, $d^* = 0.3$ and $S_0^* = 1.05$, the parameter values used for Figure 1 (b), equation (20) has zeroes when $\tau^* = 1.335$ and $\tau^* = 1.974$. These are the values of the residence time corresponding to Hopf bifurcation points.

Condition (16) shows that when $\alpha^* = 0.01$ and $S_0^* = 1.05$ an $\mathrm{H2}_1$ degeneracy occurs where

$$(d^*, \tau^*) = (0.2126, 1.396), \tag{21}$$

and we conclude from this analysis that natural oscillations do not occur when $d^* < 0.2126$. Further, from the Hopf condition (20) it is possible to deduce that Hopf bifurcations cannot occur when

$$\alpha^* \geq \frac{d^*}{1-d^*} S_0^*, \tag{22}$$

i.e., when either the feed concentration or the death rate is sufficiently small.

3.2. Numerical results

Figure 1 shows two steady-state diagrams, each exhibiting the 'death' and 'no-death' solutions. In diagram (a) the value for the death-rate (d^*) is lower than the critical value given by equation (21) and there are no Hopf points. In diagram (b) the death rate (d^*) is higher than the critical value given by equation (21) and there are two Hopf points. In Figure 1 (b) the average value of the effluent concentration on a limit cycle is higher than that at the corresponding unstable steady-state — periodic behaviour does not improve the performance of the reactor.

Figure 2 shows the curve defined by equation (16) for a fixed value of the yield constant. If the feed concentration value is above the the $\mathrm{H2}_1$ locus, then the steady-steady diagram contains *two* Hopf points. If it is below the locus, then the steady-state diagram contains *no* Hopf points. This figure shows that the range of feed concentrations over which oscillations are possible increases for higher values of the death rate.

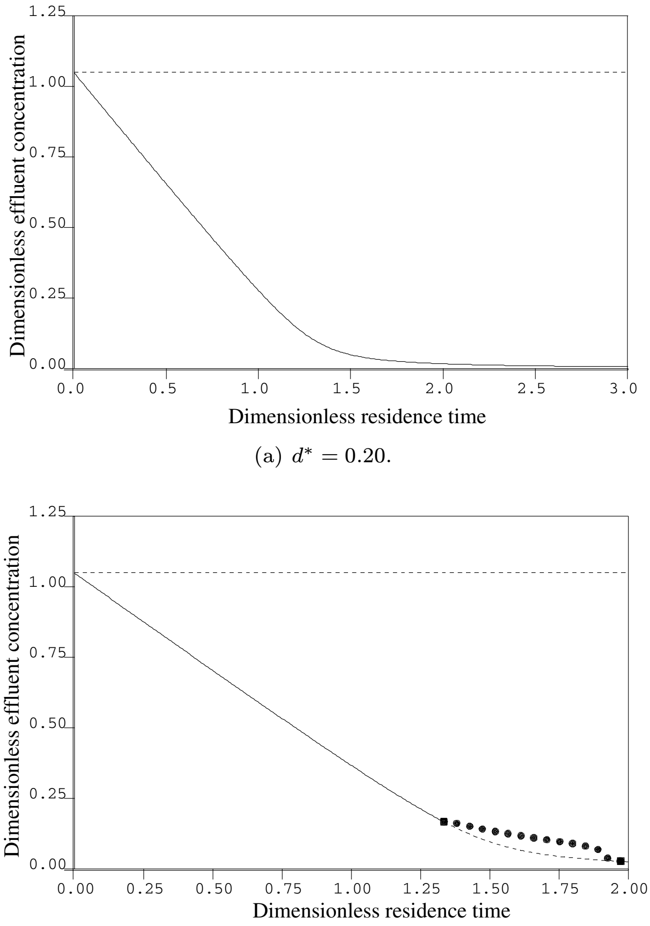
4. Results for a double-reactor

In this section we investigate the performance of a cascade containing two reactors in series. For a given total residence time (τ_t^*) we investigate how the effluent concentration (S_2^*) varies as a function of the residence time in the first reactor (τ_1^*). (The residence time in the second reactor is simply $\tau_t^* - \tau_1^*$). The limits $\tau_1^* = 0$ and $\tau_1^* = \tau_t^*$ represent the degenerate case, in which the 'cascade' reduces back to a single reactor of residence time $\tau_1^* = \tau_t^*$.

4.1. Steady-state Diagrams

Figure 3 shows a sequence of steady-state diagrams as a function of the total residence time (τ_t^*). In each figure the performance of the cascade varies considerably as the reactor-design is varied through the choice of the residence time in the first reactor.

In Figure 3 (a) the effluent concentration has a global minimum of $S_2^* = 4.19 \times 10^{-5}$ when the residence time in the first reactor is $\tau_1^* = 4.52$. The effluent profile is very flat for residence times near $\tau_1^* = 4.52$. Consequently, there is only a very small degradation in reactor performance if the residence time of the reactor is



(a) $d^* = 0.20$.

(b) $d^* = 0.30$.

FIGURE 1. Steady-state diagrams showing the variation of substrate effluent (S^*) with residence time (τ^*). Parameter values: feed concentration, $S_0^* = 1.05$; yield constant, $\alpha^* = 0.01$.

not the optimal choice — there is robustness in the optimal reactor performance. In Figure 3 (a) the two Hopf points are to the left of the global minimum. In Figure 3 (b) the Hopf points straddle the global minimum, which is now unstable. We define the best performance of this reactor to be the value associated with the Hopf bifurcation at $\tau_1^* = 1.97$. The robustness of the reactor performance to small variations in the reactor design is not as high as it is in Figure 3 (a). If the residence time in the first reactor is slightly too high, the effluent concentration increases due to periodic behaviour; whereas if it is too low, the effluent concentration is given

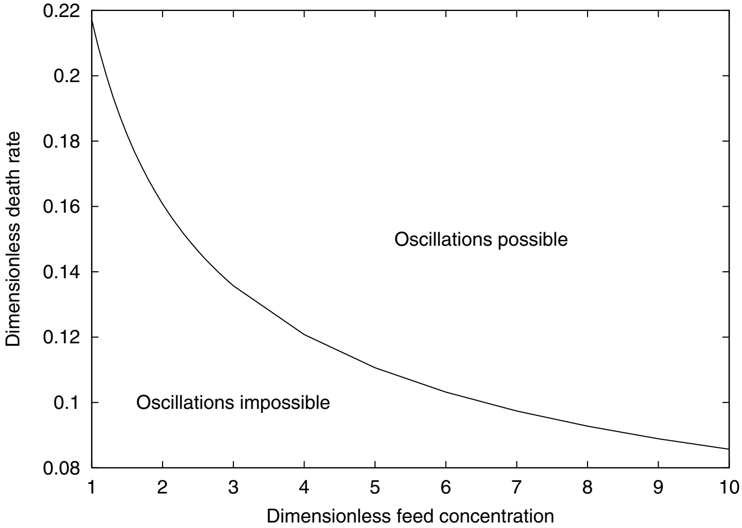
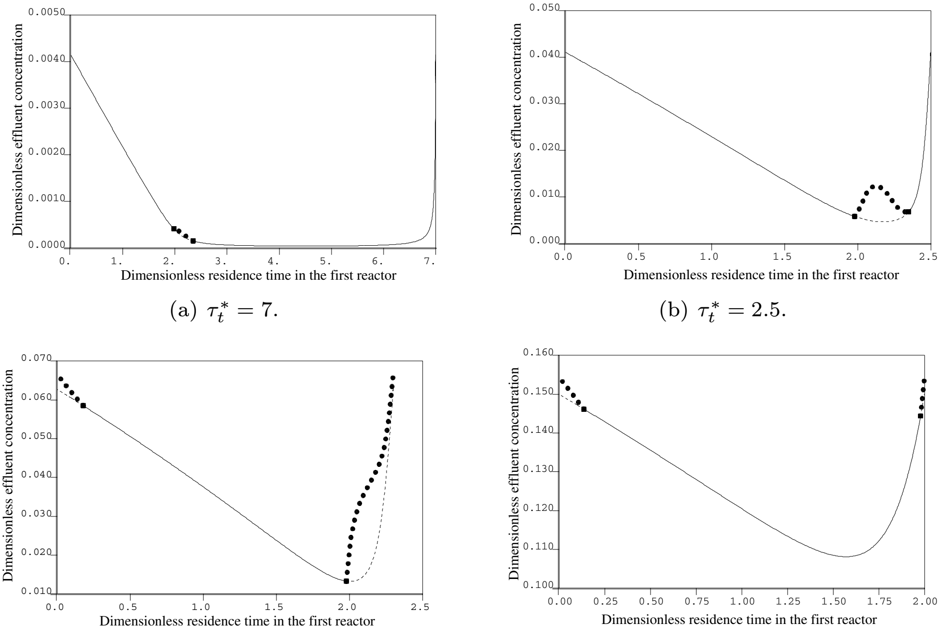


FIGURE 2. $H2_1$ degeneracy diagram. Parameter value: yield constant, $\alpha^* = 0.01$.

by a steady-state with a noticeably higher value. However, in both circumstances the performance of the double reactor is superior to that of a single reactor with a residence time $\tau_1^* = 2.5$.

Observe that in Figures 3 (a) and (b) there is a single periodic solution branch which connects the two Hopf points, whereas in Figures 3 (c) and (d) there are two disjoint periodic solution branches. The transition between these two types of behaviour can be understood by applying the Hopf bifurcation equation (20) to a single reactor with parameter values specified in the caption of Figure 3. This shows that there are Hopf bifurcation points in the single reactor when $\tau^* = 1.97$ and $\tau^* = 2.34$. These are the values of the Hopf bifurcation points exhibited in Figures 3 (a) and (b). Thus these Hopf bifurcation points are generated by the dynamics of the first reactor. In Figures 3 (c) and (d) the total residence time is smaller than 2.34 — and therefore in these figures one Hopf bifurcation point is generated by the dynamics of the first reactor (at $\tau_1^* = 1.97$), whereas the other (the one on the left) is generated by the dynamics of the second reactor.

In Figure 3 (c) the global minimum is unstable, and the same remarks regarding the operation of the reactor apply as for Figure 3 (b). At values of the residence time between the Hopf points the steady-state solutions are now stable, whereas in Figures 3 (a and b) they were unstable. Thus in Figure 3 (d), in which the two Hopf points again straddle the global minimum, the global minimum is stable. Here there is some degree of robustness in the performance of the reactor to errors in its construction, although the robustness is less than for the case shown in Figure 3 (a).



(a) $\tau_t^* = 7$.

(b) $\tau_t^* = 2.5$.

(c) $\tau_t^* = 2.3$.

(d) $\tau_t^* = 2$.

FIGURE 3. Steady-state diagrams for a double membrane reactor cascade showing the variation of substrate effluent (S_2^*) with residence time in the first reactor (τ_1^*). Parameter values: death rate, $d^* = 0.2$, feed concentration, $S_0^* = 1.65$; yield constant, $\alpha^* = 0.01$.

4.2. Comparing the Performance of a Cascade with a Single Reactor

In Figure 4 we compare the performance of the best double reactor design with that of a single reactor. The single reactor performance line is found as the stable component of a steady-state diagram for a single reactor. Thus depending upon the total residence time, the best single reactor performance is either given by a steady-state solution or a periodic solution, — cf. Figure 1. For a given value of the total residence time, in Figure 4 the best performance of the double reactor is obtained from the corresponding steady-state diagram. The best performance is defined to be the global minimum, if that is stable; and if it is unstable, the Hopf bifurcation that gives the lowest effluent concentration. The optimal performance was never found to be associated with a stable periodic solution.

Figure 4 shows that at sufficiently small residence times there is little difference between an optimised cascade and a single reactor. However, at higher residence times there is a considerable difference between their performance: an optimised cascade can be superior by two orders of magnitude. Figure 4 also shows that the performance of a single reactor can sometimes be replicated by a double

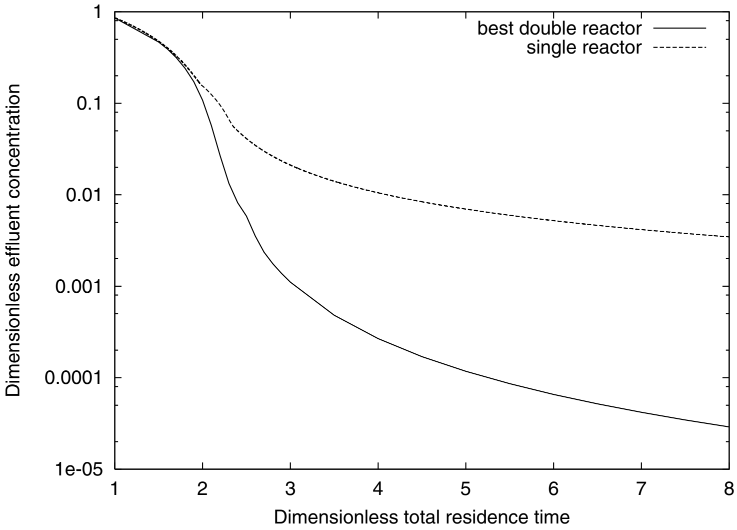


FIGURE 4. Comparison of performance in the best double membrane cascade against the performance in a single reactor. Parameter values: death rate, $d^* = 0.2$, feed concentration, $S_0^* = 1.65$; yield constant, $\alpha^* = 0.01$.

reactor having a much shorter residence time. Under such circumstances, a cascade has a considerably greater throughput of polluted wastewater.

Figures similar to Figure 4 have been obtained for different values of the feed concentration. The values of the total residence time at which the performance of the single and double reactor configurations starts to differ depends upon the feed concentration. However, the ‘selling point’ remains the same — i.e., the optimised cascade outperforms the single reactor for most residence times, two orders of magnitude for ‘moderate’ residence times and even higher for even larger residence times.

5. Conclusion

We have analysed microbial growth in a single membrane reactor, using a Contois growth model with a variable yield coefficient. We also considered a cascade of two membrane reactors, in which the residence times in the two reactors are varied whilst keeping the total residence time fixed. The optimal performance of the cascade, measured by the minimal stable effluent concentration, was determined and compared to that of a single reactor operating at the same residence time. The performance of a cascade could be two orders of magnitude better than that of a single reactor. For both the single and cascade reactor configurations, it was found that periodic behaviour did not improve the performance of the reactor.

Acknowledgement

This work was supported by a grant from the Australian Research Council (DP0559177).

Appendix A. Nomenclature

F	Flowrate.	$(\mathrm{l\,hr}^{-1})$
K_s	Contois saturation constant.	(—)
S	Substrate concentration.	$(\mathrm{g\,l}^{-1})$
S^*	Dimensionless substrate concentration. $S^* = \beta K_s S$.	(—)
S_0	Substrate concentration in the feed.	$(\mathrm{g\,l}^{-1})$
S_0^*	Dimensionless substrate concentration in the feed. $S_0^* = \beta K_s S_0$.	(—)
V	Reactor volume.	(l)
X	Cell mass concentration.	$(\mathrm{g\,l}^{-1})$
X^*	Dimensionless cell mass concentration. $X^* = \beta K_s^2 X$.	(—)
X_2^*	The dimensionless cell mass concentration in the second bioreactor in a cascade.	(—)
$Y\,(S)$	Cell mass yield coefficient.	(—)
d	Death rate of cell mass.	(h^{-1})
d^*	Dimensionless cell mass death rate. $d^* = d/\mu_m$	(—)
t	Time.	(h)
t^*	Dimensionless time. $t^* = \mu_m t$.	(—)
α	Constant in yield coefficient.	(—)
α^*	Dimensionless yield constant. $\alpha^* = \alpha K$.	(—)
β	Constant in yield coefficient.	$(\mathrm{l\,g}^{-1})$
$\mu\,(S)$	Specific growth rate.	(hr^{-1})
μ_m	Maximum specific growth rate.	(hr^{-1})
τ^*	Dimensionless residence time. $\tau^* = \mu_m \cdot V/F$.	(—)

References

[1] D. E. Contois. *Kinetics of bacterial growth: Relationship between population density and specific growth rate of continuous cultures.* J. of General Microbiology **21** (1959), 40–50.

[2] J. Beltran-Heredia, J. Torregrosa, J. R. Dominguez, and J. Garcia. *Ozonation of black-table-olive industrial wastewaters: effect of an aerobic biological pretreatment.* J. of Chemical Technology and Biotechnology **75** (2000), 561–568.

[3] A. E. Ghaly, S. S. Sadaka, and A. Hazza'a. *Kinetics of an intermittent-flow, continuous-mix anaerobic reactor.* Energy Sources **22** (2000), 525–542.

[4] W. C. Hu *et al* K. Thayanithy, and C. F. Forster. *A kinetic study of the anaerobic digestion of ice-cream wastewater.* Process Biochemistry **37** (2002), 965–971.

[5] M. Işik and D. T. Sponza. *Substrate removal kinetics in an upflow anaerobic sludge blanket reactor decolorising simulated textile wastewater.* Process Biochemistry **40** (2005), 1189–1198.

[6] L. Krzystek, S. Ledakowicz, H-J. Kahle, and K. Kaczorek. *Degradation of household biowaste in reactors.* J. of Biotechnology **92** (2001), 103–112.

[7] J. Czeczot, M. Metzger, J. P. Babary, and M. Nihtilä. *Filtering in adaptive control of distributed parameter bioreactors in the presence of noisy measurements.* Simulation Practice and Theory **8** (2000), 39–56.

[8] C. K. Essajee and R. D. Tanner. *The effect of extracelluluar variables on the stability of the continuous baker's yeast-ethanol fermentation process.* Process Biochemistry **14** (1979), 16–25.

[9] B. F. Gray and M. J. Roberts. *A method for the complete qualitative analysis of two coupled ordinary differential equations dependent on three parameters.* Proceedings of the Royal Society A **416** (1988), 361–389.

[10] E. J. Doedel, T. F. Fairgrieve, B. Sandstede, A. R. Champneys, Y. A. Kuznetsov, and X. Wang. *AUTO 97: Continuation and bifurcation software for Ordinary Differential Equations (with HomCont)*, March 1998. Available by anonymous ftp from ftp.cs.concordia.ca/pub/doedel/auto.

Mark I. Nelson
School of Mathematics and Applied Statistics
The University of Wollongong
Wollongong
NSW 2522
Australia
e-mail: `nelsonm@member.ams.org`

X. Dong Chen
Bioproduct and Food Engineering
Department of Chemical Engineering
Monash University
Clayton Campus
VIC 3800
Australia
e-mail: `dong.chen@eng.monash.edu.au`

Harvinder S. Sidhu
School of Physical, Environmental and Mathematical Science
UNSW at ADFA
Canberra
ACT 2600
Australia
e-mail: `h.sidhu@adfa.edu.au`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 109–141

Model Hysteresis Dimer Molecule. I. Equilibrium Properties

Christopher G. Jesudason

Abstract. A Hamiltonian system describing hysteresis behavior in a dimeric chemical reaction is modeled in a MD simulation utilizing novel two-body potentials with switches, a technique that is particularly suitable for numerical thermodynamical investigations. It is surmised that such reaction mechanisms could exist in nature on the basis of recent experiments which indicate that electromagnetic hysteresis behavior is exhibited at the molecular level, although experimental interpretations tend to construct models that avoid such mechanisms. Numerical results of various common equilibrium thermodynamical and kinetic properties are presented together with new algorithms that were implemented to compute these quantities. No unusual thermodynamics was observed for the chemical reaction whose hysteresis potential is 'time reversible invariant'. A revision of the concept of 'time reversibility' to accommodate the above results is suggested. The general design of the reaction mechanism also allows for the use of conventional potentials and, by the utilization of switches, overcomes the bottleneck of computations which involve multi-body interactions.

Mathematics Subject Classification (2000). 65-{04,Z05}, 68-{04,W01}, 70-{08,F01,F16}.

Keywords. Hysteresis chemical reaction model, Thermodynamics of reaction, Kinetic properties.

The initial theoretical, program and algorithm development was done at Norwegian University of Science and Technology-NTNU, Institute of Physical Chemistry, (Trondheim, Norway) during a sabbatical visit 2000-2001 financed by University of Malaya. The method of MD used here conforms to the periodic boundary conditions and thermostatting algorithms developed or refined by T. Ikeshoji and B. Hafskjold for standard, non-reacting particles. Their approach is non-synthetic using traditional Verlet integration of Newton's equations of motion.

1. Introduction

Recently, experiments have detected the presence of magnetic hysteresis behavior at the single molecule level [1, 2]; synthesis of such systems is also a hot topic of research [3]. Such facts suggest that non-single-valued functions are involved in the phase trajectory of the system. A rational extension of this concept, which has profound theoretical implications, is to construct a dynamical trajectory where the region of formation of the molecule does not coincide with that of its breakdown. There has been a reluctance in the past to consider such loop or hysteresis systems because of the absence of experimental evidence of hysteresis behavior at the molecular level, and because of the influence of the belief of 'time-symmetry' invariance which discourages such a view. These factor led to the construction of dynamical pathways which were both single valued and did not have any loop or circular topology; a detailed mathematical examination of these common time symmetry presuppositions — so essential to physics — has been made [4, 5] and it was shown that such views are often not warranted or were incorrect. This work reports a workable model hysteresis reaction pathway which leads to thermodynamically consistent behavior, exhibiting properties that will require new developments in reaction theory, and also predicts the feasibility of such mechanisms in nature. It suggests a re-definition and extension of the ideas of 'time reversibility' and 'microscopic reversibility' to cater for the proposed mechanism. The dimeric particle reaction simulated may be written

$$2\mathrm{A} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{A}_2 \tag{1}$$

where k_1 is the forward rate constant and k_{-1} is the backward rate constant. The reaction simulation was conducted at a very high mean temperature, about $T^*_{set} = T^* = 8.0$, well above the supercritical regime of the LJ fluid by a factor of 10 times the magnitude of normal simulation temperatures in reduced units. At these temperatures, the normal choices for time step increments do not obtain without also taking into account energy-momentum conservation algorithms in regions where there are abrupt changes of gradient. The total system temperature for this equilibrium simulation has an uncertainty of about 10^{-5} LJ units when all particles, whether atomic or dimeric, are sampled ; all other quantities determined have greater uncertainty, due to the smaller presence of the species, or if only a layer in the cell is sampled for runs of less than 5M time steps. There have been various attempts to model chemical reactions with different objectives in mind [6, 7, 8, 9, 10, 11]. Some used generalized models with few details to predict the main features experiments might reveal [6] at the reaction coordinate close to the transition state (TS), such as what might occur within a solvent-caged reaction complex: A-H$\cdots$B $\rightleftharpoons$ A$\cdots$H-B . This particular pioneering approach [6] was further elaborated by Bergsma *et al* [11] in order to examine the limits of validity of TS theory (TST) by not carrying out an ab initio study of all the possible reactive

trajectories, but by examining trajectories constrained to the TS surface because of the limits of computing power. An example of an ab initio detailed chemical reaction approach with a 1000 atom system using an assumed 3 body potential for the exchange process $F + F_2 \rightleftharpoons F_2 + F$ is that of Stillinger *et al* [9] who admit that the procedure was 'very demanding'. The current study involves 4096 particles or atoms, and therefore is much improved where statistics are concerned. At the other extreme are generalized studies of hypothetical schemes [8] such as the 'chemical reaction' $A + A \rightleftharpoons B + B$ used to elucidate some kinetic properties. Clearly in such models, species A and B must represent complex systems that can be physically distinguished; in chemical applications, they might represent for instance *cis* and *trans* isomers of some compound or they might represent mesoscopic species. Some simulations do away altogether with the details of molecular dynamics based on dynamical laws [7], replacing them with the Ansatz that the details of the interaction between individual particles are not essential in the study of the statistical evolution of the system. Such an approach would make studies attempting to correlate the details of the dynamics to macroscopic properties difficult or obscure, despite the great savings in computer time, and therefore does not suit the purposes at hand. The objectives of the present study include:
(a) Designing a mechanically well defined reaction model with low computational demands and where the averaged motions of the dimer may be correlated with the macroscopic kinetic and thermodynamical properties and where no anomalies must be observed in the macroscopic results. Such an outcome would imply that the dynamics are reliable enough to be used in other studies.
(b) Introducing some degree of complexity to the dimer such as vibrational and rotational states for more detailed dynamical investigations.
(c) Utilizing the thermodynamically consistent model (as judged by the results of an equilibrium simulation) in nonequilibrium simulations.
Here we focus primarily on (a) above. To this end a new general algorithm (which will be discussed separately in another planned work) was used to conserve momentum and energy; (b) is represented in rotational studies and (c) in an NEMD simulation.

The following essential thermokinetic parameters will be determined and discussed in the sections that follow:

- The thermodynamic equilibrium constant through extrapolating the density to zero.
- The activity coefficient ratio.
- The standard Gibbs Free energy, Enthalpy and Entropy of the reaction through extrapolation.
- The Arrhenius activation energy and pre-exponential terms, which bears no immediate connection to the potential of activation in Fig. 1, and the rate constants of the forward and reverse reactions.
- Self-diffusion and rotational diffusion constants.

- The probability distribution for the kinetic energy of a labeled atom to test the Gibbs ensemble postulate relative to the dynamic (switching) Hamiltonians used here. Within experimental error, there appears to be full conformity to the Gibbs thermodynamical postulates.

The method appears very promising for quantitative simulations of real systems, and will be utilized in the years ahead for various reaction studies, including those for conventional molecules.

2. The model

We examine the dimeric particle reaction given in (1) above

$$2\mathrm{A} \rightleftharpoons \mathrm{A}_2$$

in a range of equilibrium fluid states all well above the LJ supercritical regime. This model resembles somewhat that of ref. [8] except that a harmonic potential is coupled to the products to form the bond of the dimer whenever the internuclear distance reaches the critical value r_f between two free atoms A.

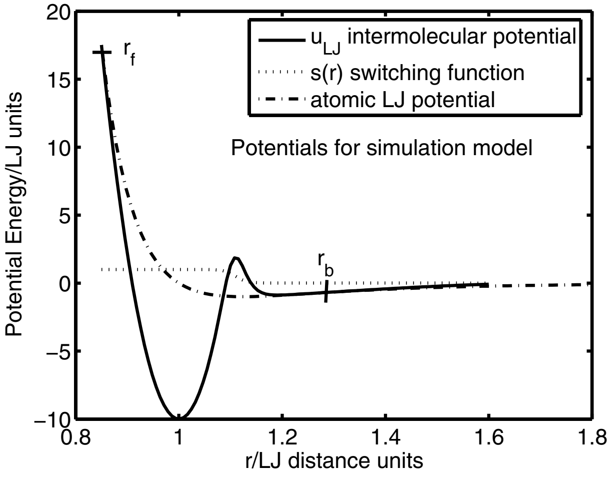


FIGURE 1. Potentials used for this work.

In the current study, the potentials as given in Fig. 1 are used, but other configurations are possible, as verified by direct simulation, such as the excited state configuration of Fig. 2 and the reduced distance model with the same spatial coordinates for the onset of the forward and reverse reactions in Fig. 3. This is a typical reaction potential and it is proposed that a quantitative simulation of a simple dissociation reaction of a diatomic gas such as H_2 be attempted. It was found that the equilibrium exchange rate of eqn. 1 was very low at lower temperatures and changed rapidly at higher temperatures to a saturation level for

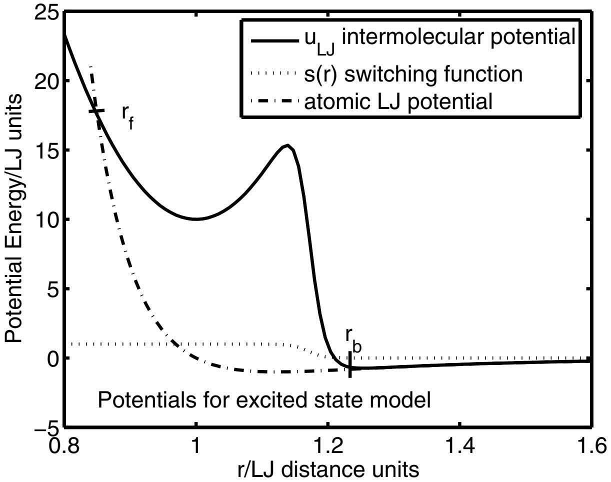


FIGURE 2. Potentials used for the excited molecular state.

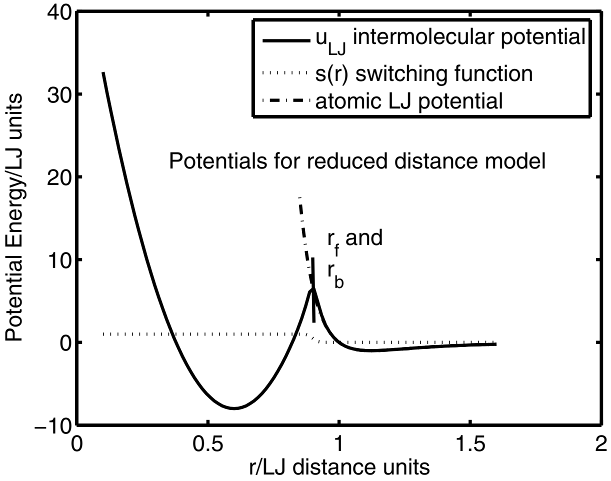


FIGURE 3. Potentials used for reduced distance molecular model.

the latter model (Fig. 3), not making it very suitable for studies where rates of formation and breakdown of bonds must be large enough for accurate statistics to be gained across the MD cell over a wide range of density and temperature ranges for a test system; the reason for the slow exchange is in part related to the small reaction or collisional cross-section of the molecule.

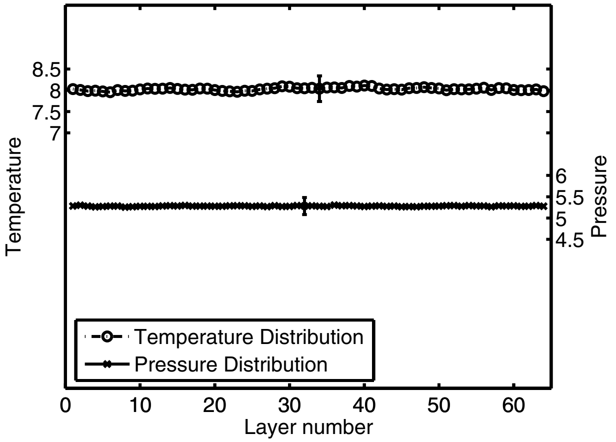


FIGURE 4. Pressure and temperature distribution across the MD cell.

The MD mechanism for bond formation and breakup is as follows. The free atoms A interact with all other particles (whether A or A_2) via a Lennard–Jones spline potential and this type of potential has been described in great detail elsewhere [12]. An atom at a distance r from another particle possesses a mutual potential energy u_{LJ} where

$$\begin{aligned} u_{LJ} &= 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] && \text{for } r \leq r_s, \qquad (2)\\ u_{LJ} &= a_{ij}(r-r_c)^2 + b_{ij}(r-r_c)^3 && \text{for } r_s \leq r \leq r_c,\\ u_{LJ} &= 0 && \text{for } r > r_c, \end{aligned}$$

and where $r_s = (26/7)^{\frac{1}{6}}\sigma$ [12]. The molecular cut-off radius r_c of the spline potential is such that $r_c = (67/48)r_s$. The sum of particle diameters is σ and ε is the potential depth for interactions of type A-A (particle-particle) or A-A_2 (particle-molecule) designated (1-1) or (1-2) respectively. The constants a_{ij} and b_{ij} were given before [12] as

$$\begin{aligned} a_{ij} &= -(24192/3211)\varepsilon/r_s^2,\\ b_{ij} &= -\,(387072/61009)\varepsilon/r_s^3. \end{aligned} \qquad (3)$$

The potentials for this system are illustrated in Fig. 1. Any two unbounded atoms interact with the above u_{LJ} (1-1) potential up to distance r_f with energy $E = u_{LJ}(r_f)$ when the potential is switched at the cross-over point to the molecular potential given by

$$u(r) = u_{vib}(r)s(r) + u_{LJ}\left[1 - s(r)\right] \qquad (4)$$

for the interaction potential between the bonded particles constituting the molecule, where $u_{vib}(r)$ is the vibrational potential given by eq. (6) below and the switching function $s(r)$ has the form given by eq. (7). LJ reduced units are used throughout this work, unless stated otherwise, by setting σ and ε to unity in the above potential description. The relationship between normal laboratory units, that of the MD cell and the LJ units, have been extensively tabulated and discussed [12] and will not be repeated here.

For the system simulated here with the potentials depicted in Fig. (1), the switching function is operative up to r_b, the distance at which the molecule ceases to exist, and where the atoms which were part of the molecule interact with the $(1-1)$ potential u_{LJ} like other free atoms; bonded atoms interact with other particles , whether bonded or free with the u_{LJ} (1-2) potential. The point r_f of formation corresponds to the intersection of the harmonic $u_{vib}(r)$ and u_{LJ} curves , and their gradients are almost the same at this point; by the Third Dynamical Law, momentum is always conserved during the cross-over despite finite changes in the gradient, since the sudden change of the force field is between only the two particles where the Third Law applies, thereby conserving momentum also. Total energy is conserved since the curves cross, and errors can only be due to the finite time step per cycle in the Verlet leap frog algorithm, which would cause the atoms to be defined as molecules at distances $r < r_f$. Similarly at the point of breakup, there is a very small ($\sim 10^{-4}$ LJ units of energy) energy difference between the LJ and molecular potentials, despite using the switching function in the vicinity of the region to smoothen and unify the curves; the small energy differences at the cross-over points are less than that due to the normal potential cut-off at distance r_c where the normal (unsplined) LJ potential is used in MD simulations.

In order to overcome this problem, a new algorithm (NEWAL) was developed (the details of which will be described in another work) which conserves momentum and energy at these two different types of cross-over points, where in one case, the switch is used (for breakup of the molecule) and not for the other during molecular formation. Briefly, if $E_p(r)$ is the inter-particle potential (energy) and $E_m(r)$ that for the molecule just after the cross-over, the algorithm promotes the particles to a molecule and rescales the particle velocities of only the two atoms forming the bond from $\mathbf{v_i}$ to $\mathbf{v'_i}$ $(i = 1, 2)$ where $\mathbf{v'_i} = (1+\alpha)\mathbf{v_i} + \beta$ such that energy and momentum is conserved, yielding $\beta = \frac{-\alpha(m_1\mathbf{v_1}+m_2\mathbf{v_2})}{(m_1+m_2)}$ (for momentum conservation), and the principle of energy conservation implies that α is determined from the quadratic equation $\alpha^2 qa + 2qa\alpha - \Delta = 0$ with $a = (\mathbf{v_1} - \mathbf{v_2})^2$,$q = \frac{m_1 m_2}{2(m_1+m_2)}$ and $\Delta = (E_p - E_m)$. Interchanging m and p allows for the same equation to be used for break-up of the molecule to free particles. For the simulations, success in real solutions for α for each instance of molecular formation is 99.9 % and 100% for breakdown, where the Δ value in this instance is very small ($\sim 1.0 \times 10^{-4}$). In these simulations, we ignored the cases when there was no solution to the quadratic equation, meaning no molecules are allowed to be formed at all, and the interactions are of the $(1-1)$ variety. This new algorithm, coupled with a shorter time step (from the typical 0.002^* for low-energy non-reacting systems to 0.00005^*), ensured excellent thermostatting, where the thermostatting was carried out at the ends of the box only, as is the case in most real systems. It should be noted that this smaller time scale is not unrealistic as the temperature for this system is of the order of $20-30$ larger than the usual values chosen, and so the translational kinetic energy of the particles would scale by the same order. In this equilibrium study, the MD cell (which is a rectangular box) is divided into 128 equal orthogonal layers in the x direction, which is of unit length in cell units. In this method of boundary conditions [12], the first 64 layers to the midpoint along the x axis are a mirror reflection about the plane parallel to the other two axes passing through this x axis mid-point. The y and z directions have length 1/16 each (cell units). This shape is chosen because non-equilibrium simulations will concentrate on imposing thermal and flux gradients along the x-axis, which would allow for more accurate sampling of steady state properties about this axis [13]. The layers that are mirror reflections about the mid-point plane are averaged for steady state thermodynamical properties, leading to effectively 64 layers. With this algorithm, with only end wall thermostatting, we sample each of the layers for temperature and pressure changes, and find that the profiles are rather constant, as shown in Fig. 4. The heat supply term (per unit time) is zero to within the error of fluctuation of energy. Without the algorithm, (but with the same time step increment) the center of the effective cell (layer 32) would have a temperature T^* higher than that of the thermostatted end layers by over 2 units, and the heat supply term would be significantly negative, implying a virtual heating up of the system at the middle due to the potential differences of the switches at the cross-over points which, because of to the finite time step increment will not conserve energy. The pressure too would be unrealistically higher at the center of the cell, which is unphysical for systems in thermodynamical equilibrium.

The algorithm above therefore is very effective in overcoming these problems. It should be noted that the uncertainty with regard to temperature for each of the layers would be about $10-100$ times larger than the total system temperature which is derived from averaging over every particle in the system, whether bonded or not. Prior to the implementation of this algorithm, each layer would be thermostatted to maintain a constant and uniform temperature and pressure profile (during the preliminary design). The non-synthetic thermostatting at only the boundaries of one direction of the cell approximates most physical systems; thermostatting each layer is used for heat of mixing studies but would not show the long-range fluctuational dynamics of energy transfer due to the thermostats, even if the noise levels are much lower. Further, for chemical reactions, there will be energy interferences due to the thermostatting of each layer, and so here, only the ends of the cell were thermostatted to eliminate any such effects, even if a greater uncertainty was introduced due to the long range bilateral transfer of energy from system to thermostat. It is found that the results of this study differ only by about 15% (for the equilibirum constant) to that found earlier when all the layers were thermostatted without implementation of the energy-momentum conservation algorithm. At regions $r < r_{sw}, s(r) \to 1$ according to (7) implying $u(r) \sim u_{vib}(r)$, i.e., the internal force field is essentially harmonic for the molecule and at distances $r > r_{sw}, u(r) \sim u_{LJ}$, so that the particle approaches that of the free LJ type as $r \to \infty$; the breakup is defined to occur at $r_b > r_{sw}$. Concerning the mechanism for the switching, in quantum mechanical kinetic descriptions, switch mechanisms are frequently used for describing potential cross-overs [14], but from a classical viewpoint one can suggest that the inductive LJ forces due to the particle potential field (with particles having a state characterized by state variables $\mathbf{s}_{LJ}$) cause the internal variables at the critical distances and energies mentioned above to switch to state $\mathbf{s}_M$ when another force field is activated for the atoms of the dimer pair. State $\mathbf{s}_M$ reverts again to state $\mathbf{s}_{LJ}$ at distances r_b.

Incidentally, the shape of the potentials and switching mechanism used here is surprisingly similar to *experimental* discussions of the charge neutralization reaction [14]

$$\mathrm{K}^+ + \mathrm{I}^- \to \mathrm{K} + \mathrm{I} \tag{5}$$

except that the discussion does not explicitly mention the crossing over of the KI and $\mathrm{K}^+\mathrm{I}^-$ potentials at short distances (high energy) due to the 'time-reversal' presuppositions referred to above. The existence of a cross-over would make the potential mathematically equivalent to the present treatment and there is good reason to suppose that such processes can and should occur in electro-magnetically induced reaction pathways (such as is manifested in charge-transfer and Harpoon mechanisms) especially since the KI potential curve exists at shorter distances well before the cross-over point. It is therefore postulated that there might well exist cross-over points not at the same vicinity for molecular formation and breakdown in actual reactions and that this simulation model is illustrative of such types of reactions. The following values were used here for the potential parameters:

(a) Current study (Fig. 1)

$u_0 = -10, r_0 = 1.0, k \sim 2446$ (exact value is determined by the other input parameters), $n = 100, r_f = 0.85, r_b = 1.20, \text{and}\, r_{sw} = 1.11$.

(b) Excited state model (Fig. 2)

$u_0 = 10, r_0 = 1.0, k \sim 2446$ (exact value is determined by the other input parameters), $n = 100, r_f = 0.85, r_b = 1.30, \text{and}\, r_{sw} = 1.17$.

(c) Reduced distance model (Fig. 3)

$u_0 = -8, r_0 = 0.6, k \sim 2446$ (exact value is determined by the other input parameters), $n = 100, r_f = 0.90, r_b = 0.90, \text{and}\, r_{sw} = 0.90$.

The intramolecular vibrational potential $u_{vib}(r)$ for a molecule is given by

$$u_{vib}(r) = u_0 + \frac{1}{2}k(r - r_0)^2. \tag{6}$$

A molecule is formed when two colliding free particles have the potential energy $u(r_f)$ whenever $r = r_f < r_0$, at the value indicated in (a) above. This value can be defined as the isolated 2-body activation energy of the reaction and has the value of 17.5153 at r_f. A molecule dissociates to two free atoms when the internuclear distance exceeds r_b (which in this case is 1.20). The switching function $s(r)$ is defined as

$$s(r) = \frac{1}{1 + \left(\frac{r}{r_{sw}}\right)^n} \tag{7}$$

where

$$\begin{cases} s(r) & \to 1 \qquad \text{if } r < r_{sw} \\ s(r) & \to 0 \qquad \text{for } r > r_{sw} \end{cases} .$$

The switching function becomes effective when the distance between the atoms approaches the value r_{sw} (see Fig. (1)).

Some comments concerning the MD potentials are in order. It is generally not correct to assume that the potentials in Fig. 1 represent the transition state theory (TST) potential surfaces; these surfaces can only be derived by computing the actual potential of the dimer or free atoms at a known internuclear distance in the presence of all the other species. The zero density limiting potentials of Fig. 1 cannot cause stable molecules to exist if they were formed by excited atoms with total kinetic energy in excess of the zero density activation energy, since if energy is conserved the formed molecule would (except for a finite number of kinetic energy values, depending on the model) have to dissociate again to the atomic states from which they were formed initially. There must be energy interchange at the potential well of the molecular species to remove energy so as to prevent dissociation. This is achieved through the presence of the temperature reservoir. This reservoir, if it is coupled to the system, would induce a system behavior whose limit at zero density would *not be* the same as an isolated mechanical system. Likewise, all standard states and other state functions of activation (free energy, entropy, etc.) must be computed as functions of all the coordinates of the particles involved in the interaction (including the reservoir). The numerical magnitude of these functions cannot be inferred only from the isolated potentials above; i.e., these potentials in conjunction with statistical mechanics should in principle yield the various system properties. Here, we extrapolate to zero density at fixed temperature to derive these functions, which cannot be inferred from mechanics only, nor from the potentials.

3. Thermodynamic results from equilibrium mixtures

The reacting mixtures considered here were in thermodynamic equilibrium with 4096 particles. The cell was thermostatted at the ends of the cell maintained at the same temperature.

3.1. Determination of accuracy of computation and convergence

It will be observed that the results provided without any adjustments are relatively smooth, even for this supercritical LJ system at relatively very high temperatures using non-synthetic thermostatting of the systems at the boundaries of the cell only. Although this method is closer to many experimental situations where thermostats are located at the boundary of the system, the transfer of energy to and from any volume element within the system to the thermostats via the molecular and particle interactions would imply a greater fluctuation in kinetic energy and possibly other forms of potential energy than if each particle were individually thermostatted through a synthetic algorithm. As mentioned before, the algorithm (where a separate study will be presented) assures of flat temperature and pressure profiles with end-point thermostatting; just reducing the time step without implementing the algorithm was inadequate in ensuring the flatness of the $P-T$ profiles. The steps and criteria used to ensure adequate sampling with energy conservation were as follows:

(1) For the time increment selected, and for runs for a particular $(\rho, T)^*$ duple combination whose system properties were to be investigated in detail, the following had to obtain:
(a) the heat supply to either of the reservoirs had to have a standard error of fluctuation about zero that was (much) less than one standard error. The actual heat supply term in an NEMD experiment is typically several orders of magnitude greater. This ensures that there is overall energy conservation. It was found that this situation obtained for the $(0.7, 8)$-duple which was used in testing various properties. In particular, the run length was varied, as follows, at 3M, 4M, 6M and 8M (where the set of values will be denoted $\mathcal{M}$)with the above criterion obtaining in each case of the set values. Hence the length of the run at 10M was chosen as a safe figure, where, incidentally, the above still obtained.
(b) the $P-T$ profile had to be flat for all these combinations of conditions.
(c) properties of interest, especially the concentration equilibrium constant, rate constants and probability distributions were also viewed at this particular duple value and for $\mathcal{M}$, and the variation was all within the vicinity of the errors given in the text for the duple concerned.

(2) For some of the algorithms, such as the ones for the diffusion coefficients, the maximum possible time prior to molecular breakdown was used (absolutely no extrapolation was attempted) in computing the coefficient from the Einstein expression, and so would be independent of $\mathcal{M}$ for large enough runs, (where the total duration of the molecule in general does not exceed about $20,000$ units of δt^*) which means that there is no problem with the choices of $\mathcal{M}$ or $10M$. Likewise, for the probability distribution, the sampling is done at each $15th$ time step, and so depicts in general very low scatter, so as to be able to discern some features such as apparent temperature differences, as discussed in the sequel to this work.

Typical runs of 10 million time steps were performed per run at each general particle density ρ (where ρ is determined as a general density irrespective of whether the particle is free or is part of a molecule), where the first $200,000$ steps were discarded so that proper equilibration could be achieved for our data samples. The sampling methods have been previously described [12] where sampling of all data variables were done each 20^{th} time step and where there were 100 dump values where each dump consists typically of 5×10^5 samples which are averaged. The 100 dump values are then averaged again to yield the standard errors of all variables. Dynamical quantities however had to be sampled at each time step $\delta t^* = 0.00005$. The thermostatting method conserves momentum and registers the energy absorbed at the thermostats [15]. All parameters given here are relative to LJ reduced units, sometimes denoted by $*$.

3.2. Equilibrium constants

There are two independent methods that are attempted here, each of which leads to the same results. The non-kinetic method (3.2.1) directly determines the concentration of reactants and products, and infers from these quantities K_{eq} at $T^*_{set} = 8.0$, whereas the kinetic method (3.2.2) infers K_{eq} from taking ratios of the computed forward and backward rate constants at the same temperature.

3.2.1. Non-kinetic method. In order to find the thermodynamic equilibrium constant, K_{eq}, the following procedure was adopted. The concentration ratio, K_c defined as

$$K_c = \frac{x_{\mathrm{A}_2}}{x^2_{\mathrm{A}}}, \tag{8}$$

was determined as a function of average system density, ρ where the x's represent number density concentrations. For this and all other equilibrium quantities, the system temperature was set at $T^*_{set} = 8.0$, with the actual temperature fluctuating error of order $< 10^{-4}$. At very small densities, the system becomes an 'ideal' mixture, but as mentioned previously, the limit of the potentials cannot be the same as the isolated potentials used in the MD calculations, since if this were the case, all the molecules would break up, yielding a net zero value for the equilibrium constant at the limit of zero density. As another project, it would be of interest to determine the limiting density and thermostatting time intervals at which the equilibrium regime breaks down in this system, and to elucidate the theory when this occurs. There may well be technical difficulties involved in computations of very low density systems though. The plot of $K_c = K_c(\rho)$ is shown in Fig. 5. The accuracy of the K_c values varies inversely with a function of ρ, where in the captions *sd* refers to the number of standard deviations of the standard error. At low densities, fluctuations in K_c implies that any extrapolative method can be ruled out, unlike previously (when NEWAL was not devised) when all the layers were individually thermostatted and where a least squares fit n-order polynomial expansion $p(x) = \sum_{i=0}^{n} a_i x^i$ to derive the zero density limit of the concentration ratio was utilized; the value of n was between 2 to 4. The zero density limit K_0 where $K_0(T^*) = K_c(\rho \to 0)$ is the true equilibrium constant. It is clear that in this system K_0 and K_c in general differ significantly; it serves as a warning that, in general, one cannot ignore activity coefficients in the calculation of such properties in model systems and theoretical demonstrations if semi-quantitative results are desired. In the present study, it was discovered that at very low densities, fluctuations are significant, as shown in Fig. 6 for the case of a run at $T^*_{set} = 8.0$. The method used in the present case is to take the mean value of K_c for very low ρ values (rarefied state) ranging from 0.03 to 0.09, for about 12 values at any one temperature and to approximate this as $K_0(T^*)$. The fluctuations show that in this range of density, the system has 'saturated' itself in that all the ρ values yield approximately the same mean K_c. Also, at such exceedingly low densities,

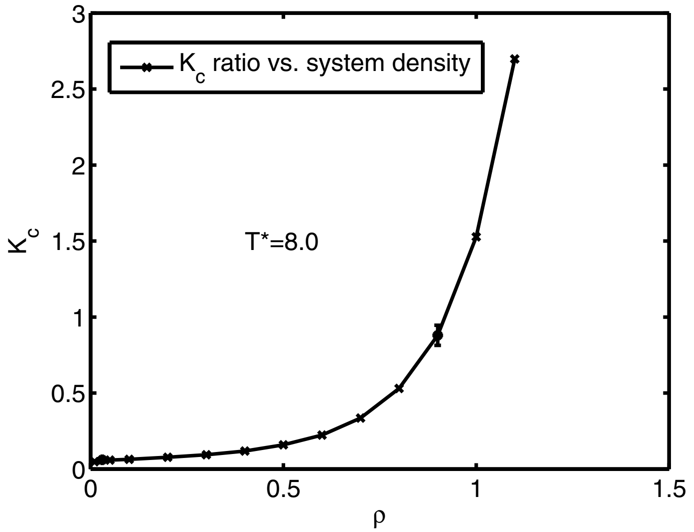


FIGURE 5. Variation of concentration ratio K_c with ρ, the system number density at LJ temperature $T^*_{set} = 8.0$ with $sd = 3$ at $\rho = .03$ and $sd = 50$ at $\rho = 0.7$.

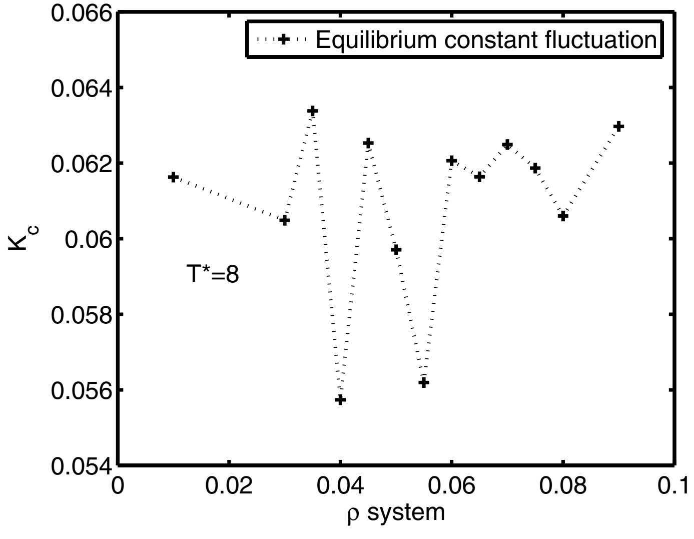


FIGURE 6. Illustration of fluctuation of individual runs at rarefied density state.

one would expect a larger fluctuation in the determination of the rate values; nevertheless, we notice a saturation, with a maximum scatter of values for K_c of about ±0.006. In view of the fact that at much higher densities, the absolute change of this constant is very much greater for unit change of density, the errors are still relatively not large. The results derived for $T^* = 8.0_{set}$ are

$$K_{eq}(T) = \lim_{\rho \to 0} K_c(T) = 0.0610 \pm .002 \text{ LJ units.} \tag{9}$$

In previous studies prior to NEWAL implementation, using polynomial extrapolation, a value of $0.050 \pm .001$ was derived. However, these two values, although close, need not coincide because the phase-space trajectory of the two systems are not the same theoretically, meaning they are not the 'same' chemical reaction system, even the only alteration here involves the time step and the thermostatting of each layer. (A change in the time step increment would alter the phase-space trajectory; so would the thermostatting mechanism.) Knowing $K_{eq}(T)$ from (9), which is an invariant quantity for any one temperature, the activity coefficient ratio, Φ can be calculated for the other densities at the same temperature by using

$$K_{eq} = K_c \frac{\gamma_{A_2}}{\gamma_A^2} = K_c \Phi. \tag{10}$$

The ratio of activity coefficients Φ is shown as a function of density in Fig. 7.

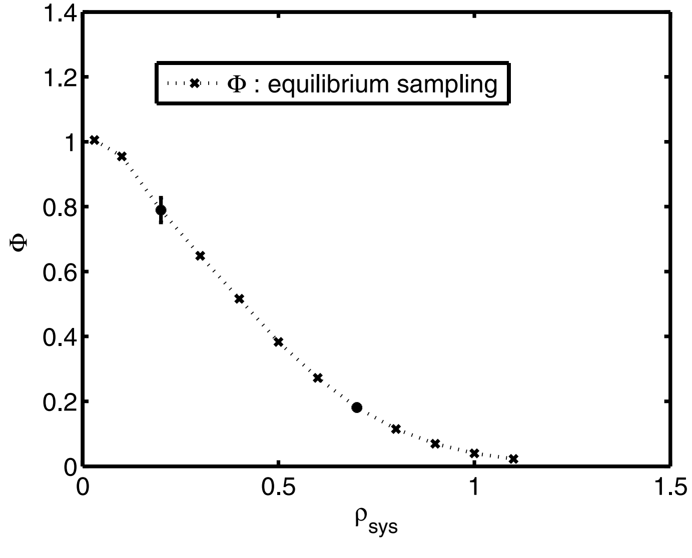


FIGURE 7. Variation of Φ with ρ, the system number density at LJ temperature $T^*_{set} = 8.0$.

It is clear from the Φ ratio that for normal densities, the equilibrium reaction mixture is highly non-ideal, which may be expected due to the large differences in the LJ energy well for the molecule and the atom (see Fig. 1). It is probably a poor approximation to use ideal models for test systems in reactor design, which is often the practice. Further, the above technique allows for the general determination of activity coefficient ratios via simulation. The determination of separate activity coefficients is a challenge. One real problem is the fact that molecules in the *equilibrium* state cannot exist in isolation. In mixtures, either the reaction goes to completion, or they do not react, as in the simple theory of mixtures. In the latter case, one might postulate separate ideal states for the 'pure' components, but in the present elementary case, for any one temperature, there is a finite value for K_0 meaning the presence of all components in a system at equilibrium. It is therefore a challenge to find a suitable model or concept to solve this problem with cycle changes. Even if a hypothetical state were defined, one must still design the route or cycle taken to the equilibrium state which consists of product and reactant species. The derivation would require a series of very elaborate and detailed computations and is not attempted here since it is not immediately relevant. Nevertheless, from the equilibrium distribution at various temperatures, the standard enthalpy, entropy and Gibbs free energy can be computed. Traditionally, many have interpreted these quantities as reflecting function changes for 'pure' component reactants to pure molecular product without any simultaneous presence (or equilibrium) between the two.

3.2.2. Kinetic rate method. The rate constant is a defined quantity, with the standard form below. The overall rate of reaction r may be written in terms of the experimentally determined forward rate ($r_1 = k_1 x_A^2$) for the process $2\mathrm{A} \overset{k_1}{\rightarrow} \mathrm{A}_2$ and backward rate ($r_{-1} = k_1 x_{\mathrm{A}_2}$) for the process $\mathrm{A}_2 \overset{k_{-1}}{\rightarrow} 2\mathrm{A}$ as $r = r_1 - r_{-1} = k_1 x_{\mathrm{A}}^2 - k_{-1} x_{\mathrm{A}_2}$; k_1 and k_{-1} are the respective rate constants.

At equilibrium $r = 0$, and so

$$\frac{x_{\mathrm{A}_2}}{x_{\mathrm{A}}^2} = \frac{k_1}{k_{-1}}. \tag{11}$$

The ratio of rate coefficients is the concentration ratio K_c where

$$K_c = \frac{k_1}{k_{-1}}. \tag{12}$$

To verify the above equilibrium constant independently from concentration measurements used in the previous section, one can extrapolate to zero density ρ the values for $r_1/x_{\mathrm{A}}^2 = Q = k_1$ and $r_{-1}/x_{\mathrm{A}_2} = R = k_{-1}$. The rates were calculated independently from the program by monitoring the number of bonds formed or broken for each time step δt^* and averaging this quantity over the $10M$ time steps. Then the relevant equations are

$$\lim(\rho \to 0)\left(\frac{Q}{R}\right) = K_{eq} = \frac{\lim\ Q(\rho \to 0)}{\lim\ R(\rho \to 0)} = \frac{Q^0}{R^0}\,. \tag{13}$$

The plots of Q and R at low densities are given in Fig. 8.

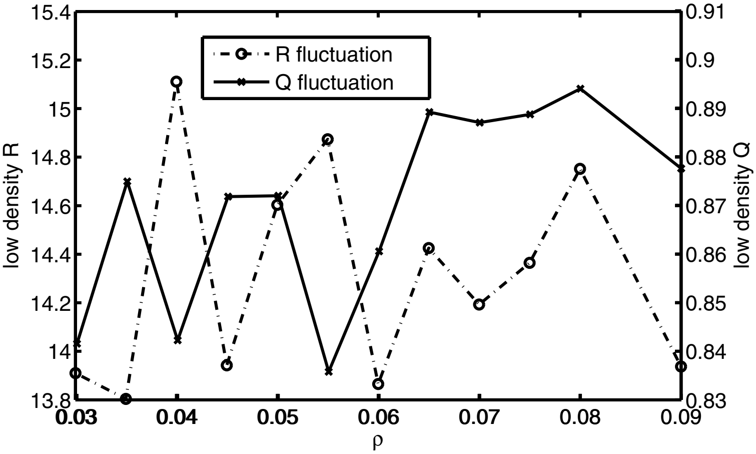


FIGURE 8. Values of Q and R variables at $T^*_{set} = 8.0$ and at rarefied densities showing "saturation" behavior to a mean value as required by the limit theorems. At such low densities, fluctuations are observed with an even scatter about the mean value.

As for the direct determination of the equilibrium constant from concentration measurements, fluctuations imply an averaging at very low densities of the values given in the figures to derive the limits. The results with the estimated errors are

$$\langle Q\rangle_{\{\rho<0.09\}} = \lim_{\rho\to 0} Q = Q^0 = 0.870 \pm .006 \text{ L.J. units}, \tag{14}$$

$$\langle R\rangle_{\{\rho<0.09\}} = \lim_{\rho\to 0} R = R^0 = 14.32 \pm .1 \text{ L.J. units}. \tag{15}$$

It will be noticed that at very low densities, we would expect the number of errors due to the breakdown process to be very much higher than that due to the formation process, since the number of dimers tends to a low number and this is reflected in the R^0 uncertainty. The ratio of the values given in (14) and (15) gives the true equilibrium constant according to (13) where

$$K_{eq}(\text{kinetic}) = \lim_{\rho\to 0} \frac{k_1}{k_{-1}} = 0.061 \pm .001\,\text{L.J. units}. \tag{16}$$

This kinetically derived result is in excellent agreement with the results from the previous method. The agreement indicates that the system is in a steady (equilibrium) state and that the simulation method is fairly coherent. The Q and R functions at other densities are given in Fig. 9.

3.3. Standard states

We use the form $\Delta G^0(T) = -kT \ln K_{eq}$ to determine the standard free energy state $\Delta G^0(T)$ of the dimer reaction. The justification is that we can choose the standard state to be at constant pressure (of zero value) for the standard state, implying

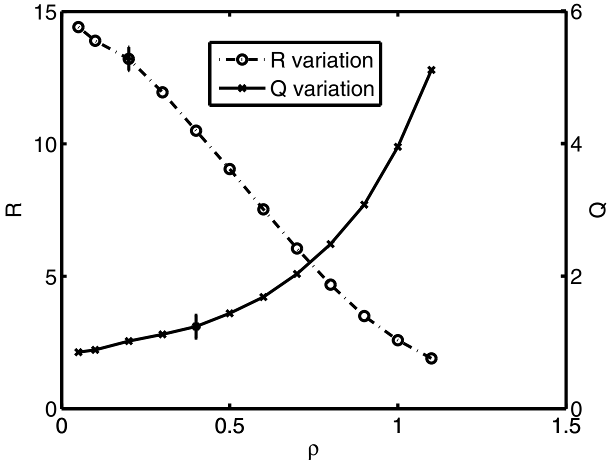


FIGURE 9. Variation of Q and R variables with density at $T^*_{set} = 8.0$.

that the chemical potential standard state for each species is only a function of temperature, so that $\Delta G^0(T)$ is strictly only a function of temperature [16, p.177–179]. We repeat the same process as described above in section (3.2) for $T^*_{set} = 8$ for different temperatures (from $T^* = 4 - 20$). Each determination required at least 8 runs at varying low densities. It was found that at low temperatures, the fluctuations were greater, as shown in Fig. 10 where the variation of K_{eq} versus $1/T$ is given. The linearity of this curve can also be used to derive an average value for each of the quantities calculated below for the entire temperature range. The curve used to determine the other standard state functions was the Gibbs free energy curve, given in Fig. 11. For this curve, the error bars (except for the first data set) all refer to the errors relative to the least squares fit of a quadratic curve to the simulation result. The fit is rather good. The standard entropy $\Delta S^0(T)$is derived from the thermodynamical entity [16, eqn. 6.34, p.182]

$$\frac{d\Delta G^0(T)}{dT} = -\Delta S^0(T). \tag{17}$$

Clearly to use (17), we must know $\Delta G^0(T)$as a function of temperature T. We write therefore a simple quadratic equation with p coefficients:

$$\Delta G^0(T) = p(1)T^2 + p(2)T + p(3). \tag{18}$$

The nonlinear least squares method yields

$$p(1) = -0.0233441, \qquad p(2) = 1.0531305, \qquad p(3) = 15.46544989$$

with an overall uncertainly of the free energy as approximately ±0.3. Differentiating (18) yields the entropy as

$$\Delta S^0 = -(2p(1)T + p(2)),$$

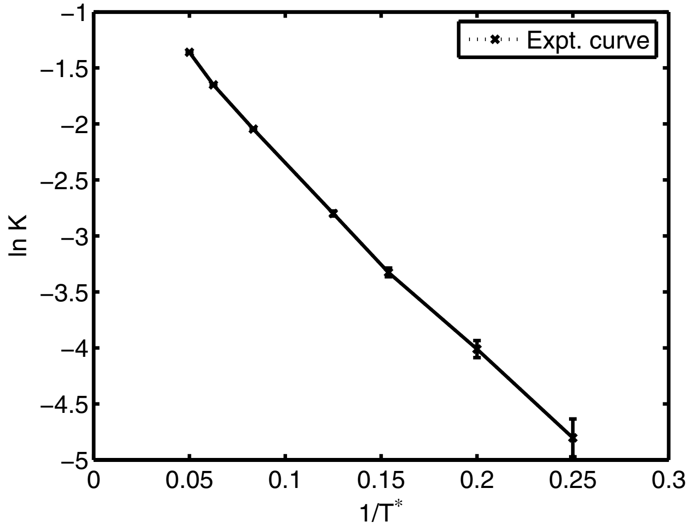


FIGURE 10. Variation of equilibrium constant K_{eq} with $1/T$ for fixed average system $\rho = 0.70$ with best fit line.

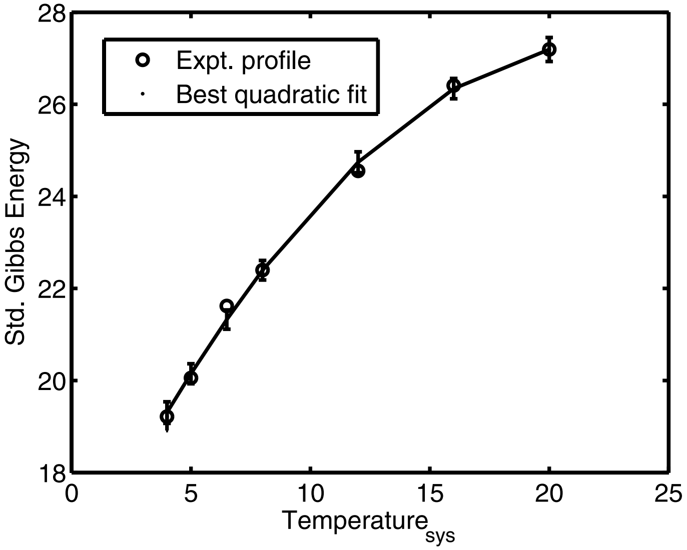


FIGURE 11. Variation of the standard Gibbs Free Energy ΔG^0 with temperature.

which is linear. The standard enthalpy ΔH^0 is given at constant temperature by the entity [16, p.183]

$$\Delta H^0(T) = \Delta G^0 - T\Delta S^0, \tag{19}$$

which therefore means that the standard enthalpy is given by

$$\Delta H^0 = -p(1)T^2 + p(3).$$

It can be verified that this expression and that for ΔS^0 recovers the quadratic (18).

The plots for the standard entropy and enthalpy as functions of temperature are given in Fig. 12.

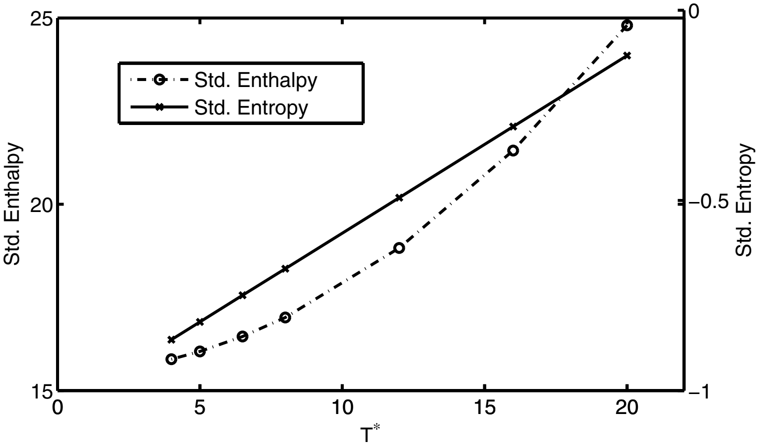


FIGURE 12. Plot of standard enthalpy $\Delta H^0(T)$ and entropy $\Delta S^0(T)$ from $\Delta G^0(T)$ quadratic curve fit with the temperature T^*.

Most experimental methods take gradients to yield average values of the standard states over a temperature range. Here, the explicit values can be calculated over the temperature range. From the calculations, we find that the standard entropy is negative, as it must be at moderate to low temperatures since the free particle state has a larger phase space than the corresponding dimer. It may appear counter-intuitive that the standard enthalpy is positive. It must be pointed out that at these temperatures, the particles are not trapped at the bottom of the potential well, and that the activation energy is positive, and that the internal potential energy at the point of formation of the molecule is not lost, but is converted to internal kinetic energy even up to the point of the break-up of the molecule, implying a positive value of this quantity relative to the dissociated particles. A quantitative treatment of these terms has been attempted [17]. It must be concluded that the simulations are able to determine the standard states without having to construct extremely detailed cycle diagrams; further, the simulation can also check on the correctness of the cycle diagrams used to determine standard state values.

3.4. Activation energies

From the way the algorithm was constructed for molecular formation, the molecularity of the elementary reaction is 2, leading to a single second-order reaction of formation, and for the dissociation of A_2, a first-order reaction results since the molecule is only allowed to exchange kinetic energy with all other particles within the system without further reactions to the dissociation limit. A frequently used model for the kinetic constant k_i for these rates, due to Arrhenius, has the form

$$k_i = A_i \exp\left(-\frac{E_i}{RT}\right) \tag{20}$$

where the rate constant is a function of the temperature only and where A_i is ideally not temperature dependent. It should be noted that the Arrhenius equation is strictly valid for 2-dimensional systems where the pre-exponential factor is independent of temperature and where the exponential factor $\exp\left(-\frac{E_i}{RT}\right)$ represents the fraction of molecules having energy in excess of E_i [18], where E_i is usually understood to be the activation energy. The reason why this form is so durable is that the exponential term represents the fraction of excited state atoms, and this term dominates over the pre-exponential term with temperature variation, which gives the impression of a constant A_i factor for the plots. The rate constants for the forward k_1 and reverse reaction k_{-1} were plotted versus $1/T$ for the given density of $\rho = 0.7$ and was found to be reasonably linear (Figs. 13, 14, with the activation energies for the forward and the backward reaction rates (E_1 and E_{-1} respectively) and the corresponding collision factors (A_1,A_{-1}) determined approximately as

$$\begin{aligned} E_1 &= 21.40 \pm .10 \text{ LJ units}, & A_1 &= 3.50 \pm 0.2 \text{ LJ units}, \\ E_{-1} &= 7.26 \pm .02 \text{ LJ units}, & A_{-1} &= 2.70 \pm .04 \text{ LJ units}. \end{aligned}$$

There are two separate rate constants here, for first and second order. The second-order forward rate constant k_1 has a form given by

$$k_1(T) = \pi b_{max}^2 \left(\frac{8kT}{\pi\mu}\right)^{1/2} \exp\left(-\frac{\epsilon^*}{kT}\right) = A_1 \exp\left(-\frac{\epsilon^*}{kT}\right) \tag{21}$$

according to 'simple collision theory' (SCT). Very roughly, if the mean temperature for the plot (which spans from 4 to 20) is 12, then (21) above yields for the given value of A_1, $b_{max} = 0.9153.....$, which is reasonably close to 0.85, the theoretical value. However, $\epsilon^* = 21.40$, which is higher than 17.5153, the set simulation potential value for the formation of a molecule. Since we can expect a yet greater accuracy for the determination of ϵ^* as compared to A_i due to the domination of the exponential terms, it may be safe to suppose that other factors contribute to the true activation energy other than what is described by SCT. Future work will attempt to determine what other energy factors are implicated in ϵ^*; currently, SCT views this energy as a pure mechanical work energy, which obtains at the molecular level. Similarly, variation of A_i with various energy terms cannot be immediately ruled out. Generally, the above values do not bear a direct relationship to the isolated 2-body potentials of Fig. 1, but nevertheless some approximate

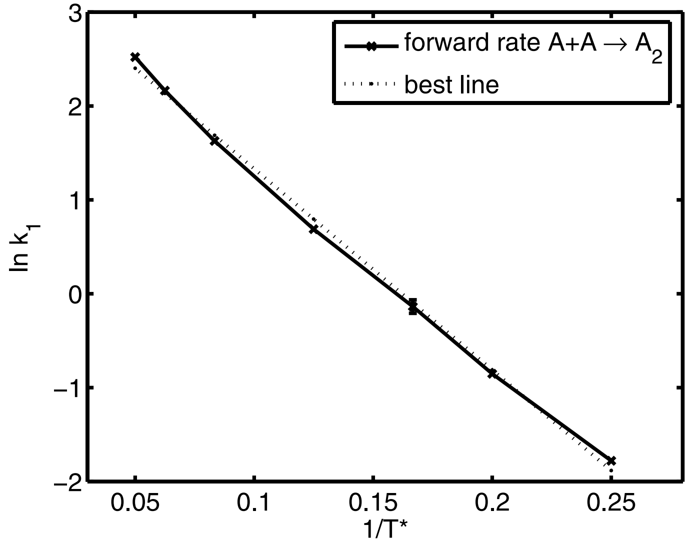


FIGURE 13. Variation of natural logarithm of forward (product forming) rate constant k_1 with reciprocal of temperature for $\rho = 0.7$.

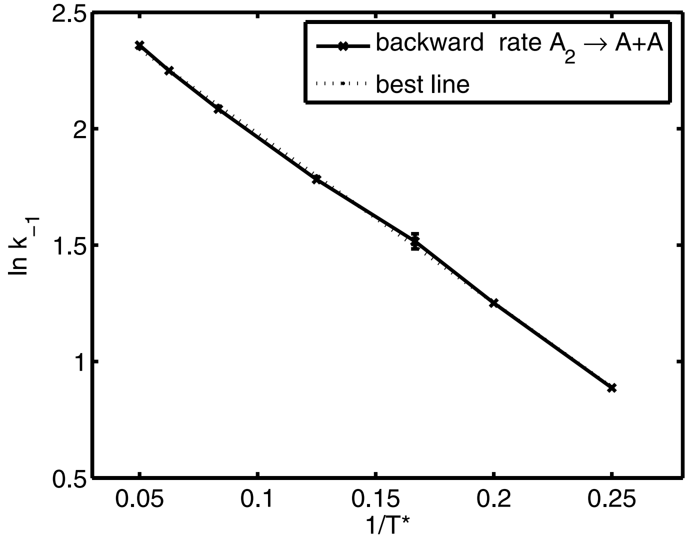


FIGURE 14. Variation of natural logarithm of backward (product disintegration) rate constant k_{-1} with reciprocal of temperature for $\rho = 0.7$.

correlations are evident; E_1 is somewhat close to the isolated activation energy 17.5153 measured from the free atomic states, and likewise E_{-1} is somewhat close to the energy difference from the bottom of the molecular potential at -10 to the potential at r_b, a distance of approximately -9 energy units. However, for a first-order reaction, a different interpretation for energy differences obtains than that due to SCT, which is concerned with bimolecular processes; the first-order interpretation is that the molecule decomposes when it overcomes an energy activation threshold, and the fraction of such molecules is reflected in the exponential term, the pre-exponential term reflecting the mechanism of the decomposition.

4. Results from equilibrium dynamical trajectory analysis

This section concentrates on variables which had to be sampled at each time step of duration $\delta t = 0.00005^*$ in order to compute the property of interest: the rate of reaction in the previous section above is also based on instantaneous sampling but more properly belongs to topics associated with equilibrium. Of importance in nonequilibrium and kinetic studies are the values of the diffusion coefficients, reaction correlation coefficients and the energy probability distributions, which if the principle of local equilibrium (PLE) obtains, imply that we may approximate the values computed in an equilibrium simulation for those in a nonequilibrium volume element having the same state variables. Examples of these quantities (which can also gauge the appropriateness of the model for nonequilibrium studies) are provided.

4.1. Rotational diffusion constants

Although connected in some ways to diffusion, a somewhat unconventional 'reorientation' diffusion function $\langle \cos\phi(t)\rangle$ has been defined [6] where $\phi(t)$ is the angle *between* $\hat{\mathbf{R}}(0)$, the unit internuclear distance vector of the dimer at $t = 0$, and $\hat{\mathbf{R}}(t)$, the same unit vector at time t. Such a definition might have applications in conjunction with their being part of transform functions [6, eqs. (17)–(20), p. 211], where the postulated exponential decay of this function when acting as a kernel of the transform could force convergence of the function being convoluted. It is found that the exponential decay assumption in $\cos(\phi(t))$is a fair but not perfect fit, perhaps implying that another type of theory for 'rotational diffusion' constants may yield even better fits with the experimental curves. We provide one such example $\langle \arccos(t)\rangle$, an approximation to $\langle\theta(t)\rangle$, which provides a far better fit and therefore is a candidate for another area of research in stochastic theory of rotational diffusion. It must be mentioned, however, that the theory of 'rotational diffusion' as developed by P. Debye and others [19, p. 81–84, esp. eq. (49)] etc. makes use of 'dissipation kinetics' where a constant torque M is balanced by an inner frictional force ζ parameter, so that $M = \zeta\frac{d\theta}{dt}$, where θ is an angular displacement. Such a theory leads to a relaxation in the distribution function f by a factor $\varphi(t)$ given by $\varphi(t) = \exp -\frac{2kT}{\zeta}t$, so that for a particular orientation angle θ, f has the form

$f = A\,[1 + C\varphi(t)\cos\theta]$. The mean dipole moment of the entire sample also decays with the same rate as with φ. It is not immediately clear that the orientation angle must also relax according to a first-order rate law. If the effect is a projection of an orientation onto an axis, then this would correspond to the result given by Allen et al. [6]. O'Konski and Haltner [20] have characterized TMV (a virus) by studying the birefringence relaxation rate, written $\delta = \delta_o \exp(-t/\tau)$ where τ_o is the initial value of birefringence [20, eq. (3), p. 3607], and the 'rotational diffusion coefficient' D_h is defined here as $D_h = 1/6\tau$ with an additional factor of $1/3$ compared to that of Allen. Most of these theories suppose that even at the molecular level, one can use frictional coefficients, as for macroscopic systems where the retarding force is linearly proportional to some form of velocity of the system, the constant of proportionality involving the frictional coefficients [21]. More recent experimental studies of rotational diffusion [22, 23] assume a first-order relaxation of fluorescent directed intensities of the chromophore of the molecule with the rotational diffusion constant defined as in [20]. To show that the results obtained are typical, we graph the functions as defined by Allen et al. [6]. The method used here to determine $\langle\cos\phi(t)\rangle$ is to create a table whenever a molecule is formed which maps out for each increment in the time step i the value of $\cos\phi(i)$ until it disintegrates: for each i^{th} time step there exists for each sampling subinterval M (M being a variable) values of $\phi(i)$ due to other molecules which have existed, and the average value for each sub-interval is computed as $\langle\cos\phi(i)\rangle = \sum_{j=1}^{M}\cos\phi_j(i)/M$. According to Allen et al. [6], the function decays as

$$\langle\cos\phi(t)\rangle = A\exp(-t/\tau_1)\ (A = 1)$$

with linearized form

$$\ln\left(\langle\cos\phi(t)\rangle\right) = -t/\tau_1 \tag{22}$$

where the 'rotational diffusion' coefficient D_r is given by $D_r = \frac{1}{2\tau_1}$. The results of the simulation are graphed in Figs. 15–17. Fig. 15 graphs the proposal found in [6]. It is clear that there is an initial chaotic regime, followed by a very slow decay of approximate form $A\exp(t/\tau_r)$,$(A = 1)$ if we measure the time from the end of the chaotic regime onwards; fitting this portion of the curve from the $400th - 800th$ time step to the above exponential yields $\tau_r = 1.38 \pm .02$LJ units. A 'rotational diffusion constant' $D_r = \frac{1}{2\tau_r}$ may be defined and the value obtained is $D_r = 0.36 \pm 0.01$LJ units. The shape of the $\langle\cos\phi(t)\rangle$curve resembles that described in [6] (where the 'initial chaotic region' is mentioned) implying a somewhat typical rotational motion, but it is clear from the figure that even in the fitting region, there is an apparent concave shape, as the tangent line makes clear. Nevertheless, for the sake of parametrization, this particular definition is used to derive the diffusion constant D_r data at other regimes of varying ρ (at constant temperature) in Fig. 17 and for varying temperature (at constant ρ) as depicted in Fig. 16, all of which are determined from the gradient between the $400th$–$800th$ time step, i.e., in these figures, the same method of determining D_r was used as for the above determination of D_r at $\rho = 0.7$ and $T^* = 8$. As with the case of rectilinear diffusion

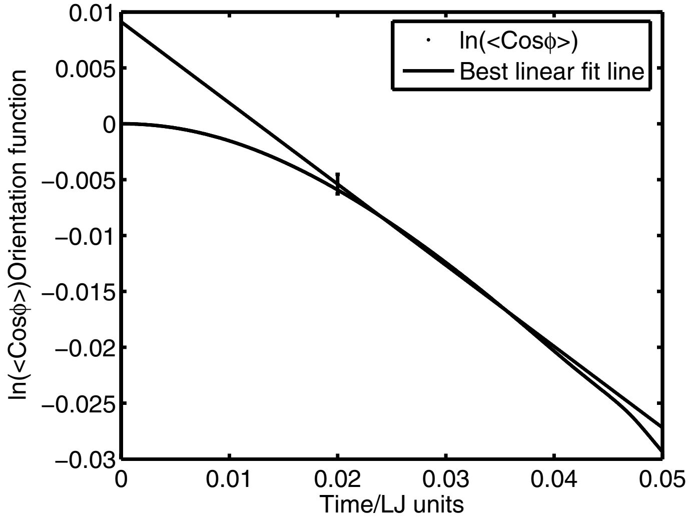


FIGURE 15. Variation of natural logarithm of $\cos\phi$ orientation function with time at $T^* = 8.0\,\text{and}\,\rho = 0.7$.

motion $D_t = BkT$, where B is the density dependent mobility coefficient, which is the steady state velocity acquired per unit external force [24, sec.14.4, eqns. (2-11), p.464-465], we obtain at fixed density ρ a linear relationship with temperature, suggesting a similarity or isomorphous theoretical construct in relation to rotational motion. Noting that different thermodynamical variable regimes are associated with different error margins when determined experimentally, we also notice an approximate linear correlation with density at fixed temperature. From the rectilinear equation, this would be the case if the mobility coefficient B were inversely linearly related to the density of the medium, which is a very reasonable assumption at higher densities $(\rho^* = 0.75 - 1.0)$. The figures show that the change of the diffusion constant with ρ at fixed temperature is much less dramatic than with temperature at fixed ρ.

Fig. 18 gives a clear indication that the long-time correlation concerning time and the logarithm of θ shows a very good linear fit (i.e., $\ln\theta vs.\ t)$, and so one can also derive a rotational diffusion coefficient where the actual angular distance relaxation is a first-order process by creating an appropriate theory as suggested by the computations (at least for the model adopted here). Lynden-Bell (R.M.) [25] has written an extensive review of the theoretical underpinnings of molecular reorientations; she concentrates on the concept of angular momentum as an indicator of reorientations. Many possibilities present themselves concerning the reorientation correlation function relaxation in time, which has the form $C_{J\alpha}(t) =< \mathbf{J}_\alpha(t) \,.\, \mathbf{J}_\alpha(0) > / < \mathbf{J}_\alpha(0) \,.\, \mathbf{J}_\alpha(0) >$ with α denoting the orientation with respect to a particular molecular axis. $\mathbf{J}_\alpha$ denotes the angular momentum about the designated axis of rotation. This $C_{J\alpha}(t)$ correlation function can have

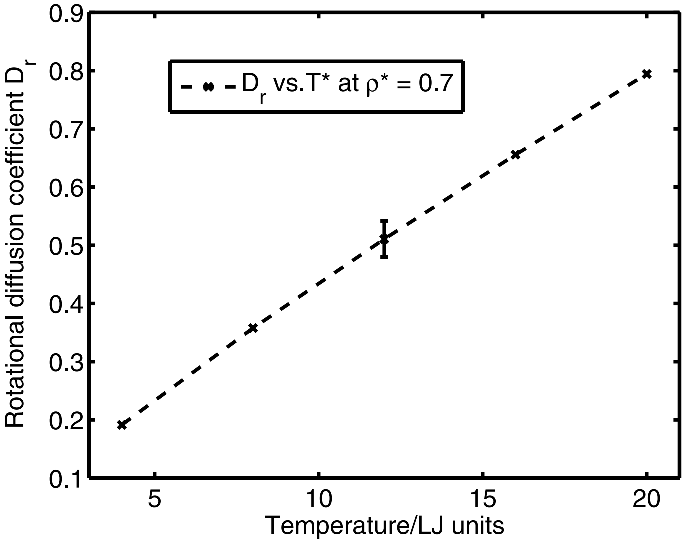


FIGURE 16. Variation of D_r with temperature at constant $\rho = 0.7$.

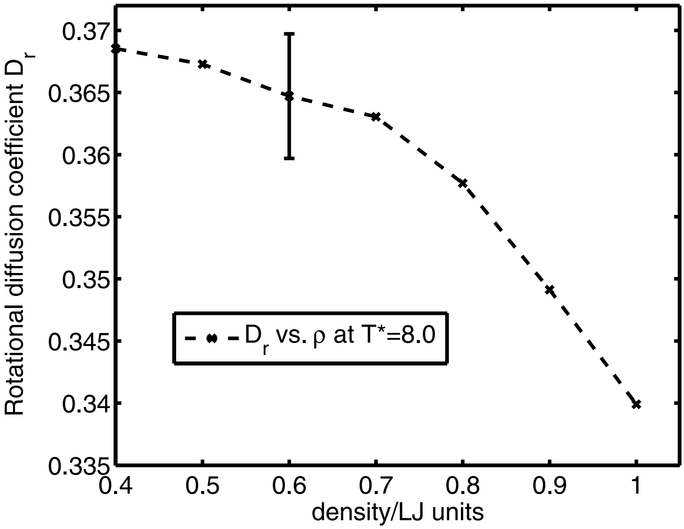


FIGURE 17. Variation of D_r with density at constant temperature $T^* = 8$.

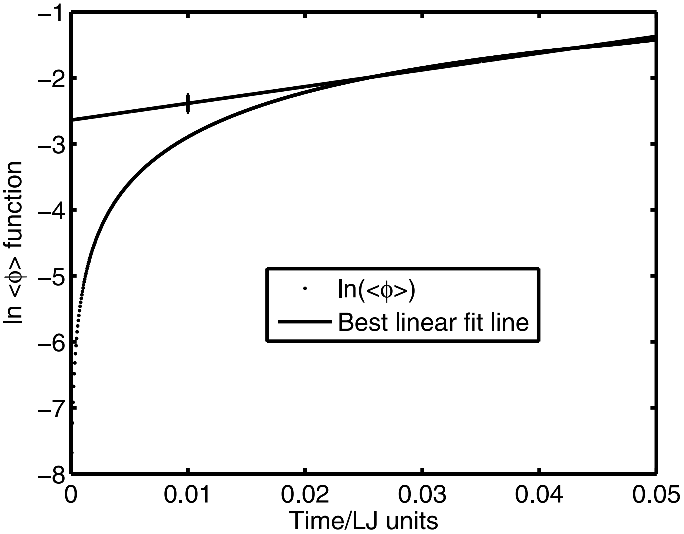


FIGURE 18. Variation of natural logarithm of arccos(cos ϕ) orientation function with time at $T^* = 8.0 \, \text{and} \, \rho = 0.7$.

an exponentially decaying form if the Fokker–Planck or J diffusion model is used. This is one of several possibilities [25, Fig. 1, p. 503] but for this dimer (a linear molecule), an exponential decay (linear in $\ln(C_{J\alpha})$ may be expected, where the angular velocity ω correlation is identical to $C_J(t)$. Here we note a long time linear correlation with $\ln\theta(t)$ and not its derivative. It could very well be that if the actual angular momentum were monitored, then an exponential decay with time would be observed in the current model, or that relative to those theories which predict an exponential decay with the ω correlation function, the current result for the evolution of $\theta(t)$ is in accordance with it; if not, then another theoretical approach may be feasible, complementing those given by others, such as Steele or Powles [25, Conclusions, p.517] and the many others since that time.

4.2. Self-diffusion coefficients

In these simulations, the mean lifetime of the molecules varies broadly in the region of 24,000 to 2400 time steps as the corresponding temperature varies from $T = 4.0$ to $T = 8.0$. The accurate determination of the three-dimensional (3-D) self-diffusion coefficient D_s for any particle requires the determination of the integral of the long time limit of the velocity autocorrelation function, or the equivalent Einstein expression of the mean square displacement at infinite time with respective forms

$$D_s = \frac{1}{3}\int_0^\infty dt \, \langle \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \rangle \tag{23}$$

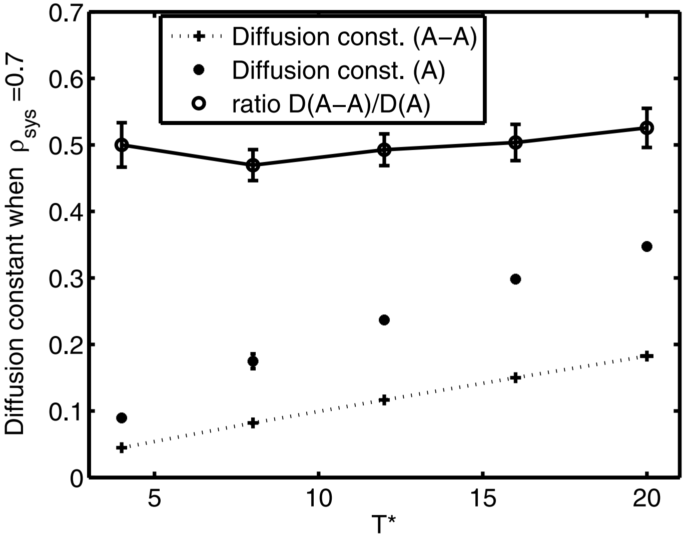


FIGURE 19. Self-diffusion coefficients at varying temperature fixed $\rho = 0.7$, where A-A refers to the dimer and A to the atom, and D denotes the self-diffusion coefficient.

and

$$2tD_s = \frac{1}{3}\left\langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \right\rangle \; (t \to \infty). \tag{24}$$

We overcome the infinite time problem here by determining the diffusion coefficient according to (24) at the time of breakup $t_{br,i}$ of molecule i (where the time is 0 when the molecule is formed), thus allowing for the maximum time possible before $D_{s,m,i}$ is computed (where m refers to the dimer.) Likewise, we can monitor the time spent as a free particle of any labeled atomic species (j), and determine the self-diffusion coefficient $D_{s,a,j}$ (where a refers to the atomic state). The molecular self-diffusion coefficient is the average of all molecules determined during the dump interval, and lastly the 100 dump values for the entire run is averaged to provide an estimate of uncertainty. Similarly, a labeled particle is used to determine the atomic diffusion coefficient based on the time spent as a free – non-bonded particle. The results for this supercritical fluid are given in Figs. 19–20. The curves in Fig. 19 appear very linear, verifying the formula $D_s = BkT$, according to previously developed theories [26, eq. (49)] especially at lower temperatures. In non-reactive systems with spherical particles, the Stokes–Einstein law for diffusion of species i in a liquid j of viscosity η_j is $D_{i,j} \approx \frac{kT}{6\pi\eta_j r_i}$. The viscosity is independent of density, and so very approximately, one might be able to interpret the reaction as one species A-A moving within the matrix of the other atomic species A where the Stoke's law for the force acting on each of the species is viscosity dependent, and obtains for both. Under this assumption and approximation, since the viscosity is independent of temperature, a first-order linear relationship with the temperature

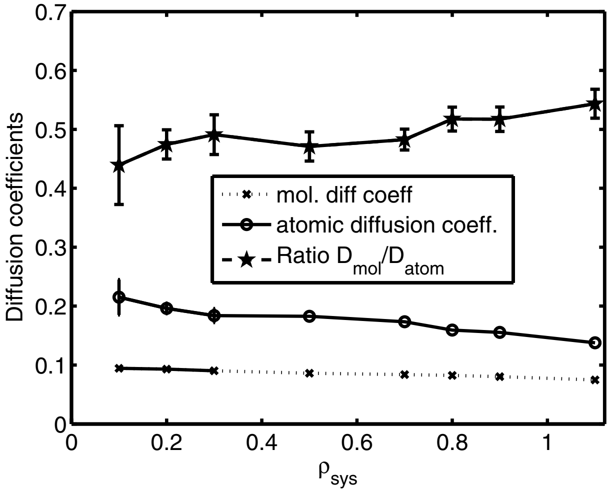


FIGURE 20. Diffusion coefficients at varying ρ and fixed $T^*_{set} = 8.0$.

is predicted for both species, as observed. Furthermore, this result is rather normal experimentally [16, p. 494, Fig. 16.7] for fluids (e.g., Ar(g) or H_2O(l)). The ratio of molecular to atomic diffusion constant is relatively close to 0.50 everywhere. The mass of the molecule is twice that of the atom and approximately twice the diameter, leading to this approximate ratio.The actual theoretical prediction due to size, energy interaction and mass effects is not well developed, and no extensive data are available for even non-reacting systems. The reactive system here depicts values of the diffusion coefficient which do not differ significantly for systems which do not react. In one study [27, p.2044 Table V] of solute diffusion in a solvent, where interactions are solvent-solvent (1-1) and solvent-solute (1-2) only, (i.e., no (2-2) interactions) the L_2 system has the following Lennard–Jones parameters $\frac{m_2}{m_1} = 2; \frac{\epsilon_{22}}{\epsilon_{11}} = 4; \frac{\sigma_{22}}{\sigma_{11}} = 2$ leading to the diffusion coefficients $D_1 = 0.063$ and $D_2 = 0.017$ (accuracy not specified) and for the S_2 system, the Lennard–Jones parameters $\frac{m_2}{m_1} = \frac{1}{2}; \frac{\epsilon_{22}}{\epsilon_{11}} = \frac{1}{4}; \frac{\sigma_{22}}{\sigma_{11}} = \frac{1}{2}$ leading to the diffusion coefficients $D_1 = 0.082$ and $D_2 = 0.190$. For the same mass ratio, the diffusion constant ratios vary from 0.27 to 0.43 for very different and extreme ($\epsilon\ \sigma$) combinations where the variation with temperature is not significant for these ratios based on the scanty information of the graphs drawn; however, for the work of this paper, $\epsilon = 1$, and $\sigma = 1$ throughout. The ratios from the above literature are not too different from the ones reported here. The variation of the diffusion constant with density is much less dramatic than for the temperature according to Fig. 20 with a slight decline in diffusion constants with increasing density, as is to be expected since the mobility would decrease. The errors appear large because the variation of the coefficients with varying density is relatively slight for fixed temperature unlike that for the variation with temperature. Resorting again to the Stokes–Einstein equation with the presuppositions above, we would expect at constant temperature for there to be no change; the variation is due to the fact that the two fluids do not approximate as two fluids where one fluid serves as a solvent for the other. For hard spheres, the self-diffusion coefficient varies inversely with the density of the gas. If this is another effect, combining this with the Stokes-Einstein expression and its assumptions above would lead to the prediction of a weak inverse dependence of the diffusion coefficient with density, which is precisely what is observed. Fig. 19 yields the dimer (A-A) diffusion coefficient as ≈ 0.09 and the atomic (A) self-diffusion coefficient as ≈ 0.18, which is very close to the mean value found in the graph of Fig. 20. It appears that first order kinetic theory, and the Maxwellian prediction of invariance of viscosity with density (pressure) is verified in these results.

4.3. Kinetic energy probability histogram

The potentials described in eqs. (2–7) have the form

$$H = \sum\nolimits_{i=1}^{m} p_i^2/2m + \sum\nolimits_{i<j} V(r_i - r_j) \tag{25}$$

together with switches operating to determine the type of potential operating. This difference might conceivably alter the Gibbs postulate in the following manner. One of the postulates states that the time average of a particular system equals the ensemble average. The density-in-phase ρ corresponding to the probability of a particular state is given by $\rho \propto \exp -\beta H$ and so the kinetic energy of any particle would be Boltzmannized for any particular system and therefore, if a labeled particle is monitored throughout the whole simulation, from the time it is bonded and when it is not, then the above postulate demands a Boltzmannized kinetic energy distribution. The Gibbs postulate can be directly tested for the chemical reaction system to verify whether or not the switching mechanism modifies or contradicts the Gibbs postulate. Experimentally, (Fig. 21), it is found that switches that lead to non-single-valued Hamiltonians *do not* affect the Gibbs postulate. If this postulate is valid for loop-like hysteresis systems, then the time trajectory of any indexed particle I must also yield, when averaged over a very long time, the result (in 3-D) $3kT/2 = \overline{p_I^2/(2m_I)}$ whether the particle is bonded or not over the trajectory equally weighted for all the states that it traverses. The Maxwellian probability density function per unit energy increment is given by

$$P = 2\pi \left(\frac{1}{\pi}kT\right)^{3/2} \epsilon^{1/2} \exp -(\frac{\epsilon}{kT}). \tag{26}$$

Eq. (26) is the standard form used for the absolute velocity distribution function since the energy $\epsilon \propto v^2$ for velocity v and this form tests for the Boltzmann distribution for kinetic energies. Noting that the accuracy of the single particle is reduced by a factor of ≈ 4000 (the number of particles in this simulation), we find that the Gibbs postulate seems to be verified, in terms of the shape of the P function (which appears Maxwellian) as well as the computed value of the

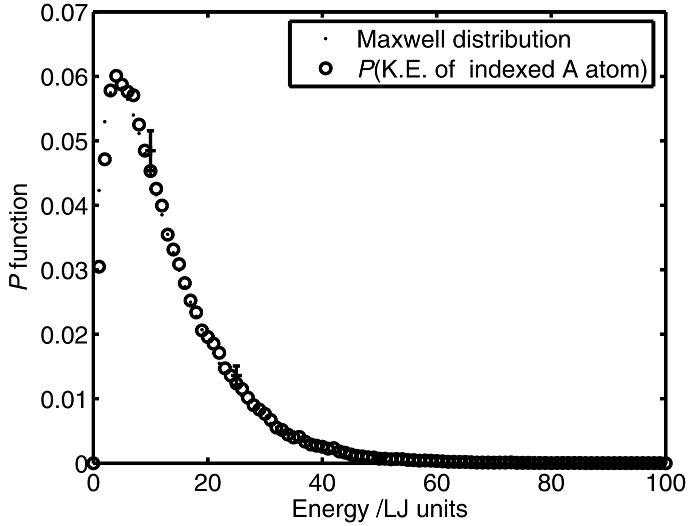


FIGURE 21. P functions for kinetic energy of fixed indexed atom A which either is bonded to some A_2 dimer or not at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$ with apparent temperature of atom $< T >_{atom} = 8.1 \pm 0.2$.The uncertainty here is 3 standard error units.

temperature with the error estimated as ±0.1, by studying an atom of fixed label as it forms and breaks bonds with neighboring molecules, as shown in Fig. 21. Clearly the time average of dynamical properties for this particle would equal the ensemble average. We notice that the reduced accuracy of the sampling is reflected in the greater scatter of the P function points.

5. Conclusion

The main thrust of this work is to depict a modeling method using exclusively two-body potentials that might be useful for thermodynamical simulation. In the course of the demonstration, the study shows that the model of the molecule utilizing switching potentials does lead to typical behavior predicted from standard thermodynamics even for these unusual hysteresis-type reaction mechanisms which theorists have largely ignored, due perhaps to the influence of 'time-reversible' symmetry concepts. It is demonstrated that microscopic loop-like pathways do not influence the macroscopic thermodynamical results in any fundamental way. In particular, the Gibbs ensemble postulate is obeyed, implying that the thermodynamics is well-behaved.

The method used here to reduce expensive 3-body calculations to easier 2-body calculations may be used as a basis for non-equilibrium simulation applications, which will be the subject of further investigations. The two-body potentials yield extremely good thermodynamic results, whilst being super-efficient in reducing computational costs, because the use of switches and algorithms that can preserve momentum and energy during potential transitions, and it is expected that semi-quantitative results at least can be determined for any known molecular potential. The NEWAL algorithm is effective for the extreme conditions of the simulation, and would prove to be a valuable tool in reducing errors attributable to switching potentials and high velocities and energies. This reduction in error would be even more evident at more 'normal' conditions with the temperature parameter scaled $10 - 20$ times less than those used here.

A whole generation of scientific literature has been devoted to establishing necessary connections between the direction of material flow (microscopic reversibility or 'time reversibility') and thermodynamics, but the results here suggest that there need not be any necessary connection between the two. In other words, although the point of breakdown of the molecule is different from the point of formation,which may be considered irreversible in terms of path, causing a hysteresis loop in the potential (which is a common phenomena in science, e.g., the various ferromagnetic and ferroelectric hysteresis curves in solid state theory), yet we observe *no unusual thermochemistry* for the macroscopic properties that result from the simulation; thus the two concepts can be decoupled to some extent because thermodynamics relates to an averaging process, whereas in pure dynamics, the temperature parameter does not have a role. (I am excluding the synthetic Hamiltonians used in simulations of constant temperature systems without explicit use of a perturbing thermal reservoir that introduces indeterminacy to the system by the introduction of random forces.) Further, pure dynamical motion is not the result of any averaging procedure where classical theory is concerned. In other words, the degree of decoupling of the above two concepts can be decided upon on the basis of the conjunctions of the following propositions being true due to the results presented here:

A.

Non-reversible *mechanical dynamical pathway* $\wedge$
Simulation of system at zero net flux steady state
with defined temperature $\Rightarrow$
Existence of thermodynamical equilibrium.

On the other hand, the conventional assumptions may be stated thus:

B.

Non-reversible *mechanical dynamical pathway* $\wedge$
Simulation of system at zero net flux steady state
with defined temperature $\Rightarrow$
Non-existence *of thermodynamical equilibrium.*

The problems associated in giving a precise characterization of A above and its elucidation for thermodynamical equilibrium states may well prove to be a worthwhile challenge. It would be of interest to repeat and compare some of the above calculations for a conventional system without hysteresis to rule out any necessary connection between dynamics and equilibrium thermodynamic properties.

Acknowledgment

C.G.J would like to thank (a) University of Malaya, Kuala Lumpur for financing a sabbatical visit to NTNU (2000-2001), and (b) availability of Intensification of Research in Priority Areas (I.R.P.A) grant no. 09-02-03-1031 from the Malaysian Government which was used to finance the computer system for the project and for conference and research visits to Europe and America where portions of this work were discussed [28].

References

[1] D. N. Hendrickson, Single-molecule magnets. In *Abstracts of Papers, 225th ACS National Meeting 2003*;American Chemical Society:Washington D.C., 2003.

[2] D. Gatteschi, From molecular magnets to magnetic molecules. *Actual. Chimique. 6* (**2001**), 21-26.

[3] E. Sanudo, E. Carolina, W. Wernsdorfer, K. A. Abboud, G. Christou, Synthesis, structure, and magnetic properties of a Mn_{21} single-molecule magnet. *Inorg. Chem.*, *43(14)* (**2004**), 4137-4144.

[4] C. G. Jesudason, I. Time's Arrow, detail balance, Onsager reciprocity and mechanical reversibility. Basic Considerations, *Apeiron* (**1999**), *6(1-2)*, 9-24.

[5] C. G. Jesudason, II. Time's Arrow, detail balance, Onsager reciprocity and mechanical reversibility. Thermodynamical Illustrations. *Apeiron 6(1-2)*, (**1999**), 172-185.

[6] M. P. Allen, P. Schofield, Molecular dynamics simulation of a chemical reaction in solution. *Mol Phys.*, *39(1)*, (**1980**), 207-215.

[7] Y. Zeiri, E. S. Hood, Nonequilibrium Distributions in Reactive Systems. *Phys. Rev. Letts.*, *55(6)*, (**1985**), 634-637.

[8] J. Gorecki, J. Gryko, Molecular Dynamics simulation Of A Chemical Reaction. *Comput. Phys. Commun. 54*, (**1989**), 245-249.

[9] F. H. Stillinger, T. A. Weber, Molecular dynamics simulation for a chemically reactive substances.Fluorine. *J. Chem. Phys. 88(8)*, (**1988**), 5123-5133.

[10] I. Benjamin, B. J. Gertner, N. J. Tang, K. R. Wilson, Energy Flow in an Atom Exchange Chemical Reaction in Solution. *J. Am. Chem. Soc. 112*, (**1990**), 524-530.

[11] J. P. Bergsma, J. R. Reimers, K. R. Wilson, J. T. Hynes, Molecular dynamics of the A+BC reaction in a rare gas solution. *J. Chem. Phys.*, *85(10)*, (**1986**), 5625-5643.

[12] B. Hafskjold, T. Ikeshoji, Partial specific quantities computed by nonequilibrium molecular dynamics. *Fluid Phase Equilibr 104*, (**1995**), 173-184.

[13] C. G. Jesudason, The Clausius inequality: Implications for non-equilibrium thermodynamic steady states with NEMD corroboration. *Nonlinear Analysis*, *63(5-7)* (**2005**) e541-e553.

[14] R. D. Levine, R. B. Bernstein, *Molecular Reaction Dynamics and Chemical Reactivity*; Oxford University Press: Oxford, 1987 ; esp. pp. 375-376 and Fig. 6.60.

[15] T. Ikeshoji, B. Hafskjold, Non-equilibrium molecular dynamics calculation of heat conduction in liquid and through liquid gas interface.*Mol. Phys*, *81(2)*, (**1994**), 251-261.

[16] I. N. Levine, *Physical Chemistry.* (5th Edition) , McGraw-Hill (Singapore) 2003.

[17] C. G. Jesudason, An energy interconversion principle applied in reaction dynamics for the determination of equilibrium standard states.*J. Math. Chem.* (JOMC) 39(1) (**2006**) 201-230).

[18] K. J. Laidler, *Chemical Kinetics.* (Third Edition),Harper & Row, New York 1987 esp. pp.74-75.

[19] P. Debye, *Polar Molecules*-1929 reprint edition, Dover Publications Inc: New York, 1929.

[20] C. T. O'Konski, A. J. Haltner, Characterization of the monomer and dimer of tobacco mosaic virus by transient elastic birefringence relaxation of optically anisotropic crystals.*J. Am. Chem. Soc. 78*,(**1956**), 3604-3610.

[21] S. Broersma, Rotational Diffusion Constant of a Cylindrical Particle. *J. Chem. Phys. 32(6)*, (**1960**), 1626-1631.

[22] R. Moog, D. Bankert, M. Maroncelli, Rotational diffusion of Coumarin 102 in Trifluoroethanol:the case for solvent attachment.*J. Phys. Chem. 97*, (**1993**), 1496-1501

[23] A. Srivastava, S. Doraiswamy, Rotational diffusion of Rose Bengal. *J. Chem. Phys.,103(14)*, (**1995**), 6197-6205.

[24] R. K. Pathria, *Statistical Mechanics (2/e)*; Butterworth-Heinemann: Oxford, 2001.

[25] R. M. Lynden-Bell, Comparison of the results from simulations with the predictions of models for molecular reorientation. In A.J. Barnes; W.J. Orville Thomas; J. Yarwood, eds.,*Molecular Liquids - Dynamics and Interactions*, D. Reidel Publishing Co. , Dordrecht/Boston/Lancaster, 1984, pp. 501-518.

[26] D. Levesque, L. Verlet, Computer "Experiments" on Classical Fluids. III. Time Dependent Self-correlation Functions. *Phys. Rev. A 2(6)*, (**1970**), 2514-2528.

[27] K. Nakanishi, K. Toukubo, N. Watanabe, Molecular dynamics studies of Lennard-Jones liquid mixtures. Further calculation on the behavior of one different particle as a model of real fluid systems.,*J. Chem. Phys. 68(5)*, (**1978**), 2041-2045.

[28] C. G Jesudason, examples include W.C.N.A 2004 (Orlando, Florida, U.S.A,2004) and in presentation entitled Equilibrium properties of a hysteresis dimer molecule from MD simulations using two-body potentials. In T.E. Simos; G. Psihoyios; Ch. Tsitouras, ed., *International Conference on Numerical Analysis and Applied Mathematics 2005*, Wiley-VCH, Weinheim, 2005, pp.287-290

Christopher G. Jesudason
Chemistry Department, University of Malaya
50603 Kuala Lumpur, Peninsula Malaysia
e-mail: `jesu@um.edu.my`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 143–160

Model Hysteresis Dimer Molecule. II. Deductions from Probability Profiles

Christopher G. Jesudason

Abstract. The hysteresis dimer reaction of Part I is applied to test the Gibbs density-in-phase hypothesis for a canonical distribution at equilibrium. The probability distribution of variously defined internal and external variables is probed using the algorithms described, in particular the novel probing of the energy states of a labeled particle where it is found that there is compliance with the Gibbs' hypothesis for the stated equilibrium condition and where the probability data strongly suggests that an extended equipartition principle may be formulated for some specific molecular coordinates. The possible ambiguity of internal variables as described in mesoscopic nonequilibrium thermodynamics (MNET) is very briefly discussed in relation to Hamiltonian variables, and a canonical distribution for a certain class of internal variables is observed and described, and plausible reasons outlined, where it is found that the always free dimer and atom particle kinetic energy distributions agree fully with Maxwell–Boltzmann statistics but the distribution for the relative kinetic energy of bonded atoms does not, even when all of these coordinates are not canonical variables. The principle of local equilibrium (PLE) commonly used in nonequilibrium theories to model irreversible systems is investigated through NEMD simulation at extreme conditions of bond formation and breakup at the reservoir ends in the presence of a temperature gradient, where for this study a simple and novel difference equation algorithm to test the divergence theorem for mass conservation is utilized, where mass is found to be conserved from the algorithm in the presence of flux currents, in contradiction to at least one aspect of PLE in the linear domain. It is concluded therefore that this principle can be a good approximation at best, corroborating previous purely theoretical results derived from the generalized Clausius Inequality which proved that the PLE cannot be an exact principle for nonequilibrium systems.

Mathematics Subject Classification (2000). 65-{04,Z05}, 68-{04,W01} 70-{08,F01,F16}.

Keywords. Kinetic energy probability profile, Gibbs ensemble hypotheses, Extended equipartition principle, NEMD, Principle of local equilibrium, Clausius inequality.

1. Introduction

The previous chapter provided details of the method and characteristics of the dimer reaction system

$$2\mathrm{A} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{A}_2 \tag{1}$$

where k_1 and k_{-1} are the respective rate constants. The methods used to ensure accuracy and convergence of results for the time steps used were also discussed in that chapter. In this work, the system is probed for probability distribution functions, and an NEMD simulation of the dimer system is also carried out to study the applicability of one aspect of PLE. The details of the computations, together with the precautions used to ensure reproducibility of results are discussed in [1]. Indeed, for the NEMD portion, the verification here of conservation of mass can only imply convergence of the system to a steady state. Comparisons between the theoretical Maxwell distribution to that derived from equilibrium simulations is carried out in Sect. 2 because fundamental deductions can be made concerning the theory and applications of the canonical distribution. Additional results are presented in Sect. 3 from NEMD using a novel difference equation which can be used to check for conservation of matter. Here, it is found that current fluxes exist in regions when this would not be expected according to one aspect of PLE. The NEMD runs were used to ascertain whether PLE is indeed a principle or merely a good approximation for describing general thermodynamical systems (whether reversible or not). It is concluded that simulations provide examples that go beyond linear and local equilibrium theories.

2. Probability histograms

These are provided in Figs. 1–7 for the translational kinetic energies of the different species probed as well as the total internal energy of the dimer, plotted with the Maxwell distribution relative to the apparent temperature determined from (4) below. The comparisons provide clues to the following:

- Shape of the probability function P, which could perhaps be used to determine whether the assumptions used in theories are reasonable or not. The shape even for this equilibrium system is not always Gaussian, and so there is no reason to assume *a priori* that nonequilibrium systems must conform to a Gaussian distribution where certain internal variables are concerned.
- A rationale for extending the theory of equipartition in an equilibrium system where the temperature relative to a particular kinetic energy coordinate is not the same as for the total system temperature determined from standard equipartition. Such a possibility seems to be supported by the evidence below.

The method of determining these probability histograms involves sampling at each time step the respective quantities, binning the values of the particular distribution, followed by normalization. Such a method ensures that an accuracy is obtained that is able, for instance, to discriminate between the different apparent species temperatures. For a given Hamiltonian $\mathcal{H}$ weakly coupled to a heat bath, written

$$\mathcal{H} = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + \mathcal{V}(r_1, r_2, \ldots, r_n) \tag{2}$$

where $\mathcal{V}$ is the potential that is position variable $\mathbf{r}$ dependent, the probability density function per unit area of phase space $(\mathbf{p}, \mathbf{q})$ is

$$\mathcal{P}(\mathbf{p}, \mathbf{q}) = \frac{\exp -\beta\mathcal{H}}{\mathcal{Z}} \tag{3}$$

where the partition function $\mathcal{Z}$ has the form

$$\mathcal{Z} = \frac{\int e^{-\beta\mathcal{H}}\mathbf{dpdr}}{N!}.$$

The separability of the Hamiltonian above for the momentum $\mathbf{p}$ and position variables $\mathbf{r}$ which is of the same form as our chemical system Hamiltonian (augmented by switches) leads for large N to the exact result (in 3-dimensional systems) (usual laboratory units)

$$N\left(\frac{3kT}{2}\right) = \overline{\sum_i p_i^2/(2m_i)} \tag{4}$$

which is the method used to determine the system temperature here. The momentum coordinates p_i refer to all atomic species, whether bonded or not. The Gibbs postulate can be directly tested for the chemical reaction system to verify whether or not the switching mechanism modifies or contradicts the postulate, which refers to the time average of a system property being equal to the ensemble average when these limits exist. Experimentally, (Fig. 6), it is found that switches in non-single-valued Hamiltonians *does not* affect the Gibbs postulate. Furthermore, over the time of the simulation, for the indexed particle I, the following (3-D) result must hold so that the particle and system temperature is defined:

$$3kT_t/2 = \langle p_I^2/(2m_I)\rangle \tag{5}$$

where the brackets represent the time average. It is found that the temperature T_t above of this single indexed particle coincides with the mean system temperature whether the particle is bonded or not over the trajectory, equally weighted in time for all the states that it traverses. Integrating the $\mathcal{P}$ function in (3)above over all equal energy values, the Maxwellian probability density function results, and is given per unit energy increment by

$$P = 2\pi\left(\frac{1}{\pi}kT\right)^{3/2}\epsilon^{1/2}\exp -(\frac{\epsilon}{kT})\,. \tag{6}$$

Eq. (6) is the standard form used for the absolute velocity distribution function. The above form is still derived from the quantum probability operator/function

$$\hat{P}(\Omega) \propto \exp -\beta\hat{H}(\mathbf{p}, \mathbf{q}) \tag{7}$$

where the phase space is averaged over equi-energy surfaces. Eq.(7) also makes the definition of the partition function Q possible as

$$Q = \sum_i \omega_i \exp -\beta E_i(\mathbf{p}, \mathbf{q})$$

for a system where even for Q, $(\mathbf{p}, \mathbf{q})$ represents the canonical coordinates only. The Gibbsian and other thermodynamical state functions are derived strictly from operations on Q, e.g., $U = kT^2 \left(\frac{\partial \ln Z}{\partial T}\right)_{V,N}$ for the energy and $P = kT \left(\frac{\partial \ln Z}{\partial V}\right)_{T,N}$ for the pressure. Other internal coordinates cannot (unless proved otherwise) give rise to state functions where standard statistical mechanics is concerned.

An apparent temperature parameter $< T >_X$ is computed here for some species X and is defined such that

$$\frac{3 < T >_X}{2} = \left\langle \frac{p_X^2}{2m_X} \right\rangle \tag{8}$$

where m_X is the mass of species X and p_X is its momentum variable. This parameter is clearly not well defined as a temperature if it does not obey the equipartition result above, for the obvious reasons connected to conjugate transforms. In statistical thermodynamics, the total system Hamiltonian

$$H = \sum\nolimits_{i=1}^{m} p_i^2/2m + \sum\nolimits_{i<j} V(r_i - r_j) \tag{9}$$

leads to the density-in-phase having form

$$\rho(\mathbf{p}, \mathbf{q}) \propto \exp[-H(\mathbf{p}, \mathbf{q})/kT] \tag{10}$$

and so for systems with separable coordinates, each kinetic energy coordinate $E_{k,i} = p_i^2/2m$ and potential form $V(|r_i - r_j|)$ will have the above Boltzmann distribution. The $(\mathbf{p}, \mathbf{q})$ coordinates are termed 'canonical' and equipartition and the distribution laws are derived relative to these coordinates only [2]. The development of Statistical Mechanics by Gibbs very clearly relates the density-in-phase probability to these Hamiltonian coordinates only and nothing else [3, Chap. 5, p. 46], so that no other coordinates are mentioned. In particular, the thermodynamical properties are (average) integrals over the coordinates, and therefore no dynamics are to be inferred from this development. Furthermore, the 'Gibbsian' entropy, a strictly equilibrium concept, is more of an after-development, for in the original Gibbsian development, the entropy for the petite and grand ensemble is respectively the average [3, p. 203, eq. (545)] of quantities η and H where H has additive terms in the chemical potential; if these terms represent work, then the form

$$S = -k_B \int d\mathbf{X}^{\mathbf{N}} \rho(\mathbf{X}^{\mathbf{N}}) \ln[C_N \rho(\mathbf{X}^{\mathbf{N}})]$$

results, called the Gibbs entropy [4, p.342, eq. (7.2)]. No exact theory of entropy with the same logical clarity, breadth and application of the classical forms has been produced, and consequently, all forms proposed are assumptions with regard not just to the form of the entropy expression, but also its statistical analog.

And these forms are for strictly equilibrium thermodynamics. Even using rigorous classical principles, it was shown recently that there exists, on topological grounds, another type of entropy called the surface entropy, different from the one developed by the Clausius definition [5]. One form proposed in MNET [6, eq. (7)] is

$$S = S_{eq} - k_B \int P(\gamma, t) \ln \frac{P(\gamma, t)}{P_{eq}(\gamma, t)} d\gamma$$

where S_{eq} is the entropy of the system when the 'nonequilibrated' degrees of freedom γ are those that obtain at equilibrium. This entropy form is fundamental to the entire MNET theory [6, sec.4] Conventional treatments of statistical mechanics of the Gibbsian kind do not refer to the γ coefficients, for these are integrated out (such as the p and q Hamiltonian variables, as in the Gibbs entropy arising from the average of the H and η function). One other curiosity, from the conventional point of view of equilibrium statistical mechanics is that it is the probability distribution function which determines the other Gibbsian potentials, such as the chemical potential, so that variation of the probability distribution function would also vary the potential, unless it is *assumed* that to first order, such a potential is not varied and that the variation is significant only for the probability distribution, so that in general [6, eq. (18)] one writes

$$\delta S = -\frac{1}{T} \int \mu(\gamma) \delta P(\gamma, t) d\gamma.$$

The above form should be contrasted to non-additive entropies that have found widespread application in recent times, such as the Tsallis non-normalized entropy [7] where uniqueness is proven relative to some axioms in conditional probability. Here the entropy is postulated to have the form

$$S_q(p) = \frac{1}{q-1} \left(1 - \int p^q(x) dx \right)$$

where p is the probability distribution, and q a real parameter, with recovery of the Gibbs form when $q \to 1$. The debate concerning its general validity still rages, and widespread applications have been attempted for this non-extensive entropy. This entropy and the MNET one do not have much resemblance at first sight; perhaps they belong to different regimes of applicability, but interpretations that the Tsallis entropy is relevant to nonequilibrium theory abound. The thrust of this work is such as to neither support nor contradict MNET, but to give an example of 'internal coordinates' from an exact point of view with reference to the process Hamiltonian which precisely determines the trajectory of the simulation between times when the thermal reservoir is not interfering with the motion. But the definition of internal and canonical variables is at all times uniquely defined. However, the 'internal coordinates' during a chemical reaction or other process refer to an artificial aggregation — meaning they are transient species — such as the center of mass (C.M.) velocity and position for particles k, l forming a molecule which is not permanent, e.g., $\mathbf{P}_j = \mathbf{p}_k + \mathbf{p}_l$ $(k \neq l)$, $\mathbf{R}_j = \frac{1}{m_k + m_l} (\mathbf{r}_k + \mathbf{r}_l)$, and so these are not canonical coordinates in the defined sense [2] and there is no immediate reason *a priori* that these coordinates for the internal energy or potential must have Boltzmannized distributions. It could well be that if the mean lifetime τ of the species obeyed $\tau \to \infty$, then they might qualify as a pseudo-canonical variable, but a theory for such limits does not seem available. Permanent aggregated states can be expressed in terms of canonical transformations $\mathbf{Q} = \mathbf{Q}(\mathbf{p}, \mathbf{q})$, $\mathbf{P} = (\mathbf{p}, \mathbf{q})$ [2, Chap. VII] and the new Hamiltonian that results must by ensemble theory be subjected to the density distribution described above. But for systems which are described by 'internal' coordinates of a non-permanent nature (in the sense that the forces between the particles cease when the molecule decomposes) and which does not refer to the system Hamiltonian, no general theory exists, and no presuppositions can be made regarding their density distributions. It may be remarked that, in statistical mechanics for canonical distributions, average quantities $\overline{M}$ (corresponding to classical thermodynamical state functions) are defined as $\overline{M} = \sum_{i=1}^{N} M_i P_i$ where P_i is the probability of state i with value M_i. The partition function $Q = \sum_{i=1}^{N} g_i \exp -\frac{\epsilon_i(\mathbf{p},\mathbf{q})}{kT}$ for the system has been defined so that operations on it, $\hat{O}_M[Q]$, yield the average value for property M [8, p. 422], e.g.,

$$\hat{O}_E[Q] = NkT^2 \frac{d \ln Q}{dT} = \overline{E} \tag{11}$$

yields the total energy due to translational kinetic energy for systems that conform to the canonical probability law. It follows that the density-in-phase probabilities are correlated to the $(\mathbf{p}, \mathbf{q})$ phase space volume elements, and that the canonical $(\mathbf{p}, \mathbf{q})$ coordinates or their equivalents are central to the above procedure. Clearly, when the process defined by the coordinates is not canonical, then it is not in general correct to insist *by necessity* that any coordinate combination is 'self-similar' to a canonical coordinate set, with a canonical probability distribution. Nevertheless, theories have been created that *assume* in certain situations that the (Gaussian) density for internal variables is true [9, 10] without clear qualification concerning the situation when this condition obtains in terms of the Hamiltonian of the system. How does one specify so-called 'nonequilibrated degrees of freedom' γ variables which still exist for equilibrium systems [6, p.21504, sec.3], and what is the relation of these variables to the Hamiltonian variables? In equilibrium statistical thermodynamics, the Gibbs energy and all other thermodynamical state functions are derived from the partition function through averaging with the canonical distribution, which pertains to the entire system taken as a whole; to infer that each microscopic portion of the system at a particular phase-space coordinate is necessarily self-similar to the whole is incorrect, as a few counter-examples to this proposition show below (Figs. 2–4). Furthermore, the PLE has been proposed as useful [9] for these new theories, and another counter-example to this is also provided, this time from a NEMD simulation. In other words, basic simulation is able to refine the assumptions used for theories, and in particular, the hysteresis system described here refers to internal coordinates and variables which are not of the same variety as those that pertain to such 'mesoscopic' level thermodynamics.

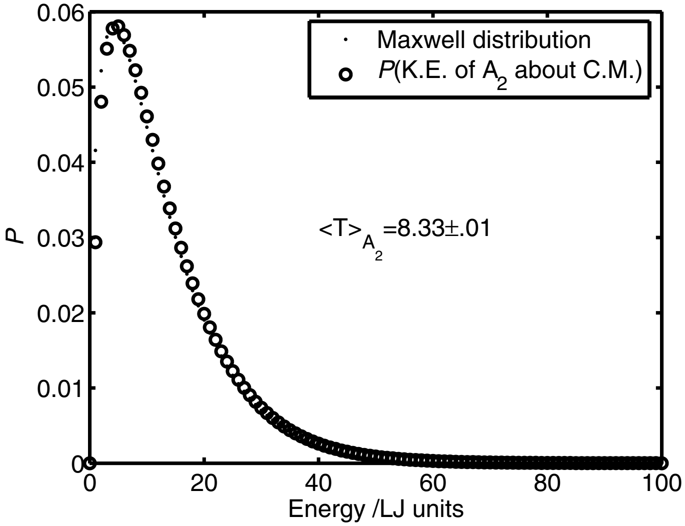


FIGURE 1. P functions for translational kinetic energy of A_2 about center of mass at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$, with apparent temperature of molecule indicated.

The total internal energy coordinate (TIEC) and the internal kinetic energy coordinate (IKE) are not Gaussian distributions for equilibrium systems according to the simulation results below. Of great theoretical interest is that, for cases of nonpermanent coordinates, some types of distributions are essentially Bolztmannized, others are not, even for an equilibrium system. It would be of great significance and interest to provide criteria which can predict when a Boltzmann distribution can be expected. The apparent temperature parameter $< T >_X$ may well qualify as a temperature in an extended equipartition scheme if there is agreement with the Maxwellian distribution *even if this temperature does not correspond to the unique system temperature* $< T >_{sys}$. Here the degree of agreement with the Maxwell distribution is either very good (in some cases), or rather bad. It would be of great theoretical interest if some form of relationship between the apparent temperatures could be made on the basis of internal energetics. The uncertainty (unless stated otherwise) is of the order as given in the error bars of Fig. 5 which is at 100 standard error units and which would not feature in any figure where errors are typically quoted at 3 standard error units. This figure corresponds to the TIEC distribution. The errors in the temperature are given in Figs. 1–7. Fig. 1 shows that the center of mass (C.M.) kinetic energy follows quite accurately a Maxwellian P function with a temperature parameter *higher* ($T^* = 8.33$) than the system temperature ($T^*_{set} = 8.0$). The fact that the shape is Maxwellian at the indicated temperature parameter does seem to imply that theories may be developed *within* an equilibrium system with different coexisting temperatures, provided that these parameters require that a Maxwellian form regarding shape

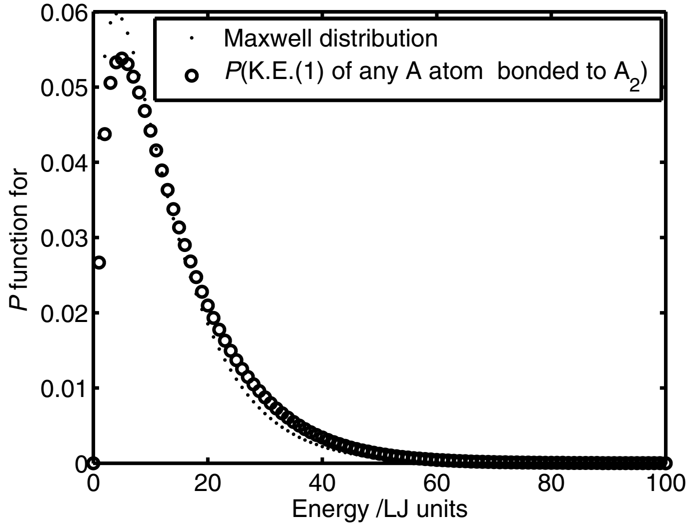


FIGURE 2. P functions for kinetic energy of any atom A bonded to A_2 at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$. The apparent non-Boltzmann kinetic energy temperature of these bonded atoms is $< T >_{kin} = 8.10 \pm .01$.

prevails; after that first stage one perhaps might also be able to propose generalizations to temperature not requiring a Maxwellian distribution. But a proper theory would have to begin from first principles which can subsume without contradiction the previous axiomatics, including the Zeroth Law. Another inference is that these non-standard 'temperatures' have definite values (or limits), where the degree of scattering is relatively low; hence one might expect some type of stochastic averaging which yields exact values (limits). How these averages are performed, and the theoretical justification for these averages, remain significant challenges. The other important scientific question is the explanation of the shift of 'temperature' $< T >_X$ for such Boltzmann distributions for non-permanent aggregates.

An atom bonded to a molecule does not have a clear Maxwellian shape, as is evident from Figs. 2–3 since there is interference from the internuclear potentials. The graph in Fig. 2 computes the absolute kinetic energy AKE (also denoted K.E.(1))of the particle with respect to the MD cell, whereas Fig. 3 refers to half the relative kinetic energy and half the translational kinetic energy about the C.M. of the bonded pair, where the relative kinetic energy $\epsilon_{k.e.rel.}$ is written as

$$\epsilon_{k.e.rel.} = \frac{1}{2}\mu(\dot{\mathbf{r}}_1 - \dot{\mathbf{r}}_2)^2 = \frac{1}{2}\mu\dot{\mathbf{r}}^2 \tag{12}$$

for any two bonded atoms 1 and 2, where the reduced mass μ is given as $\frac{1}{\mu} = \frac{1}{m_1} + \frac{1}{m_2}$ and where the intermolecular axis vector is $\mathbf{r} = \mathbf{r_1} - \mathbf{r_2}$. The total internal kinetic energy IKE is also defined as the relative kinetic energy of a bonded pair,

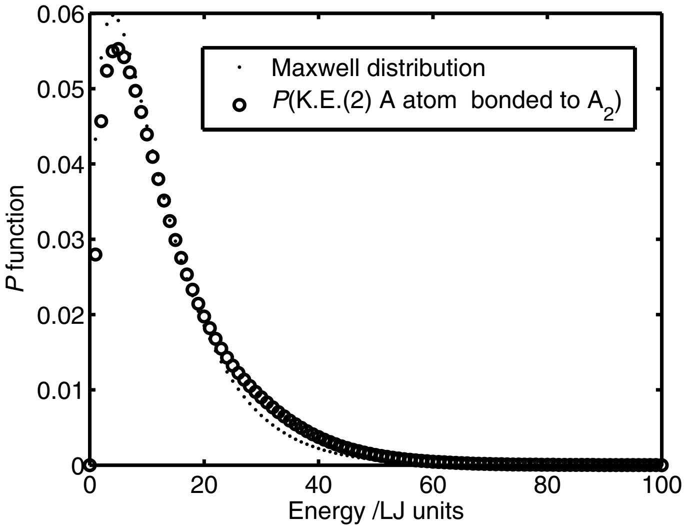


FIGURE 3. P functions for average kinetic energy of atom A by K.E.(2) method at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$ with total system temperature indicated where apparent non-Boltzmann kinetic energy temperature of these bonded atoms is $< T >_{kin} = 8.10 \pm .01$.

given as $\epsilon_{k.e.rel.}$ as above. The AKE averages

$$\frac{1}{2}.\frac{1}{2}(\mathbf{v_1} - \mathbf{v_2})^2 = \frac{1}{4}\left\{\mathbf{v_1}^2 + \mathbf{v_2}^2 - \mathbf{v_1}.\mathbf{v_2}\right\}$$

whereas the kinetic energy about the C.M. (KCM) averages the expression

$$\frac{1}{2}.2\frac{(\mathbf{v_1} + \mathbf{v_2})^2}{2^2} = \frac{1}{4}\left\{\mathbf{v_1}^2 + \mathbf{v_2}^2 + \mathbf{v_1}.\mathbf{v_2}\right\}.$$

Adding these expressions and then dividing by 2 would lead to convergence of the result to that for AKE, which is what is presented in Fig. 3 as K.E.(2), which is almost the same graph as for Fig. 2. The reason for this computation was to check for consistency of result for the two different sampling techniques.

The IKE distribution, that of an internal coordinate, is clearly non-Gaussian, as depicted in Fig. 4. This result is not consistent with any theory which assumes that motion along these coordinates is in local equilibrium with Gibbsian-like equations (which would demand a canonical energy weightage from elementary statistical mechanics) along the entire trajectory, as has been suggested in some applications of nonequilibrium thermodynamics. [9, 10].

TIEC defined above refers essentially to the vibrational and rotational kinetic energy of the molecule E_{tiec}, since the translational kinetic energy about the C.M. has been factored away where

$$E_{tiec} = V(|\mathbf{r_i} - \mathbf{r_j}|) + \frac{\mu\dot{\mathbf{r}}^2}{2} \qquad (13)$$

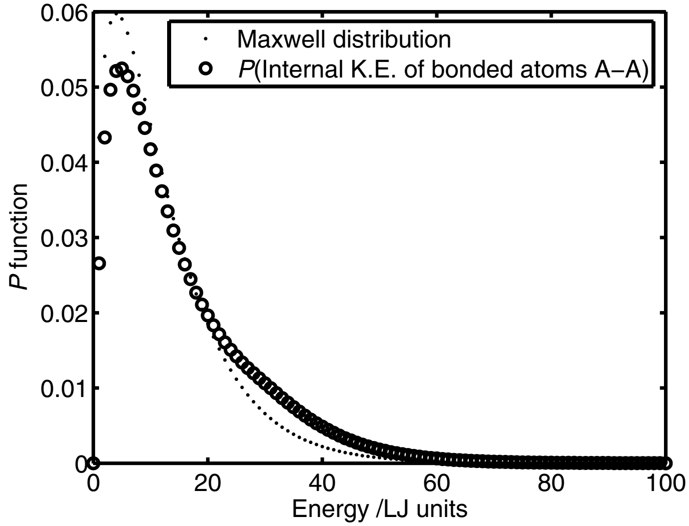


FIGURE 4. P functions for total internal kinetic energy (IKE) of the two bonded A atoms about the internuclear axis at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$.

with $V(|\mathbf{r_i}-\mathbf{r_j}|) = u_0+\frac{1}{2}k(r-r_0)^2$. Hence the intermolecular potential would play an important part in determining the motion along the internuclear axis, with the environmental potential due to other particles playing a moderating role by introducing stochasticity to an otherwise plainly mechanical system. The probability of occurrence of a state is proportional to the time spent at any configuration according to Gibbs, and with a harmonic potential, most of the time spent will be at the turning points in simple harmonic motion (SHM). In the molecular potential used there is a 'dissociation hump' just prior to the dissociation limit, leading to a departure from the Maxwell (M) distribution; other reasons for departure from the M distribution include the dissociation itself, precluding higher energy states from being accessed. It is clear that the distribution in Fig. 5 is non-Maxwellian and corresponds faintly with the shape of the molecular potential energy function, with its humped potential near the distance of dissociation. SHM in conjunction with permanent canonical coordinates has been used as a classic description of equipartition. If the particles were bonded permanently, this quantity would have a canonical distribution, which it clearly does not, because bonds are formed and broken at a rate that precludes adjustment to a Gaussian probability factor. This distribution, which also refers to an internal coordinate for total internal molecular energy, is not consistent with some recent nonequilibrium theories [9, 10] which assumes without proof that these Gaussian factors must obtain.

Noting that the accuracy of the single particle is reduced by a factor of ≈ 4000 (the number of particles in this simulation), we find that the Gibbs postulate seems

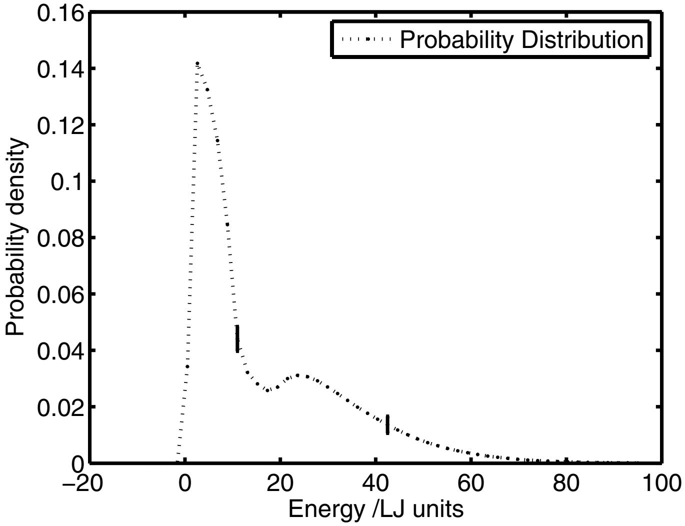


FIGURE 5. The total internal energy coordinate (TIEC) distribution as given in the text. The error bars are for 100 standard error units at $T^*_{set} = 8.0$ and $\rho = 0.7$.

to be verified in terms of the shape of the P function (which appears Maxwellian) as well as the computed value of the temperature with the error estimated as ±0.2 by studying an atom of fixed label (no. 29) as it forms and breaks bonds with neighboring molecules, as shown in Fig. 6. Clearly the time average of dynamical properties for this particle would equal the ensemble average. We notice that the reduced accuracy of the sampling is reflected in the greater scatter of the P function points. The time averaged particle temperature corresponds within error to the system temperature.

Finally, since the molecular P function has been mentioned, it would be interesting to compare it to the case of a random, but always free A particle which is given in Fig. 7, where the determined temperature is slightly *lower*, (to within the error limits) than the system temperature, and where the shape of the P curve is Maxwellian. This particular species type cannot fulfill the Gibbs postulate because its trajectory is confined to those areas where there is no molecular formation, and so its time averaged properties, like the temperature, need not necessarily equal that for the system as a whole as determined from the equipartition principle. We can conclude that the energy subsystems that can be chosen for devising a theory of unequal temperature distributions in an equilibrium system all of which have a Maxwellian probability profile include at least the following candidates:

- Translational k.e. about C.M. for A_2,
- Fixed indexed k.e. of particle A (in both free and bonded state),
- Random, always unbonded k.e. of particle A.

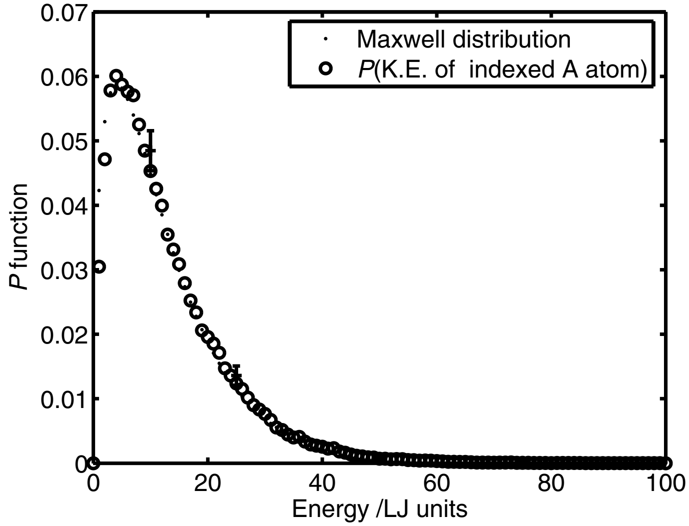


FIGURE 6. P functions for kinetic energy of fixed indexed atom A which either is bonded to some A_2 dimer or not at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$ with apparent temperature of atom $< T >_{atom} = 8.1 \pm 0.2$. The uncertainty here is 3 standard error units.

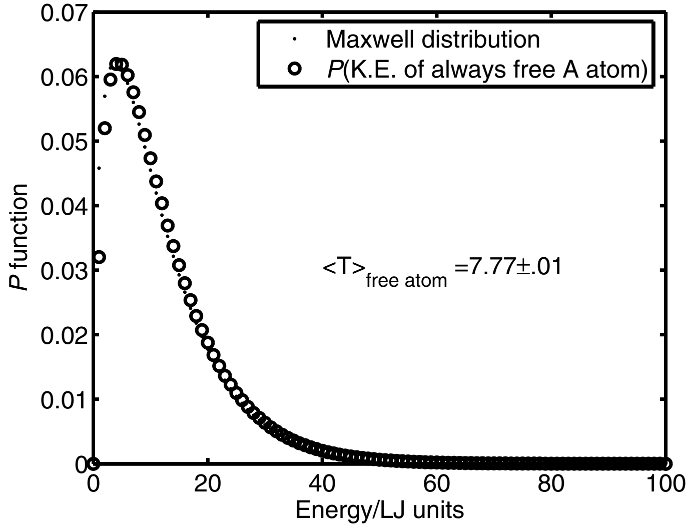


FIGURE 7. P functions for kinetic energy of a free (unbonded) random atom at system temperature $T^*_{set} = 8.0$ and $\rho = 0.7$ with apparent temperature of the random atom indicated.

The following is suggested as a result of the above observations.

Conjecture 1. *If the random forces are external to the subsystem, and they all have the same force law when acting on the particles of the system which may be different from the force law for internal forces acting on the particles of the same subsystem, then the kinetic energy of the C.M. of the subsystem would have a probability distribution that is Maxwellian.*

The above conjecture is weak as it stands and should be supported by a theoretical approach using stochastic calculus.

3. NEMD results

A NEMD simulation was conducted with the thermodynamical variable distribution for temperature and number density depicted in Fig. 8. The results presented here are additional to the results presented elsewhere [1, Case 2 simulation] for the same thermodynamical conditions, where this time, we concentrate on the flow properties of the system, rather than the static property of the equilibrium constant variation across the cell given previously. Figs. 9, 10 are the flux and divergence of the flux for 'Case 2' simulation where a temperature gradient across the MD cell is imposed together with the making and breaking of bonds at the ends of the cell leading to a molecular flux according to the thermodynamical conditions and rate details of the breaking and formation of bonds as given in [1].

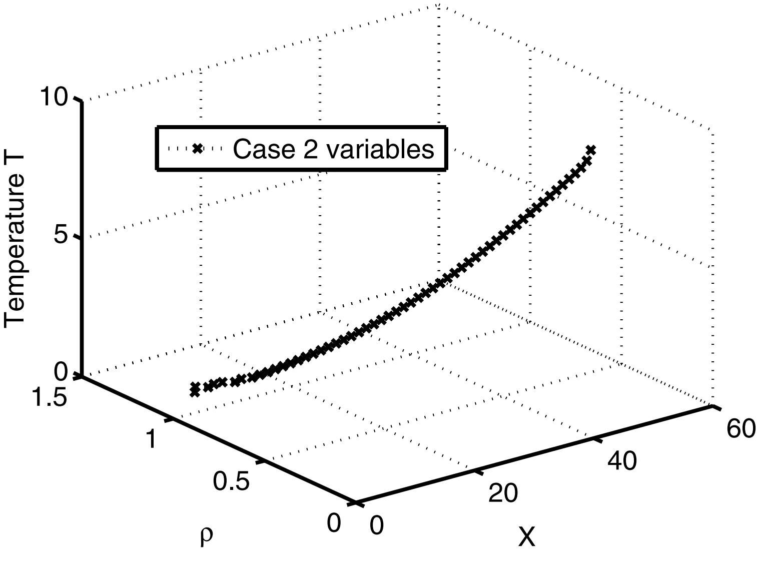


FIGURE 8. Temperature and density profile for Case 2 simulation along the MD cell which was divided up into 64 layers in the X axis direction.

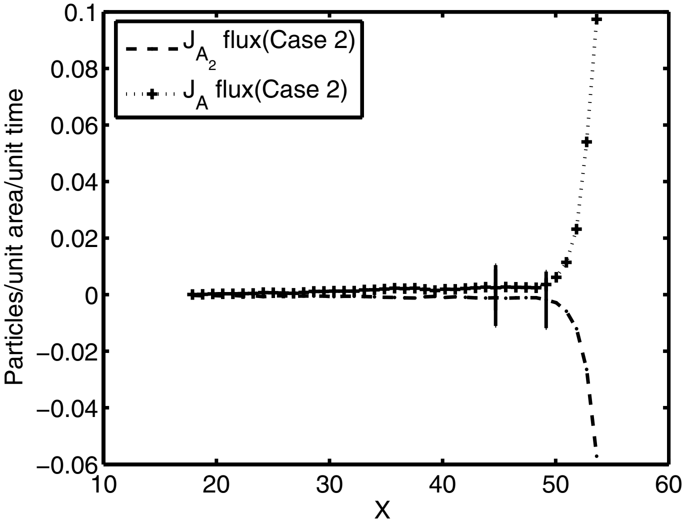


FIGURE 9. Extreme thermodynamical conditions leading to the presence of steady state atomic and dimer fluxes. The data points are used to construct the difference equation in the text to verify the conservation law.

The cell is broken up into 64 layers along the X-direction and the thermostats are placed at the ends of the layers. Fig. 9 has overlapping error bars with magnitudes that do not change significantly over the range where the fluxes are evident. The stationary source and sink quantities are denoted σ (σ_f and σ_b are the rate of formation and breakdown of the dimer in unit time and unit volume respectively throughout the cell). The conservation of mass equation for atoms and dimers reads as follows, where the subscripts refer to the species label for the flow vector J and the concentration c:

$$\begin{aligned} dc_{A_2}/dt &= -\nabla . J_{A_2} + \sigma_f - \sigma_b\,, \\ dc_A/dt &= -\nabla . J_A - 2\sigma_f + 2\sigma_b\,. \end{aligned} \tag{14}$$

The steady state conditions are

$$\begin{aligned} \nabla . J_A &= -2(\sigma_f - \sigma_b) = -2\sigma_r\,, \\ \nabla . J_{A_2} &= \sigma_r\,, \end{aligned} \tag{15}$$

where $(\sigma_f - \sigma_b) = \sigma_r$ and σ_r is a scalar flux. At thermodynamical equilibrium, $\sigma_r = 0$. If the PLE were valid in the sense that for chemical reactions which are in a state of local equilibrium, the affinity of the reaction A is zero leading to zero σ_r, then the J_A, J_{A_2} fluxes must vanish; clearly here, this is not the case. Some elaboration seems necessary. The affinity is defined as $A = \sum_{i=1}^{n} \nu_i \mu_i$ where μ_i is the chemical potential. The Gibbs equilibrium criterion is equivalent to the affinity vanishing at constant pressure and temperature. The rate σ_r cannot be linearly proportional to the temperature gradient, if the common understanding of the Curie symmetry principle is used [11, p. 21]. One can couple a flow J_i of substance i and rate v according to the linear thermodynamic equations

$$\begin{aligned} J_i &= L_{ii}\frac{d}{dx}\frac{1}{T} + L_{ic}\frac{A}{T}\,, \\ v &= L_{ci}\frac{d}{dx}\frac{1}{T} + L_{cc}\frac{A}{T}\,. \end{aligned} \tag{16}$$

In such an understanding, $L_{ic} = 0$ or else a scalar cause A/T produces a vector effect J_i. So $L_{ci} = 0$ also by the reciprocity condition. Hence in the naive sense above, one would not expect flows to be present along the cell where there is no artificial (externally imposed) formation or breaking of bonds for the reasons that follow. One might argue that $L_{ii} \neq 0$ induces the flow; but it was found that no perceptible flow was observed when there was no breaking or forming of bonds at the reservoir ends; but in any case (16) suggests that the rate is due only to the affinity not being zero, and the conservation equations show that the divergence of the flow is related to the chemical rate $v = \sigma_r$. If the rate were zero over the whole length of the cell, then if the flow rate $J_A(J_{A_2})$ were zero at one end of the cell, then by integration it would also be zero along the whole cell length; experimentally the flow is zero at one of the ends (at colder temperature), so a zero reaction rate v everywhere implies zero flow rate elsewhere under these conditions. To check for flux conservation, the divergence term is discretized by integration over one layer, using the trapezoidal rule, where for any layer i,

$$\int_{i-1}^{i} \nabla.J_{A_2} dV = \frac{(\sigma_r(i) - \sigma_r(i-1))\Delta V}{2} = J_{A_2,dif}(i) = J_{A_2}(i) - J_{A_2}(i-1) \tag{17}$$

where the layer has volume ΔV. Similarly, for the atomic fluxes,

$$J_{A,dif}(i) = J_A(i) - J_A(i-1) = -(\sigma_r(i) + \sigma_r(i-1))\Delta V. \tag{18}$$

Eq.(15) says that

$$2\nabla.J_{A_2} + \nabla.J_A = 0 \tag{19}$$

which may be expressed as

$$J_d(i) = 2J_{A_2,dif}(i) + J_{A,dif}(i) = 0. \tag{20}$$

The plot of J_d given in (20) in Fig. 10 complies with the conservation law rather well, within statistical error. We have therefore shown that PLE in the above sense is not a rigorous principle from numerical simulation with this counterexample. Another result from NEMD concerning equilibrium constants has been reported [1]. It has been shown that local stochastic equilibrium dynamical variables do not necessarily have Gaussian (canonical) distributions. Both these conditions are demanded as being essential by some specialists [9, 10, 12] in their theories. The theoretical developments concerning PLE begins with the generalized Clausius Inequality,

$$\oint \frac{dQ_{[q]}}{T} \leq 0\ , \quad q \in \{adia,\ tot\} \tag{21}$$

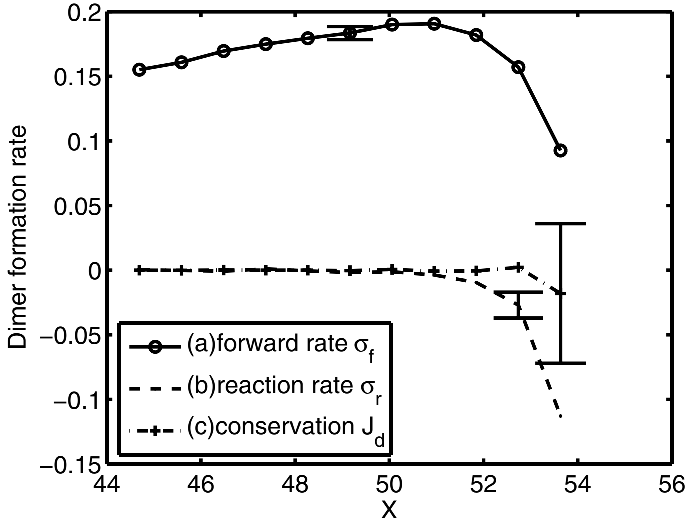


Figure 10. Test of divergence theorem for mass conservation via a difference equation.

where [1] two separate forms of heat obtain: (i) *adia* refers to a diathermal heat transfer across the primary system boundary whereas (ii) *tot* refers to a nonlocal heat term which includes various heat transfer terms due to standard state substance and thermal reservoirs, heat pumps and the primary system. The inequality above holds for both types of heat transfer separately. It is deduced [1, Corollary 1] that it is impossible for any irreversible pathway which exchanges heat P'_{BA} connecting two equilibrium states A, B to contain the same sequence of points as P_{BA}, that is, for the equilibrium pathway for all P_{BA}. The set of all equilibrium states (which are points in the thermodynamical space) is Σ where $P_{AB} \subset \Sigma$ and where the set of points in P_{BA} (or P_{AB}) is denoted $\{\omega\}$. It is shown that $P'_{BA} = \{\omega\} \cup \{\Delta\}$ where $\Delta \notin \Sigma$. The theoretical development [1] does not provide a specific form for Δ but physical considerations suggest that this variable includes spatial gradients and time derivatives of Σ. As such, the theoretical development states that the use of simple differentials of equilibrium state functions used routinely to describe nonequilibrium systems is incomplete. It is suggested here that this incompleteness shows up in the NEMD simulation results provided here, which is not well described by the first order linear thermodynamics theory in conjunction with the Gibbs equilibrium criterion.

4. Conclusion

We have shown through numerical counter-examples that the PLE and the canonical averaging assumption used in recent thermodynamical theories as fundamental and required assumptions are approximate in nature, at best. In canonical averaging, the internal variables do not have the same algebraic structure as the variables that are explicitly featured in the system Hamiltonian. A previous work [1] shows that the PLE neglects other variables not found in the equilibrium state space. It would be of interest to repeat and compare some of the above calculations for a more conventional system without hysteresis, to definitively rule on the effects of artifacts due to the use of these novel potentials. The NEMD simulation provides an example of a system that may be better described by theories that go beyond linear and local equilibrium theories.

Acknowledgment

C.G.J would like to thank the Malaysian Government for availability of Intensification of Research in Priority Areas (I.R.P.A) grant no. 09-02-03-1031 for financial assistance for the entire project, which included research visits.

References

[1] C. G. Jesudason, The Clausius inequality: Implications for non-equilibrium thermodynamic steady states with NEMD corroboration. *Nonlinear Analysis*, **2005**, *63(5-7)* e541-e553.

[2] M. G. Calkin, *Lagrangian and Hamiltonian Mechanics*; World Scientific: Singapore, 2001.

[3] J. W. Gibbs, *Elementary Principles in Statistical Mechanics*; Dover 1902 reprint, New York, 1960.

[4] L. E. Reichl, *A Modern Course in Statistical Physics*; Wiley-Interscience, New York, 1998.

[5] C. G. Jesudason, Analysis of open system Carnot cycle and state functions,*Nonlinear Analysis- Real World Applications*, **2004**, *5(4)* 695-710.

[6] D. Reguera, J. M. Rubi, J. M. G. Vilar, The Mesoscopic Dynamics of Thermodynamic Systems, *J. Phys. Chem. B*, **2005**, *109*, 21502-21515.

[7] S. Abe, Axioms and uniqueness theorem for Tsallis entropy, *Phys. Lett. A*, **2000**, *271(1)* ,74-79.

[8] H. Kuhn, H-D. Försterling, *Principles of Physical Chemistry*; J. Wiley & Sons: Chichester, 2000.

[9] J. M. G. Vilar, J. M. Rubi, Thermodynamics “beyond” local equilibrium, *PNAS*, **2001**, *98(20)*, 11081-11084.

[10] I. Pagonabarraga, A. P. Madrid, J. Rubi, Fluctuating hydrodynamics approach to chemical reactions *Physica A*, **1997**, *A337*, 205-219.

[11] W. Yourgrau, A. van der Merwe, G. Raw, *Treatise On Irreversible And Statistical Thermophysics*; Dover Publ. Inc. : New York, 1982.

[12] J. Keizer, *Statistical Thermodynamics of Nonequilibrium Processes* ; Springer: Berlin, 1987 and the many publications of the same author cited within.

Christopher G. Jesudason
Chemistry Department, University of Malaya
50603 Kuala Lumpur, Peninsula Malaysia
e-mail: jesu@um.edu.my

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 161–193
© 2008 Birkhäuser Verlag Basel/Switzerland

Mathematical Modelling and Simulation of Coronary Blood Flow

Bernhard Quatember and Martin Mayr

Abstract. As *in vivo* observations and measurements of flow conditions in the coronary circulation are extremely difficult and clinically almost infeasible, the use of mathematical models and the performance of simulation studies are the only practical way for a good understanding of coronary haemodynamics. As the range of clinical applicability of known models in this field is rather limited, we developed a detailed lumped parameter model of the coronary haemodynamics. From the beginning we aimed at its use as an aid for interventional cardiologists and heart surgeons. At this stage of development, our lumped parameter model can already be of some help to physicians, when appraising the adverse effects of stenoses on myocardial blood supply and assessing the attainable improvement of the supply by therapy. The crude approximations inherent in lumped parameter modelling restrict the applicability of this unsophisticated approach to an imprecise global assessment of the blood supply to the myocardium and its detoriation in the case of coronary artery disease. However, to be able to assess other specific pathophysiological processes, such as thrombus development and stenoses growth, for instance, we need precise knowledge of the local three-dimensional flow pattern in a stenosed section of the coronary artery tree, especially around the apex of the stenosis. We present simulation studies of the three-dimensional flow in stenosed sections of the coronary arteries, based on a distributed parameter modelling approach. In these studies, the governing partial differential equations are solved with the finite element method. Particular attention is given to the acquisition of patient-specific data, especially of data describing the geometry of the patient's epicardial arteries, derived from medical images. These data are required not only for the patient-specific adaptation of our lumped parameter model but also, in the case of our three-dimensional flow simulations, for the generation of a finite element mesh in the flow domain under investigation. However, we deal with the mesh generation issues only very briefly.

Mathematics Subject Classification (2000). 81T80, 37N10, 76D05, 35Q30, 65N50.

Keywords. Mathematical modelling; Simulation model; Lumped parameter model; Nonlinear model; Coronary haemodynamics; Coronary stenoses.

1. Introduction

In this chapter, we deal with modelling approaches and simulation studies of the blood flow of the human cardiovascular system, especially in the coronary vessels. The fluid mechanics of blood circulation, or haemodynamics, is a rather difficult scientific domain — mainly due to the complicated structure and function of the cardiovascular system and considerable differences between individuals. In the human body, the cardiovascular system performs several essential physiological functions. An important one is to transport the substances involved in metabolic processes that take place in the cells (tissues and organs). Substances that play a major role in these processes are oxygen, various nutrients, and carbon dioxide. The cardiovascular system consists of the heart, the pulmonary circulation, and the systemic circulation. A relatively small but very important part of the systemic circulation is the coronary circulation which is responsible for the supply of blood to the myocardium. The coronary vessels comprise the epicardial arteries, the intramyocardial arteries and arterioles, the capillary bed, the intramyocardial venules and veins, and the epicardial veins. The entire system of coronary vessels is subdivided into the left and the right coronary network, since both of them have a tree-like structure. The coronary blood flow must be strong enough to meet the metabolic requirements of the heart, for otherwise the myocardium becomes vulnerable to ischemia [1–3]. The coronary blood flow can be reduced by diffuse narrowing (atherosclerotic plaques), and especially by stenoses in the epicardial arteries as well as by thrombi. Thrombus formation and development usually takes place in the area of a stenosis in the epicardial arteries. The adverse effects of these pathological changes are of particular importance for interventional cardiology and coronary surgery.

Measuring the blood flow within the coronary vessels is fraught with considerable difficulties, and *in vivo* measurements of pressure and flow pose especially severe problems. Given these difficulties, it is obvious that the use of simulation models of the coronary haemodynamics is extraordinarily important and almost indispensable [4–8] for:

- facilitation of the detailed study of the flow behaviour in the coronary circulation easily and thoroughly;
- assessment of the adverse effects of pathological changes (e.g., stenoses), especially in the field of coronary artery disease; and
- predicting the success of planned therapeutical measures.

It is thus very important to provide physicians with appropriate simulation models.

To quantitatively assess the supply of blood to the myocardium and its impairment that results from pathological changes, we need simulation studies of the entire coronary network which thus have a global character [9]. Due to the complexity of the coronary vessels, the blood flow in the entire coronary network can only be simulated by employing lumped parameter models. We developed such a model, which allows a quantitative assessment of the blood supply to the myocardium, especially the reduction that results from stenoses and other obstructions in the epicardial arteries. Our lumped parameter model of coronary haemodynamics is described in § 2 and § 3. The results of simulation studies with this model are presented in § 4.

Severe stenoses in the epicardial arteries not only reduce the blood supply to the myocardium but may also stimulate pathological processes at the molecular and cellular level, mainly due to irregular flow conditions around a stenosis. In these processes, activities in the blood cells and their immediate vicinities play a predominant role. These include cell-cell interactions, cell-plasma interactions and interactions between the blood cells and the arterial wall. The most interesting sites of these activities are the region around the apex of the stenosis and the flow domain downstream from the centre of the stenosis. These processes are of considerable clinical interest, since they are responsible for the growth of the stenosis and the eventual formation and development of thrombi. A sound knowledge of the relevant three-dimensional flow conditions is required in all investigations of these pathophysiological processes. A lumped parameter modelling approach of course cannot provide the sufficiently detailed haemodynamic knowledge that would be required for investigations of this kind, since this medical problem area requires a fair knowledge of the three-dimensional flow pattern in an epicardial artery. We describe computer simulations of the three-dimensional flow field around a severe eccentric stenosis in § 4.

At this stage of development, our modelling approaches are solely based on the morphological and physiological data found in the literature. They are therefore not patient-specific, so the range of clinical applicability is rather limited. However, we are now developing methods and software modules to adapt our lumped parameter model to patient-specific data, and give some relevant details in § 6.

2. Basics of our lumped parameter modelling approach

We constructed a simulation model of the coronary flow dynamics that can be used to investigate the blood supply to the myocardium under physiological conditions, and under impaired perfusion conditions that result from coronary artery disease. In our modelling efforts, we concentrated on the adverse effects of stenoses (local narrowing) and diffuse narrowing in the epicardial arteries. However, our modelling concept could easily be adapted to other pathophysiological changes.

The simulation model described here has been designed on the basis of the geometric (morphometric) data, mechanical properties and values of other functional parameters relevant for a hypothetical average adult. In this model, it is possible to define geometrical changes that involve stenoses and diffuse narrowing in the epicardial arteries, for studies of coronary artery disease. At this stage of development, as mentioned earlier, our modelling approach is not yet patient-specific, but we are attempting to acquire the patient-specific data needed for adaptation of the simulation model to individual patients.

We confined our modelling to the left coronary artery system, i.e., to that part of the coronary network belonging to the left coronary artery, since this subsystem plays an especially important role in our medical problem area. Nevertheless, the model could of course also be applied to the right coronary arterial system.

2.1. General characteristics

As previously mentioned, we have chosen a lumped parameter modelling approach. Lumped parameter models only permit simulation studies that very roughly approximate the real behaviour, but have the following distinct advantages:

- the model equations are a system of differential algebraic equations, whose solution is much less computationally expensive than the solution of the partial differential equations used in the case of the more accurate distributed parameter models; and
- system identification tasks that would enable the determination of patient-specific parameters by means of measurement results remain tractable.

Due to these advantages, lumped parameter models are frequently used in current cardiovascular research [10–17].

Our modelling is based on several simplifying assumptions regarding the structure and geometry of the coronary vessels. We consider the arterial and venous system to have a strict tree-like structure. Collateral conduits have not yet been taken into account. The flow domain in each segment of the epicardial arteries and veins is modelled as a cylinder with a perfectly circular luminal cross-sectional area. We regard the tree-like structure of the intramyocardial arteries, arterioles and venules, veins as being symmetric — and model each intramyocardial generation of the tubular tree as a parallel connection between identical tubes. The capillary bed is treated as an intramyocardial generation of the tree-like structure.

Apart from these coarse approximations of the structure and geometry of the coronary vessels, a further simplification regarding the mechanics of flow in the tubular vessels has been made. Our modelling is based on assuming that the flow through each tubular segment is a "fully developed" incompressible laminar viscous flow – i.e., classical Hagen–Poiseuille flow. The velocity profile across the luminal cross-sectional area is thus a paraboloid, with zero velocity at the vessel wall and maximal velocity at the centreline of the vessel. One may however consider that, at each and every point in time, blood is flowing uniformly across the entire luminal cross-sectional area at an average velocity. Thus the flow velocity is independent of spatial coordinates, and the blood flow can be considered to be a process with lumped parameters [18].

The coronary artery tree originates from the aorta, in close proximity to the aortic valve. From there, the blood flows through the arterial subsystem, the capillary bed and the venous subsystem of the coronary circulation. Finally, it is drained into the right atrium, mainly via the coronary sinus. The pressure at the inlet of the coronary vessels is thus the aortic pressure $p_A(t)$, which in our model is thus the input pressure. On the other hand, the pressure at the outlet of the coronary circulation is virtually identical to the pressure $p_R(t)$ in the right atrium, and so the output pressure in our model. The total perfusion pressure is

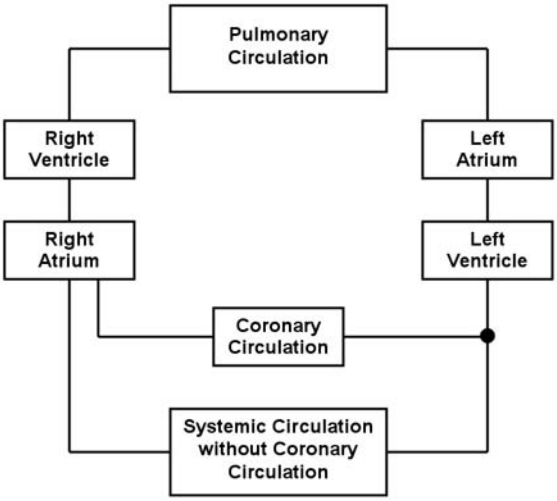


FIGURE 1. Block diagram of the auxiliary model of the entire cardiovascular system, with embedded model of the coronary circulation.

thus $p_T(t) = p_A(t) - p_R(t)$, which can be regarded as the driving pressure of the coronary circulation.

A fair knowledge of the pressures $p_A(t)$ and $p_R(t)$ and their variation in time are a prerequisite for carrying out any simulation runs of our model. To generate the values of $p_A(t)$ and $p_R(t)$, we developed an auxiliary model of the entire cardiovascular system, with our model of the coronary circulation embedded in this system. This auxiliary model is also a lumped parameter model. Figure 1 shows the structure of the auxiliary system, together with the embedded lumped parameter model of the coronary circulation. The auxiliary model comprises the following submodels:

- submodel of the left atrium (with mitral valve);
- submodel of the left ventricle (with aortic valve);
- submodel of the right atrium (with tricuspid valve);
- submodel of the right ventricle (with pulmonary valve);
- submodel of the pulmonary circulation; and
- submodel of the systemic circulation excluding the coronary circulation, which is a small but very important part of the systemic circulation.

Moreover, the auxiliary model is based on an especially crude modelling approach. For each submodel, we provided only a small number of lumped components. These have been formulated in conformity with the description and the data given in [19].

The total perfusion pressure $p_T(t)$, the structure and geometry of the coronary vessels, and the mechanical properties of the vessel walls all have substantial influence on the coronary blood flow. However, in a thorough analysis of the coronary blood flow, we also have to consider other phenomena and quantities. Various aspects of the regulation of the coronary blood flow play a decisive role in coronary haemodynamics, and we must also take into account the interactions between coronary blood flow and myocardial mechanics, especially the effects of the intramyocardial pressure on the coronary blood flow [20, 21].

2.2. Morphological and physiological basis

At this stage of development, our modelling approach is based solely on morphometric and physiological data found in the literature [3, 8, 21–26]. Collecting the information required was a difficult task, since articles with appropriate data are rare. Moreover, we have to consider that measurement data very often refer to animal experiments. Additional difficulties are caused by the variation of important characteristics in coronary networks, even among healthy adults. This is particularly true of the structure and geometry of the epicardial arteries. We considered a number of relevant descriptions in the literature, and made every effort to define a structure and geometry for the epicardial arteries that can be regarded as representative for a hypothetical average adult, as the basis for our modelling approach. We deal with the issue of patient-specific data in § 6 in detail, but at some future time we intend to consider the patient-specific geometry of the epicardial arteries.

2.2.1. Morphology of the coronary vessels. As previously mentioned, we developed an adequate representation of the structure and geometry of the epicardial arteries [27]. In doing so, we simplified the complex network of the epicardial arteries by combining smaller arteries to form a tree-like structure with eight arterial branches. In coronary artery disease, stenoses and diffuse narrowing almost exclusively appear in the epicardial arteries. In our model, we are already able to specify stenoses with various degrees of severity, and to determine the site where they appear. However, we are not yet able to account for the formation of collateral conduits.

In modelling the intramyocardial arteries, we took the morphometric investigations of Spaan [21] completely into account, by providing a separate lumped component for each of the 10 layers of the morphometric scheme in our model. Detailed descriptions of the capillary bed in the literature [23, 28–30] served as the foundation of this part of the coronary network, which plays a key role in the supply of the myocardium with oxygen and nutrients.

Very little precise morphometric data of the venous system could be found in the literature. Our modelling of the venous system assumes that it is more or less a mirror image of the whole arterial tree. However, we accounted for the higher blood volumes in the venous system by making an appropriate adaptation in the dimensions.

2.2.2. Mechanical properties of the coronary vessels. The specification of the mechanical properties of the coronary vessels was based on carefully selected data from the literature [21–23, 25, 29–32].

We assumed that the walls of all blood vessels are purely elastic. This assumption is of course a simplification, since blood vessels exhibit viscoelastic properties, but we consider that ignoring the viscoelastic behaviour should not have significant influence on the results.

However, we allowed for the nonlinear character of the assumed purely elastic behaviour of the vessels, in terms of ”transmural pressure-luminal area” relations, by using distensibility diagrams from the literature [21, 23, 25, 28] in formulating the model equations.

2.2.3. Effects of intramyocardial pressure on coronary blood flow. The luminal cross-sectional area of an intramyocardial vessel is dependent on the transmural pressure, which is the difference between the pressure within the vessel and the extravascular pressure. The extravascular pressure in the myocardium varies, due to periodic compression within the myocardium generated by myocardial contraction. The investigation of the extravascular pressure is a current research topic [20, 21, 31, 33–38], and precise results for this quantity are not yet available. However, in conformity with comments elsewhere in the literature, we made the following simplifying assumptions — viz,

- the hydrostatic pressure of the interstitial fluid around the vessel, the so-called intramyocardial pressure, is identical to the extravascular pressure (e.g., the effects of the attachment of connective tissue to blood vessels are ignored); and
- the intramyocardial pressure at the endocardial surface is identical to the left ventricle pressure — but on the other hand, at the epicardial surface it is zero (ambient pressure, or more specifically, intrathoracic pressure), and decreases linearly from the endocardial surface to the epicardial surface.

2.2.4. Coronary control. In the field of coronary control, we distinguish between relatively large conducting vessels and resistance vessels, primarily arterioles and pre-arteriolar intramyocardial arteries. The resistance vessels have an autoregulatory function — i.e., they enable the oxygen consumption (metabolic rate) of the myocardium to remain constant when the total perfusion pressure in the coronary system changes. Moreover, they are even able to adapt the coronary blood flow to the higher oxygen demand that occurs during increasingly vigorous physical exercise — i.e., they can increase it. The coronary control mainly serves to regulate the increase, whereby vasodilatory effects in the resistance vessels play a predominant role. However, pharmaceuticals such as Dipyridamole also have vasodilatory effects on the resistance vessels.

Under physiological perfusion conditions at rest, the perfusion pressure of the intramyocardial vessels has a value within the autoregulatory range. When that is the case, an impairment of the perfusion pressure does not cause a significant deterioration of the perfusion conditions (the blood flow). However, in coronary artery disease the range of coronary control may be exceeded as soon as the resistance vessels are fully relaxed (maximally dilated).

At present, the mechanisms responsible for coronary control are not very well understood [39–42]. For this reason, we decided not to include a detailed description of all the regulatory processes in our model. At this stage of development, we have chosen a straightforward approach to this problem area, where we confine ourselves to the above-mentioned clinically important effects — viz. the coronary flow under basal perfusion conditions, and the flow in the case of maximally dilated resistance vessels. There are the following reliable data for these two effects in the literature:

- distensibility diagrams for the resistance vessels (arterioles and also other small intramyocardial arteries) under physiological perfusion conditions at rest, for which we introduce the term "distensibility diagrams of Type A" [21, 23, 25]; and
- distensibility diagrams for fully relaxed (maximally dilated) resistance vessels, for which we will introduce the term "distensibility diagrams of Type B" [21, 23, 25].

We are thus able to carry out simulations

- based on the distensibility diagrams of Type A, which enable us to investigate blood flow under normal perfusion pressure and basal perfusion conditions; and
- based on the distensibility diagrams of Type B, which permit investigations of the blood flow under conditions of severe stenosis in an arterial branch, leading to a maximal dilation of the resistance vessels in the perfusion territory of the stenosis.

3. Description of our lumped parameter model of coronary haemodynamics

Our lumped parameter model of the dynamics of coronary blood flow comprises a large number of lumped components for the individual segments of the network of the coronary vessels, rendering a fair representation of the inhomogeneity within the coronary network. The model consists of two main parts — viz.

- the submodel of the epicardial vessels; and
- the submodel of the intramyocardial vessels.

As already mentioned, we embed our model of the dynamics of coronary blood flow in a coarse model of the entire cardiovascular system.

3.1. Block diagram

A block diagram of the model of the coronary blood flow dynamics and its connection with the auxiliary model of the whole cardiovascular system is shown in Figure 2, where $p_A(t)$ is the aortic pressure and $p_R(t)$ is the pressure in the right atrium. The total perfusion pressure $p_T(t)$ of the whole system of the coronary vessels is thus

$$p_T(t) = p_A(t) - p_R(t) \; . \tag{3.1}$$

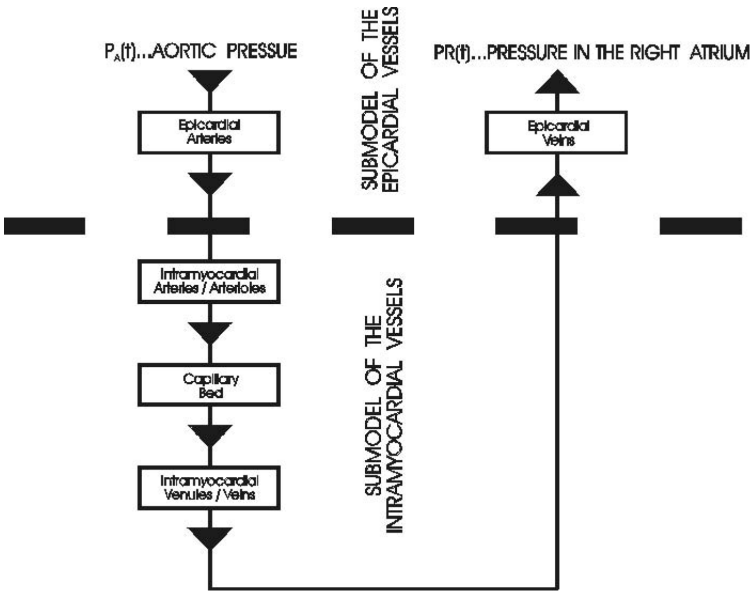

P_A(t)...AORTIC PRESSUE
SUBMODEL OF THE EPICARDIAL VESSELS
PR(t)...PRESSURE IN THE RIGHT ATRIUM
Epicardial Arteries
Epicardial Veins
Intramyocardial Arteries / Arterioles
Capillary Bed
Intramyocardial Venules / Veins
SUBMODEL OF THE INTRAMYOCARDIAL VESSELS

FIGURE 2. Block diagram of the entire system of the coronary vessels.

3.2. Pivotal points of the modelling approach

Let us now summarise the main features of our modelling, as introduced above.

1. We treat the blood as an incompressible Newtonian fluid.
2. We assume "Hagen–Poiseuille" flow, and moreover, assume that the vessels are cylindrical — i.e., circular cylinders. (Entrance effects and specific effects in bifurcation areas are of course present, but nevertheless ignored in our coarse approximation.)
3. We assume that the coronary blood flow is laminar throughout the whole cardiac cycle.
4. In the field of coronary haemodynamics, the inertia of the blood is not very important, and we consider this only in the epicardial arteries. (However, we do not consider any changes in the inertia due to volume changes of the epicardial arteries during a cardiac cycle, and in all other sections of the coronary vessels the inertia of the blood can justifiably be ignored.)
5. Body forces (such as gravity or Coriolis force) have justifiably been neglected.
6. As discussed in § 2.2.2, the viscous properties of the vascular walls are not taken into account, and we assume purely elastic, isotropic and homogeneous vascular walls.
7. As explained in § 2.2.2, we avoid the oversimplification that would result from a linear treatment — and fully consider nonlinear elastic properties of the coronary vessels, expressed in terms of the relationship between the pressure and the luminal cross-sectional area from distensibility diagrams.

8. Changes in the luminal cross-sectional area of the arteries, the arterioles, the venules and the veins affecting the flow resistance are considered; but not the changes in the lengths of these vessels (except those of the capillaries) during the cardiac cycles, since these changes do not have a significant influence on the flow resistance.
9. A rough approximation is made to consider the influences that the very complicated changes in the cross-sectional area, and in the length of the capillaries [29], have on the flow resistance.

3.3. Submodel of the epicardial vessels

3.3.1. Epicardial arteries. The structure and behaviour of the epicardial arteries are illustrated in Figure 3. The model equations for our lumped parameter model of haemodynamics are analogous to those for an electrical circuit (also lumped parameter model). As no graphic symbols for the lumped components of our fluid-mechanical model exist, we make use of this analogy and employ electrical circuit diagrams in graphical representations of our model of coronary haemodynamics — e.g., the diagrams in Figure 3 and Figure 4. In this circuit analogy:

- pressure corresponds to voltage;
- volume flow corresponds to electrical current;
- "inertance" (inertia of the fluid) corresponds to inductance;
- compliance corresponds to capacity; and
- flow resistance corresponds to electrical resistance.

Segments in Figure 3 (the lumped components) embody circular cylindrical vessels, but their luminal cross-sectional areas change with transmural pressure during cardiac cycles. At the epicardial arteries (also epicardial veins), recall that the transmural pressure is assumed equal to the pressure in the lumen (luminal pressure), since the extravascular pressure is taken to be zero. In the lumped-circuit equivalent of Figure 3 for the representation of the tree-like structured epicardial arteries, the flow resistances and the compliances associated with the individual segments are depicted by symbols for variable resistors and variable capacitors, since their values change over time throughout a cardiac cycle. On the other hand, the inertances are depicted by symbols for non-variable inductances, under our simplifying assumption that they are constant.

In the electrical circuit diagram (electrical analogue) of coronary haemodynamics shown in Figure 3:

- the pressures $p_i(t)$ and volumetric rates of flow $f_i(t)$ behave like voltages and currents;
- the inertances L_i are akin to the inductances;
- the compliances $C_i(t)$ are akin to the capacities; and
- the flow resistances $R_i(t)$ are akin to the electrical (ohmic) resistances.

In conformity with the electric circuit diagram in Figure 3, we formulate our model equations for the epicardial arteries in 14 segments.

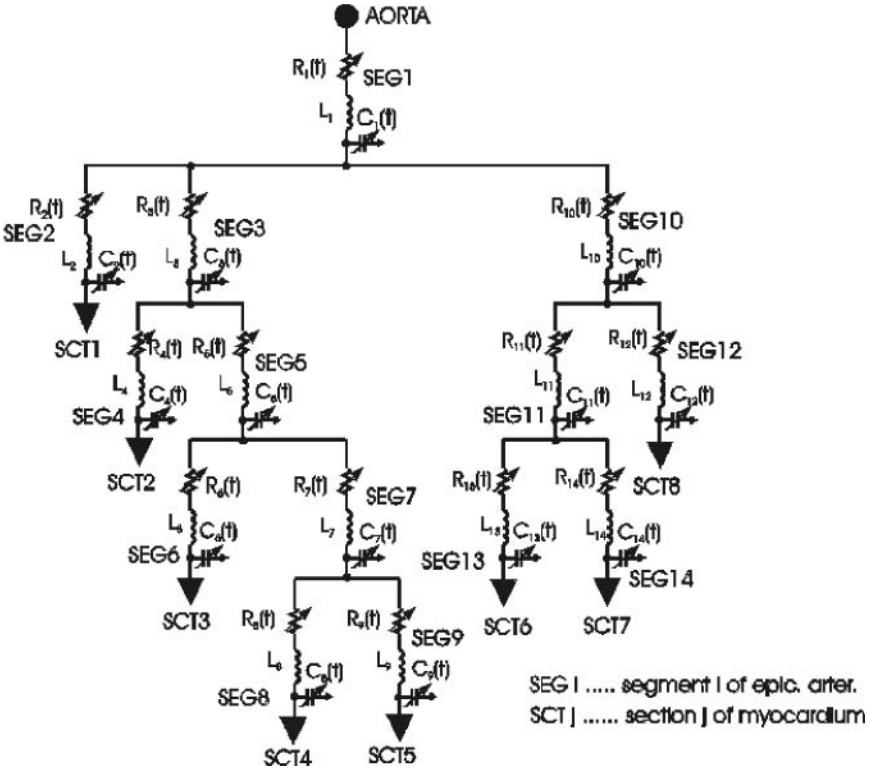


FIGURE 3. Lumped circuit diagram for the representation of the tree-like structured epicardial arteries (j-th section of the myocardium)

The equations of continuity for the individual segments are:

$$q_1(t) = \int_0^t (f_1(t') - f_2(t') - f_3(t') - f_{10}(t'))dt' + q_1^{unstr} \tag{3.2}$$

$$q_2(t) = \int_0^t (f_2(t') - f_1^{sct}(t'))dt' + q_2^{unstr} \tag{3.3}$$

$$q_i(t) = \int_0^t (f_i(t') - f_{i+1}(t') - f_{i+2}(t'))dt' + q_i^{unstr} \tag{3.4}$$

$$\text{for } i = 3, 5, 7, 10, 11 \tag{3.5}$$

$$q_4(t) = \int_0^t (f_4(t') - f_2^{sct}(t'))dt' + q_4^{unstr} \tag{3.6}$$

$$q_6(t) = \int_0^t (f_6(t') - f_3^{sct}(t'))dt' + q_6^{unstr} \tag{3.7}$$

$$q_8(t) = \int_0^t (f_8(t') - f_4^{sct}(t'))dt' + q_8^{unstr} \tag{3.8}$$

$$q_9(t) = \int_0^t (f_9(t') - f_5^{sct}(t'))dt' + q_9^{unstr} \tag{3.9}$$

$$q_{12}(t) = \int_0^t (f_{12}(t') - f_8^{sct}(t'))dt' + q_{12}^{unstr} \tag{3.10}$$

$$q_{13}(t) = \int_0^t (f_{13}(t') - f_6^{sct}(t'))dt' + q_{13}^{unstr} \tag{3.11}$$

$$q_{14}(t) = \int_0^t (f_{14}(t') - f_7^{sct}(t'))dt' + q_{14}^{unstr} \tag{3.12}$$

for $i = 1, 2, \ldots, 14$ (in 14 segments, cf. Figure 3) and $j = 1, 2, \ldots, 8$ (in 8 sections of the myocardium, cf. Figure 3) where

$q_i(t)$	is the blood volume of segment i,
q_i^{unstr}	is the unstressed blood volume of segment i,
$f_i(t)$,	is the volumetric rate of blood flow into segment i, and
$f_j^{sct}(t)$	is the volumetric rate of blood flow into the intramyocardial arteries of section j of the myocardium.

For the 14 segments in Figure 3, the nonlinear relationships between

- the luminal pressure $p_i(t)$ of segment i (equal to the transmural pressure, as mentioned above) and
- the luminal cross-sectional area $a_i(t)$ of segment i, as well as the blood volume $q_i(t)$ of this segment

are respectively expressed (for $i = 1, 2, ..., 14$) as

$$p_i(t) = \lambda_i(a_i(t)) \tag{3.13}$$

and

$$a_i(t) = \frac{q_i(t)}{l_i} , \tag{3.14}$$

where each λ_i is a nonlinear function depicted in the distensibility diagrams contained in the literature [21,23,25,28], and the length l_i of any segment i is assumed to be constant (a justifiable simplification, as mentioned above). The total pressure drop in each of the 14 segments ($i = 1, 2, ..., 14$) is

$$p_{i-1}(t) - p_i(t) = R_i(t)f_i(t) + L_i\frac{df_i(t)}{dt} , \tag{3.15}$$

$$for\ i = 1, 2, 4, 5, 8, 11$$

$$p_{i-2}(t) - p_i(t) = R_i(t)f_i(t) + L_i\frac{df_i(t)}{dt} , \tag{3.16}$$

$$for\ i = 3, 7, 9, 12, 13$$

$$p_{11}(t) - p_{14}(t) = R_{14}(t)f_{14}(t) + L_{14}\frac{df_{14}(t)}{dt} , \tag{3.17}$$

$$p_1(t) - p_{10}(t) = R_{10}(t)f_{10}(t) + L_{10}\frac{df_{10}(t)}{dt} , \tag{3.18}$$

or in integral form

$$f_i(t) = \frac{1}{L_i}\int_0^t (p_{i-1}(t') - p_i(t') - R_i(t')f_i(t'))dt' + f_i(0)\ , \tag{3.19}$$

$$\text{for } i = 1, 2, 4, 5, 8, 11$$

$$f_i(t) = \frac{1}{L_i}\int_0^t (p_{i-2}(t') - p_i(t') - R_i(t')f_i(t'))dt' + f_i(0)\ , \tag{3.20}$$

$$\text{for } i = 3, 7, 9, 12, 13$$

$$f_{14}(t) = \frac{1}{L_{14}}\int_0^t (p_{11}(t') - p_{14}(t') - R_{14}(t')f_{14}(t'))dt' + f_{14}(0)\ , \tag{3.21}$$

$$f_{10}(t) = \frac{1}{L_{10}}\int_0^t (p_1(t') - p_{10}(t') - R_{10}(t')f_{10}(t'))dt' + f_{10}(0)\ , \tag{3.22}$$

with

$$f_i(0) = 0 \qquad \text{and}$$

$$p_0(t) = p_A(t)$$

and where

L_i is the inertance of segment i which, as already mentioned, is assumed as being constant,

$p_A(t)$ is the aortic pressure, and

$R_i(t)$ is the (viscous) resistance to blood flow of segment i, given (approximately) by the Poiseuille formula

$$R_i(t) = 8\pi\mu\frac{l_i}{a_i(t)^2} \tag{3.23}$$

where

μ is the viscosity of the blood,

l_i is the length of the segment i, and

$a_i(t)$ is the luminal cross-sectional area of segment.

In the numerical simulation of $f_i(t)$ (Equations 3.19 to 3.22), $f_i(0) = 0$ has been arbitrarily chosen, and consequently some settling time must elapse before the numerical simulation produces meaningful results — i.e., the simulation procedure must progress for a considerable number of timesteps, to produce meaningful results.

3.3.2. Epicardial veins. We treated the epicardial veins in a manner similar to the epicardial arteries. Our modelling approach is based on the simplifying assumption mentioned in § 2.2.1 — viz. that the epicardial veins can be regarded in the same way as the epicardial arteries. However, a few adaptations in respect to the dimensions had to be made, and of course we had to employ different distensibility diagrams.

3.3.3. Stenoses. A stenosis in an epicardial artery causes increased flow resistance due to the pressure loss that develops across the stenosis, and the extent of this loss depends strongly on the geometry of the stenosis. Frequently, the severity of stenoses is expressed in terms of "percent stenosis", which refers to the reduction in luminal cross-sectional area, usually expressed as the percent of the luminal cross-sectional area that is unobstructed. We used the method described in [43], to calculate pressure losses across stenoses in the epicardial arteries. Stenoses can be arbitrarily chosen (defined) in all sections of the tree-like structure of the epicardial arteries in the model.

3.4. Submodel of the intramyocardial vessels

The submodel of the intramyocardial blood vessels was formulated on the basis of data found in the literature, especially morphometric data and descriptions of the mechanical properties of the vessels.

3.4.1. Intramyocardial arteries and arterioles. Our modelling of the intramyocardial arteries is based upon the morphometric scheme of Spaan [21] and distensiblity diagrams found in [23]. In accordance with Spaan's scheme, in each of the eight sections of the myocardium (left ventricle) we regarded the whole system of the intramyocardial arteries as a symmetrical tree-like structure with 10 generations.

Figure 4 shows an electric circuit diagram that represents the structure of the arteries and arterioles. Unlike the electric circuit diagram for the epicardial arteries in Figure 3, no symbols for the inductance (representing the inertance) appear in Figure 4, since the effects of inertia are insignificant in all intramyocardial blood vessels (inertance need not be considered).

However, there is an even much more important difference between epicardial arteries and intramyocardial blood vessels, since the extravascular pressure in those vessels does not have the value zero, but is equal to the intramyocardial pressure at the particular position of the individual artery (arteriole) within the myocardium.

Moreover, we have to bear in mind that the intramyocardial pressure differs considerably throughout the myocardium, and also changes over time during a cardiac cycle. As discussed in § 2.2.3, the intramyocardial pressure will be high in the zone close to the endocardial surface, but much lower near the epicardial surface. For this reason, we divided each of the eight sections of the myocardium into three layers of equal size — viz.

- a subendocardial layer,
- a mid-myocardial layer, and
- a subepicardial layer.

We treated the arteries of each generation in an individual layer as an array of identical parallel tubes, with each array lumped together to form a single lumped component of our lumped parameter model.

In § 2.2.3, we made the simplifying assumptions that the intramyocardial pressure increases linearly from the epicardium to the endocardium. At the epicardial surface, its value is equal to zero; and at the endocardial surface, it is equal

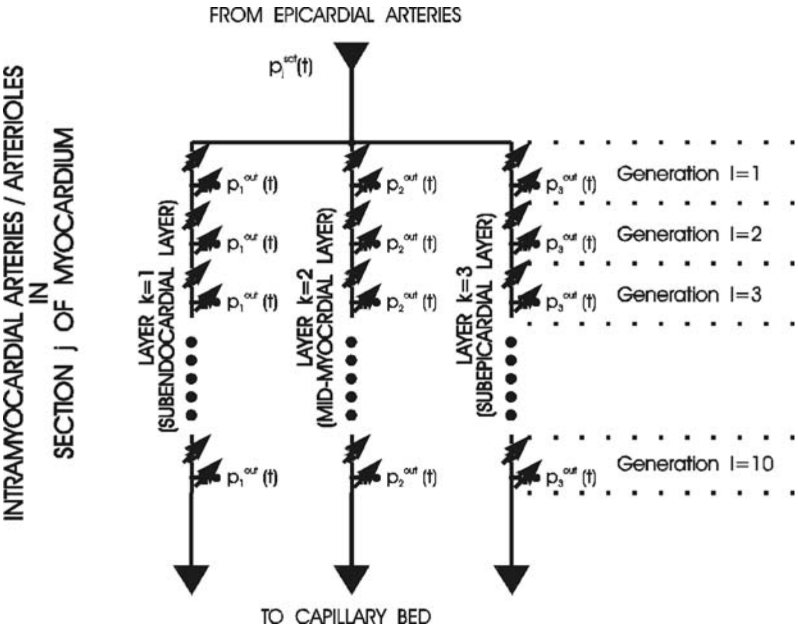


FIGURE 4. Lumped circuit diagram for the representation of the intramyocardial arteries.

to the pressure within the left ventricle. Moreover, we assume that the extravascular pressure within each layer has the same value for all arteries and arterioles within this layer. As previously mentioned, the arteries and arterioles of each layer are represented by 10 lumped components. and the extravascular pressure of these 10 components will thus be the same. For the k individual layers ($k = 1, 2, 3$), we assume the extravascular pressure in the subendocardial layer to be

$$p_1^{out}(t) = 0.875 P_L(t), \tag{3.24}$$

the extravascular pressure in the mid-myocardial layer to be

$$p_2^{out}(t) = 0.5 P_L(t), \tag{3.25}$$

and the extravascular pressure in the subepicardial layer to be

$$p_3^{out}(t) = 0.125 P_L(t), \tag{3.26}$$

where $P_L(t)$ is the pressure in the left ventricle.

In conformity with the electric circuit diagram in Figure 4, we formulated the model equations for the individual lumped components as follows.
The equations of continuity are of the form

$$q_{j,k,l}^{lay}(t) = \int_0^t (f_{j,k,l}^{lay}(t') - f_{j,k,l+1}^{lay}(t'))dt' + q_{j,k,l}^{layunstr} \tag{3.27}$$

with

$$f^{lay}_{j,k,11} = f^{cap}_{j,k}$$

where

$q^{lay}_{j,k,l}(t)$ is the blood volume of lumped component l in layer k of section j,
$q^{layunstr}_{j,k,l}$ is the unstressed blood volume of lumped component l in layer k of section j, and
$f^{cap}_{j,k}$ is the volumetric rate of blood flow from the arterial system in layer k of section j of the myocardium (left ventricle) into the capillary bed, for

$j = 1, 2, ..., 8$ (8 sections of the myocardium, c.f. Figure 4),
$k = 1, 2, 3$ (3 layers in each section, c.f. Figure 4) and
$l = 1, 2, ..., 10$ (10 generations, c.f. Figure 4).

The transmural pressure of all arteries (arterioles) belonging to the lumped component l in layer k of section j (the array of parallel tubes) is the difference

$$p^{lay}_{j,k,l}(t) - p^{out}_k(t)$$

between

- luminal pressure $p^{lay}_{j,k,l}(t)$ of all arteries (arterioles) that belong to the lumped component l in layer k of section j (the array of parallel tubes that is lumped together) and
- the extravascular pressure $p^{out}_k(t)$ of layer k.

For each of the 10 generations in Figure 4, the nonlinear relationships between

- the transmural pressures $p^{lay}_{j,k,l}(t) - p^{out}_k(t)$, which refer to all arteries (arterioles) that belong to lumped components of generation l, and
- the luminal cross-sectional area $a^{lay}_{j,k,l}(t)$ of all (identical) arteries (arterioles), which belong to these lumped components as well as the blood volume $q^{lay}_{j,k,l}(t)$ of these lumped components,

are expressed

- in the case of generations l=1 to l=8, which contain conducting non-resistance vessels, as

$$p^{lay}_{j,k,l}(t) - p^{out}_k(t) = \lambda^{gen}_l \left(a^{lay}_{j,k,l}(t) \right), \tag{3.28}$$

and

$$a^{lay}_{j,k,l}(t) = \frac{q^{lay}_{j,k,l}(t)}{n^{lay}_{j,k,l}\ l^{gen}_l} \tag{3.29}$$

for
$j = 1, 2, ..., 8$ (8 sections of the myocardium (cf. Figures 3 and 4))
$k = 1, 2, 3$ (3 layers in each section (cf. Figure 4))
$l = 1, 2, ..., 8$ (generations which contain conducting non-resistance vessels. (cf. Figure 4)),

where

λ_l^{gen} is a nonlinear function as depicted in the distensibility diagrams contained in the literature [23],

l_l^{gen} is the length of all the (identical) arteries (arterioles) which belong to the lumped components of generation l; this length, as mentioned above, is assumed as being constant and

$n_{j,k,l}^{lay}$ is the number of all the (identical) arteries (arterioles) which belong to the lumped component l in layer k of section j (array of parallel tubes which is lumped together); and

- in the case of generations l=9 and l=10, which contain resistance vessels, as

$$p_{j,k,l}^{lay}(t) - p_k^{out} = \lambda_l^{genTypA}\left(a_{j,k,l}^{lay}(t)\right) , \tag{3.30}$$

$$p_{j,k,l}^{lay}(t) - p_k^{out} = \lambda_l^{genTypB}\left(a_{j,k,l}^{lay}(t)\right) \tag{3.31}$$

and

$$a_{j,k,l}^{lay}(t) = \frac{q_{j,k,l}^{lay}(t)}{n_{j,k,l}^{lay}\ l_l^{gen}} \tag{3.32}$$

for

$j = 1, 2, ..., 8$ (8 sections of the myocardium, c.f. Figure 4)
$k = 1, 2, 3$ (3 layers in each section, c.f. Figure 4)
$l = 9, 10$ (generations which contain resistance vessels, c.f. Figure 4)
where

$\lambda_l^{genTypA}$ is a nonlinear function as depicted in the distensibility diagrams in the literature which refers to physiological conditions at rest and

$\lambda_l^{genTypB}$ is a nonlinear function as depicted in the distensibility diagrams contained in the literature which refers to maximally dilated resistance vessels.

The total pressure drop in a segment is

$$p_{j,k,l-1}^{lay}(t) - p_{j,k,l}^{lay}(t) = R_{j,k,l}^{lay}(t)\ f_{j,k,l}^{lay}(t) \tag{3.33}$$

and

$$f_{j,k,l}^{lay}(t) = \frac{1}{R_{j,k,l}^{lay}}\left(p_{j,k,l-1}^{lay}(t) - p_{j,k,l}^{lay}\right) \tag{3.34}$$

for

$j = 1, 2, ..., 8$ (8 sections of the myocardium (cf. Figures 3 and 4))
$k = 1, 2, 3$ (3 layers in each section (cf. Figure 4))
$l = 1, 2, ..., 10$ (10 generations (cf. Figure 4))
with

$$p_{j,k,0}^{lay}(t) = p_j^{sct}(t)$$

where

$p_j^{sct}(t)$ is the luminal pressure of that segment (branch) j of the epicardial arterial system which supplies its section (perfusion territory) of the myocardium and

$R_{j,k,l}^{imlay}(t)$ is the (viscous) resistance to blood flow of lumped component l in layer k of section j, which is given (approximately) by the Poiseuille formula

$$R_{j,k,l}^{lay}(t) = 8\pi\mu \frac{l_l^{gen}}{n_{j,k,l}^{lay}\ a_{j,k,l}^{lay}(t)^2} \tag{3.35}$$

where μ is the viscosity of the blood (assumed as being a constant).

3.4.2. Capillary bed. We treated the capillary bed as if it were an individual generation of the arterial system with one exception — viz. that the length of the capillaries is no longer regarded as being constant. Thus we considered the changes in length that occur during a cardiac cycle, based on the questionable assumption that these changes are in proportion to the changes of length of the neighbouring myocytes, to which the capillaries are connected [29]. Our depiction of the capillary bed as arrays of parallel cylindrical tubes with laminar flow is of course an oversimplification, which we had to make at this stage of development. In a future version of the model, we intend to take into account the specific characteristics of the blood flow in the capillary bed, which differs considerably from that in the arterial system.

3.4.3. Intramyocardial veins and venules. As in the case of the epicardial vessels, we treated the intramyocardial veins and venules in a manner similar to the intramyocardial arteries and arterioles, again making the simplifying assumption that the intramyocardial veins and venules can be regarded in a similar way as the intramyocardial arterial system. We also had to make a few adaptations in the dimensions, by employing distensibility diagrams of the intramyocardial veins and venules, which differ from those of the intramyocardial arterial system.

4. Simulation studies with our lumped parameter model

We have already carried out several simulation studies with our model. In the following, we present simulation results of the flow in the capillary bed (perfusion territory) belonging to an arbitrarily selected branch of the epicardial arteries.

In these simulation studies, the perfusion pressures always remain within the autoregulatory range.

We performed simulation runs under physiological conditions at rest, and for two purely hypothetical cases of stenoses, assuming (unrealistically) that the autoregulatory mechanisms do not come into play. By comparing the flow patterns in the two cases with the flow pattern under physiological conditions, we gain a

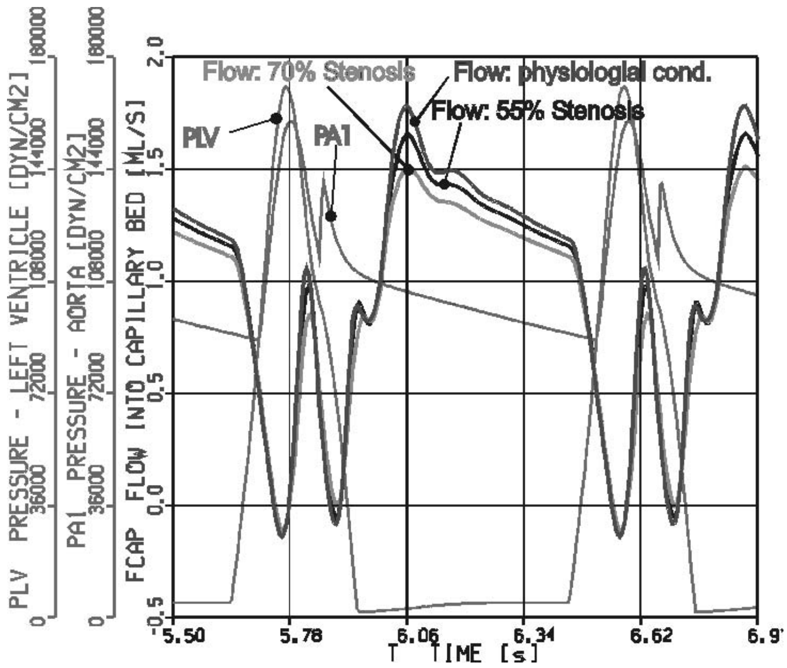


FIGURE 5. Blood flow within the capillary bed belonging to the obtuse marginal branch: comparison between physiological (normal) conditions and the purely hypothetical cases of a 55%-stenosis and a 70%-stenosis (aortic pressure curve "PA1", represented by the thin line).

better insight into the adverse aspects of stenoses, and into the extent of the necessary autoregulation, as described in § 2.2.4.

We have chosen the section (perfusion territory) of the myocardium that is supplied with blood by the obtuse marginal branch of the epicardial arterial tree. The flow into the capillary bed has been taken into consideration in presenting the following simulation results, since the supply processes to the myocardium (oxygen and nutrients) only takes place in the capillary bed.

For the hypothetical average adult, Figure 5 shows the volume flow into the capillary bed belonging to the chosen obtuse marginal branch. This figure contains three graphs, displaying:

- the volume flow in the case of physiological conditions at rest; and
- the volume flows in the two purely hypothetical cases of a single moderate stenosis in the obtuse marginal branch (case 1: 55% stenosis, 0.45 cm long and case 2: 70% stenosis, 0.45 cm long). [We thereby assume that the autoregulatory mechanisms do not come into play.]

As mentioned before, the above specification of the stenoses in terms of "percent stenoses" refers to the reduction in the luminal cross-sectional area, expressed as

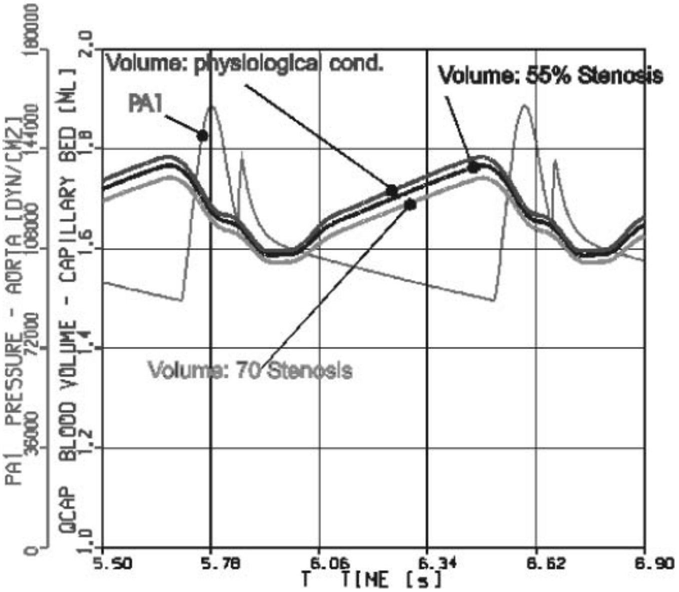


FIGURE 6. Blood volume within the capillary bed belonging to the obtuse marginal branch: comparison between physiological (normal) conditions and the purely hypothetical cases of a 55%-stenosis and a 70%-stenosis (aortic pressure curve "PA1", represented by the thin line).

the percent of the cross-sectional area of the lumen that is unobstructed. Thus the remaining luminal cross-sectional area at the position of the stenosis is respectively 45% or 30% of the unobstructed luminal cross-sectional area. For the hypothetical average adult, Figure 6 shows the variation of blood volume with time in the capillary bed belonging to the chosen obtuse marginal branch — viz.

- the blood volume in the case of physiological conditions at rest; and
- the blood volume in two purely hypothetical cases, involving a single moderate stenosis in the obtuse marginal branch (case 1: 55% stenosis, 0.45 cm long and case 2: 70% stenosis, 0.45 cm long), assuming that the autoregulatory mechanisms do not come into play.

5. Simulation studies of the three-dimensional flow patterns around an eccentric stenosis

As discussed in the Introduction, certain medical problems require simulation models produced via a distributed parameter modelling approach. A medical problem area that requires a fair knowledge of the three-dimensional pattern of the disturbed blood flow around stenoses is, for instance, the risk assessment and prevention of thrombotic and thromboembolic events [3, 44, 45].

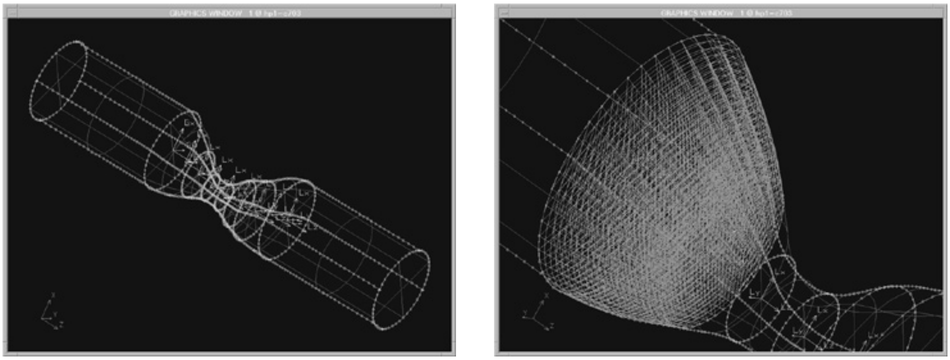
5.1. Governing equations and their solution with the finite element method

We consider blood flow within the entire domain of the coronary arteries to be laminar, and confine ourselves to steady state flow conditions at the end of the diastole. As mentioned earlier, the model equations are now partial differential equations — viz.

- the continuity equation and
- the Navier-Stokes equation,

and their three-dimensional versions are required. We solved these equations by using the finite element method. In doing so, we treated the blood as an incompressible and homogeneous non-Newtonian fluid, which obeys the power law. Moreover, it is necessary to specify initial and boundary conditions. We adopted "no slip" conditions at the wall and "natural" boundary conditions at the outlet. At the inlet, a paraboloid velocity profile was assumed as the boundary condition for our case of a steady state flow at the end of the diastole. However, in our lumped parameter coronary model for the entire circulation, we are able to calculate the aforementioned boundary condition.

Geometry of the flow domain: We assumed a geometry typical for the proximal segment of a circumflex artery with a severe eccentric stenosis. We made the simplifying hypothesis that all luminal cross-sectional areas along the entire arterial segment are ellipses. Figure 7a shows a wire-frame representation of the geometry of this diseased arterial segment. This Figure also contains the required nodes for the generation of a mesh. Such a geometric model is called a meshable representation of the geometry, or an empty mesh of the flow domain [46]. Figure 7b illustrates the generation of a multi-block mesh.



(a) Meshable geometric representation of the flow domain (empty mesh) around an eccentric stenosis.

(b) Multi-block mesh generation approach: generation of the mesh in one block of the flow domain.

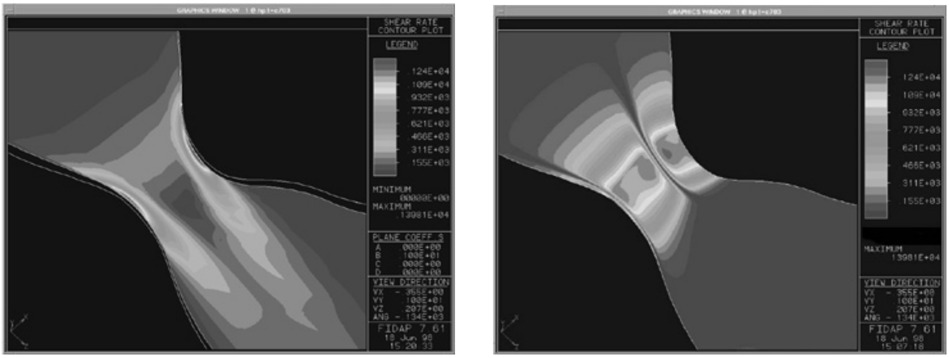
FIGURE 7. Flow domain: wireframe model and mesh generation.

5.2. Simulation results

We now present simulation results for the three-dimensional blood flow around the central region of an eccentric stenosis in the circumflex artery.

Our finite element computations of coronary haemodynamics yield a wealth of numerical data. However, excessively long lists of numerical data would be very difficult to comprehend and to analyse. For this reason, the simulation results are usually represented as surface plots, contour plots, fishnet plots, diagrams, and other graphics.

The contour plots in Figures 8a and 8b show the spatial variation of an especially important fluid mechanical quantity — viz. the shear stress in the central region of the stenosis, under steady state flow conditions at the end of the diastole and the assumption of rigid walls. Figure 7a shows the spatial variation of the shear stress within the flow domain. The spatial variation of the flow-induced shear stress along the inner arterial wall is depicted in Figure 8b. We confined ourselves to this fluid-mechanical quantity, since the shear stress plays a key role in numerous pathophysiological processes at the molecular and cellular level [45,47,48]. In both figures, we used a three-dimensional Cartesian (OXYZ) coordinate system. The Z-axis is the longitudinal axis of the stenosed artery and the X-axis lies in the longitudinal cutting plane, where the stenosis possesses greatest eccentricity. The simulation results in Figure 8a refer to the longitudinal cutting X-Z plane, whereas those of Figure 8b represent the spatial variations of the shear stress along the inner arterial wall.



(a) Contour plot of the variation of the shear stress within the flow domain on longitudinal cutting plane; range 0.00dyn/cm2 to 48.93 dyn/cm2.

(b) Contour plot of the variation of the flow-induced shear stress along the inner arterial wall; range 0.00dyn/cm2 to 48.93 dyn/cm2.

FIGURE 8. Simulation results: contour plots of the shear stress.

6. Problem area of patient-specific modelling and simulation

To be of real clinical value, simulation studies of the coronary haemodynamic must be carried out patient-specifically. Thus the simulation studies must be based on the specific geometry and the mechanical properties of a particular patient's coronary vessels, so one must acquire the necessary geometric data of the patient, which can only be derived from medical images. We confined our data acquisition to the geometry of the epicardial arteries, because this subsystem of the coronary circulation is most important for our modelling activities and shows considerable anatomical variation from individual to individual.

At present, the preferred imaging modality to fulfil this task remains biplane angiography, where the images are obtained by using a biplane angiography system that consists of two "X-ray tube – image intensifier" pairs. The X-ray tubes cast shadows of the coronary arteries, when filled with a contrast medium, onto image intensifiers. The resulting two images are called biplane angiograms. In both angiograms, the coronary artery tree must be segmented, and subsequently its structure three-dimensionally reconstructed. In clinical settings, both the segmentation procedures and the three-dimensional reconstruction should be carried out (computed) entirely or largely automatically. In the case of our simulation studies of three-dimensional blood flow around stenoses, we must also generate a high-quality mesh in order to apply the finite element method. Unfortunately, our recently developed algorithms (to undertake the tasks largely automatically) are extremely expensive in terms of computing time, and require parallel computation. Acquisition of the geometry of the intramyocardial vessels and the epicardial veins is also much more difficult than for the epicardial arteries, another topic for future research.

6.1. Segmentation methods

As previously mentioned, a fair knowledge of the geometry of the epicardial arteries is a prerequisite for our patient-specific modelling approach, and the geometry of the patient's coronary (epicardial) arteries is preferably derived from biplane angiograms. Complicated and computationally expensive image-processing methods are required to carry out a segmentation of the coronary artery tree — i.e., to separate the tree-like structure of the epicardial arteries in the angiograms from the rest of the images, especially from background structures.

Overview of our segmentation approach: Figure 9 shows a biplane angiogram (Projection A) of the contrast-medium-filled epicardial arteries. We have to allow for noise and background structures — and the finite focal size of the X-ray tube causes a considerable geometric blurring, which leads to a deterioration of the image quality. Despite the availability of advanced image-processing software, performing such a segmentation continues to be an extremely difficult and challenging task. After it became clear that the straightforward application of classical image-processing methods would fail, we developed an advanced segmentation technique using our *a priori* knowledge that the coronary (epicardial) arteries have a tree-like

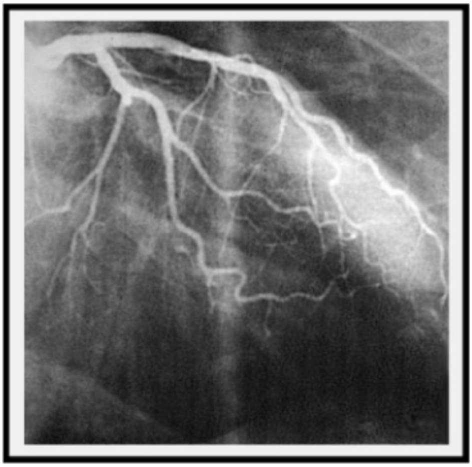


FIGURE 9. Angiogram: one projection (projection A) acquired with a biplane angiography system.

tubular structure with sections of different size. Thus we detected the representation of the coronary artery tree (X-ray shadow of the contrast-medium-filled arteries) — despite the noise, background structures, artefacts and other weak spots of the angiograms — by using special Hessian filters and applying skeletonising procedures. We devised a special software module to detect the borderlines of the coronary artery tree, which also permits a correction of geometric blurring effects. Our segmentation procedures can therefore be divided into three phases — viz.

- the Hessian artery enhancement filtering phase,
- a skeletonising phase, and
- the borderline detection phase,

described in more detail below.

Hessian artery enhancement filtering phase: Our approach is based on a differential-geometric criterion, relying on the values of so-called vesselness functions derived from Gauss-filtered angiograms. Our Hessian-based vessel enhancement filter involves a multi-scale filtering procedure, so it can be used for epicardial arteries of any size. This filtering approach involves the repeated execution of the following steps:

- Gaussian smoothing;
- calculation of a vesselness function, using the eigenvalues of the Hessian matrix of the grey scale function; and
- representation of the vesselness function as a grey scale image.

We vary the value of the standard deviation of the Gaussian filter σ_i $(i = 1, 2, ..., N)$ within an appropriately chosen range. For each i, we calculate the Hessian matrix

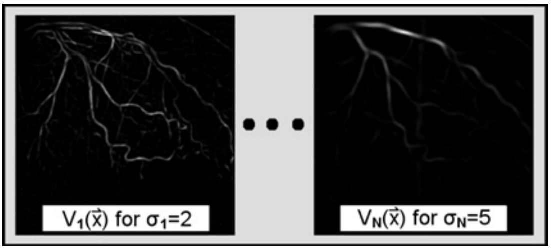


FIGURE 10. Hessian artery enhancement filtering approach: family of vesselness functions $V_i(\vec{x}),\ i = 1, 2, ..., N$.

$H_i(\vec{x})$ for each pixel $\vec{x}$ of the filtered grey scale image (filtered angiogram), and also the eigenvalues $\lambda_{1,i}(\vec{x})$ and $\lambda_{2,i}(\vec{x})$. In the region of tubular structures, we have

$$|\lambda_{1,i}(\vec{x})| \gg |\lambda_{2,i}(\vec{x})| \ . \tag{6.1}$$

The family of N vesselness functions [49] for $i = 1, 2, ..., N$ is defined as follows:

$$V_i(\vec{x}) = \begin{cases} 0 & \lambda_{1,i}(\vec{x}) < 0 \ , \\ \exp\left(-\frac{R_{B,i}(\vec{x})^2}{2\beta_1{}^2}\right)\left[1 - \exp\left(-\frac{S_i(\vec{x})^2}{2\beta_2{}^2}\right)\right] & \text{otherwise} \ , \end{cases} \tag{6.2}$$

in which

$$R_{B,i}(\vec{x}) = \frac{\lambda_{2,i}(\vec{x})}{\lambda_{1,i}(\vec{x})} \tag{6.3}$$

and

$$S_i(\vec{x}) = \sqrt{\lambda_{1,i}(\vec{x})^2 + \lambda_{2,i}(\vec{x})^2} \ . \tag{6.4}$$

The scaling factors $\beta_1 > 0$ and $\beta_2 > 0$ influence the sensitivity of $V_i(\vec{x})$ to $R_{B,i}(\vec{x})$ and $S_i(\vec{x})$, respectively. Here $V_i(\vec{x})$ is restricted to the interval [0,1], where the value 0 shows the indicated position in the image does not bear any resemblance to a tubular structure, whereas values close to 1 denote a significant similarity to the structure of a vessel. The family of N vesselness functions $V_i(\vec{x})$ is represented as a series of N grey-scale images. Figure 10 shows such individual images σ_i, each representing the vesselness function for a particular value of σ_i. This family of images is the basis for the subsequent skeletonising phase.

Skeletonising phase: From the family of filtered images just described, we extract a new grey-scale image by computing the maximum intensity projection (cf. Figure 11). This new grey-scale image reveals the crude structure of the coronary arterial tree. By thresholding, we generate a binary mask and then apply a thinning filter to obtain a skeletonised angiogram. Such a skeletonised image is shown in Figure 12.

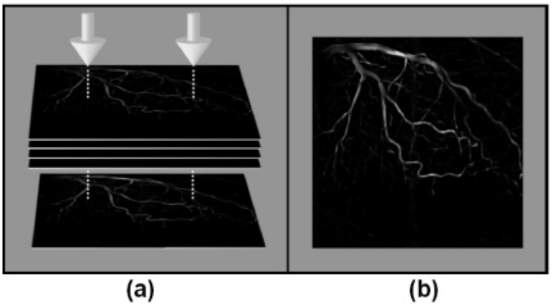


FIGURE 11. Skeletonising approach: maximum intensity projection: (a) procedure; (b) resulting image.

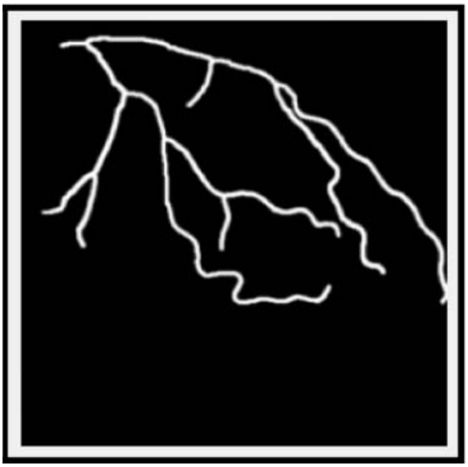


FIGURE 12. Skeletonising approach: skeleton after applying a thinning filter.

Detection of borderline phase: The third phase of our segmentation process involves the application of algorithms to detect the borderlines (edges) of the representation of the coronary arteries in our angiograms. This borderline detection is a complicated procedure, and the computationally most expensive stage of the segmentation process.

We first obtain a smooth approximation to the discrete set of points (pixels) of the skeleton image by using splines (approximating splines), which may be regarded as preliminary centrelines of the coronary arteries. The tree-like structure of these splines is superimposed onto the original image (angiogram). At relatively short intervals, we draw normals to these splines that we call scan lines. Along the scan lines we acquire the intensity profiles and then analyse each of them thoroughly, on considering an interval that is marginally larger than the expected maximal size of the coronary (epicardial) artery. Principally due to the finite focal spot size (apparent focal spot size) of the X-ray tubes, the relatively small coronary arteries become considerably distorted by blurring. We developed a special method to process these intensity profiles and so accurately determine the borderlines, using newly-developed algorithms that eliminate error caused by the blurring (cf. [50] for further details).

6.2. Three-dimensional reconstruction

We developed a special method for the three-dimensional reconstruction of the epicardial arteries, involving the following steps (cf. again [46] for more detail):

- automatically construct the centreline in the three-dimensional space, using a space curve (spline curve) that passes through the points of intersection of the back projection rays belonging to the individual pairs of corresponding points (epipolar constraints);
- automatically construct the normal planes at each of these points of intersection located on the space curve; and
- draw back projection rays from selected points of the borderlines, in both projections. The back projection rays in the vicinity of a pair of corresponding points are intersected with the normal plane that passes through the point of intersection belonging to this pair. We define four curves that pass through the aforementioned points of intersection on a normal plane. Subsequently, we construct an ellipse that approximates the aforementioned four curves. This ellipse is regarded as belonging to the inner surface of the coronary artery under consideration. In this way, we obtain a basic wire frame model of the arterial section.

6.3. Image-based generation of a high-quality mesh

We now give a brief summary of our image-based mesh generation procedures, for stenosed sections of the epicardial arteries. We used a structured mesh with hexahedra as elements, and a multi-block approach. To obtain a high-quality mesh, we adapted the size of the elements to the flow conditions. As such an adaptive procedure with an *a posteriori* error analysis would consume too much time, we decided to employ specific *a priori* criteria. Although our criteria are heuristic, they nevertheless reflect a fair quantitative *a priori* knowledge relevant to the coronary artery under investigation, derived from *a posteriori* analyses of flow patterns in so-called reference flow domains which are computed with extraordinarily high accuracy. Heuristic *a priori* criteria are frequently taken as the basis for mesh generation in computational fluid dynamics, but we recognised that such qualitative criteria often reflect engineering experience and are not always reliable — so from the outset we aimed at better criteria for the mesh adaptation, by using quantitative *a priori* knowledge. However, at this stage of development we restricted our attention to mesh generation for concentric stenoses and stenoses with a relatively low degree of eccentricity. We defined a series of reference flow domains that are axially symmetric, by systematically varying the luminal diameter of the stenosed arterial section and other characteristics of the geometry of the stenoses. In each reference domain, we generated an extremely fine mesh to compute the flow via the finite element method. Because we used such an extraordinarily fine mesh, we have confidence in the high accuracy of our computations for a particular reference domain. The obtained solution is regarded as being close to the exact solution. We performed an analysis based on this solution — specifically, we used a local error estimate for approximate solutions obtained with meshes comprising elements of a size that would allow an efficient computation of the solution with a pre-specified accuracy. (We restricted ourselves to the interpolation error, explicitly neglected all other sources of error, and applied Cea's Lemma.) Our goal was to determine the characteristics (size of the elements) of an optimal mesh for the particular reference domain and a pre-specified accuracy. From the results for all the specified reference flow domains we built a look-up table. This task can be completed before any patient-specific simulation studies are carried out. Based on the data contained in this table, we are able to generate a high-quality mesh of our stenosed sections of the coronary arteries within a relatively short period of time. This is beneficial to cardiologists who cannot wait too long on simulation results. However, we have to bear in mind that in reality the flow domains are not axially symmetric, even for stenoses which cardiologists classify as being concentric, so meshes need to be generated in genuine (not axially-symmetric) three-dimensional flow domains. We exploited results from our heuristic approach for (axially symmetric) reference flow domains as *a priori* criteria for more general mesh construction as follows. From the look-up table, we selected the data for the two reference flow domains that come closest to the previously calculated parameters of the stenosed section of the coronary artery under investigation; and from those two reference flow domains, we chose the one with the data that would produce a finer mesh. We employed this data as the control data for generation of the mesh in the flow domain within the stenosed section under investigation (cf. also [46]).

7. Implementation aspects — GRID environment

The development and implementation of our software is being done within the framework of the Austrian GRID, a newly-established computational GRID architecture. We take complete advantage of the powerful resources incorporated in this GRID, to fully exploit the possibilities afforded by parallelism. Apart from its potential for high-performance computing, the GRID can give all the cardiologists in a geographic region equal access to our software [50, 51].

The image-processing tasks and the three-dimensional reconstruction and mesh generation tasks described above are computationally extremely expensive. Fortunately, these tasks can be subdivided into numerous weakly coupled subtasks, each requiring comprehensive computations. This coarse-grain parallelism, in which the individual tasks are largely independent of one another, is well suited for the assignment of subtasks to clusters of workstations incorporated in the GRID — i.e., the supercomputers integrated in the GRID architecture are very suitable for future numerical simulation of three-dimensional blood flows based on the finite element method.

8. Concluding remarks and future work

We described a lumped parameter model and presented simulation results obtained with this model. Further, we dealt with simulation studies of the three-dimensional blood flow around an eccentric stenosis. We demonstrated that, even at this stage of development, our lumped parameter modelling and simulation facilities can help in carrying out clinical assessments.

However, the present model has various imperfections that limit its diagnostic performance capacity, so the model will be enhanced and refined in the future. We aim to make the following improvements:

- incorporate individual patient geometry of the epicardial arteries in our model (presently confined to a hypothetical average adult);
- make a detailed and more precise treatment of the control of the coronary flow, especially of the autoregulation;
- include the viscoelastic properties of the vascular walls, not yet taken into account;
- introduce a more detailed description of the coronary microcirculation, especially a more precise description of the changes of dimension of the capillary bed over time; and
- employ a more sophisticated auxiliary model for the entire cardiovascular system.

Another unsolved issue is that our model is based on the assumption that the extravascular pressure of the intramyocardial vessels is equal to the intramyocardial pressure. We also worked with the simplifying assumption that the intramyocardial pressure is strongly dependent on the position within the myocardium — at the endicardial surface, its value is almost identical to the pressure within the left ventricle, and decreases linearly to zero at the epicardial surface. However, as pointed out in [22, 37, 38], this does not exactly accord with reality. In future, we plan to provide a more realistic description of the interactions between the contracting myocardium and the coronary blood flow.

We also plan to reconsider other modelling features, especially our lumped parameter modelling approach, such as the possible formation of collaterals. Until now, we have taken for granted that the coronary arteries have a strict tree-like structure, which is essentially true under physiological conditions. However, in such a strict hierarchical system of conduits we can only identify one transport path from the inlet of the coronary artery tree to a particular capillary. Consequently, each epicardial branch of the coronary artery tree is exclusively responsible for the blood supply to a particular section of the capillary bed in the myocardium — i.e., each epicardial branch has its own perfusion territory. Usually, no efficacious collateral connections between the individual perfusion territories exist under physiological conditions. However, in coronary artery disease, an effective collateral vasculature may develop, so the number of collateral conduits and their size are strongly dependent on ischemic history. In such a meshed arterial structure, there can be two or even more paths for the transport of blood to a particular part of the capillary bed. It may then become possible to supply the primary perfusion territory of a severely obstructed or even occluded arterial branch with blood from another branch of the arterial tree via these collateral conduits. We plan to extend our lumped parameter modelling by incorporating lumped components into the model that describe the blood flow through the collateral conduits. The parameters of these lumped components can be determined on the basis of three-dimensional perfusion imagery (PET, SPECT, MRI). In doing so, we need to solve a complicated inverse problem.

Acknowledgment

The work described in this paper is partially supported by the “Austrian GRID” project, funded by the Austrian BMWF (Federal Ministry for Science and Research) under contract GZ 4003/2-VI/4c/2004./

References

[1] K. L. Gould. *Coronary artery stenosis and reversing atherosclerosis.* Arnold, London etc., second edition, 1999.

[2] J. Strackee and N. Westerhof. *The Physics of heart and circulation.* Medical science series. Institute of Physics Pub., Bristol ; Philadelphia, PA, 1993.

[3] D. P. Zipes and E. Braunwald. *Braunwald's heart disease a textbook of cardiovascular medicine.* W.B. Saunders, Philadelphia, Pa., 7th ed edition, 2005.

[4] Y. Huo and G. S. Kassab. Pulsatile blood flow in the entire coronary arterial tree: theory and experiment. *Am J Physiol Heart Circ Physiol*, 291(3):H1074–87, 2006.

[5] B. Quatember and F. Veit. Simulation model of the coronary artery flow dynamics and its applicability in the area of coronary surgery. In F. Breitenecker and I. Husinsky, editors, *Proceedings of 1995 EUROSIM Simulation Congress. Vienna, Austria. 11 15 Sept. 1995.* Elsevier, Amsterdam, Netherlands, 1995.

[6] E. B. Shim, J. Y. Sah, and C. H. Youn. Mathematical modeling of cardiovascular system dynamics using a lumped parameter method. *Jpn J Physiol*, 54(6):545–53, 2004.

[7] M. Yoshigi, G. D. Knott, and B. B. Keller. Lumped parameter estimation for the embryonic chick vascular system: a time-domain approach using mlab. *Comput Methods Programs Biomed*, 63(1):29–41, 2000.

[8] M. Zamir. *The physics of coronary blood flow.* Springer, New York, 2005.

[9] D. Manor, S. Sideman, U. Dinnar, and R. Beyar. Analysis of flow in coronary epicardial arterial tree and intramyocardial circulation. *Med Biol Eng Comput*, 32(4 Suppl):S133–43, 1994.

[10] R. T. Cole, C. L. Lucas, W. E. Cascio, and T. A. Johnson. A labview model incorporating an open-loop arterial impedance and a closed-loop circulatory system. *Ann Biomed Eng*, 33(11):1555–73, 2005.

[11] M. J. Conlon, D. L. Russell, and T. Mussivand. Development of a mathematical model of the human circulatory system. *Ann Biomed Eng*, 34(9):1400–13, 2006.

[12] V. Diaz-Zuccarini and J. LeFevre. An energetically coherent lumped parameter model of the left ventricle specially developed for educational purposes. *Comput Biol Med*, 37(6):774–84, 2007.

[13] G. Ferrari, C. De Lazzari, T. L. de Kroon, J. M. Elstrodt, G. Rakhorst, and Y. J. Gu. Numerical simulation of hemodynamic changes during beating-heart surgery: analysis of the effects of cardiac position alteration in an animal model. *Artif Organs*, 31(1):73–9, 2007.

[14] G. A. Giridharan, S. C. Koenig, M. Mitchell, M. Gartner, and G. M. Pantalos. A computer model of the pediatric circulatory system for testing pediatric assist devices. *Asaio J*, 53(1):74–81, 2007.

[15] E. Lanzarone, P. Liani, G. Baselli, and M. L. Costantino. Model of arterial tree and peripheral control for the study of physiological and assisted circulation. *Med Eng Phys*, 29(5):542–55, 2007.

[16] B. H. Maines and C. E. Brennen. Lumped parameter model for computing the minimum pressure during mechanical heart valve closure. *J Biomech Eng*, 127(4):648–55, 2005.

[17] K. Pekkan, D. Frakes, D. De Zelicourt, C. W. Lucas, W. J. Parks, and A. P. Yoganathan. Coupling pediatric ventricle assist devices to the fontan circulation: simulations with a lumped-parameter model. *Asaio J*, 51(5):618–28, 2005.

[18] V. Kecman. *State-space models of lumped and distributed systems*. Springer, Berlin etc., 1988.

[19] J. A. Negroni, E. C. Lascano, and R. H. Pichel. A computer study of the relation between chamber mechanical properties and mean pressure-mean flow of the left ventricle. *Circ Res*, 62(6):1121–33, 1988.

[20] P. H. Bovendeerd, P. Borsje, T. Arts, and F. N. van De Vosse. Dependence of intramyocardial pressure and coronary flow on ventricular loading and contractility: a model study. *Ann Biomed Eng*, 34(12):1833–45, 2006.

[21] Jos A. E. Spaan. *Coronary blood flow : mechanics, distribution, and control.* Developments in cardiovascular medicine ; v. 124. Kluwer Academic Publishers, Dordrecht; Boston, 1991.

[22] R. Beyar, D. Manor, and S. Sideman. Myocardial mechanics and coronary flow dynamics. In S. Sideman and R. Beyar, editors, *Interactive Phenomena in the Cardiac System*, volume 346 of *Advances in Experimental Medicine and Biology*, pages 125–136. Plenum Press, New York, 1993.

[23] J. Dankelman. *On the dynamics of the coronary circulation.* Delft, 1989.

[24] S. Sideman and R. Beyar. *Interactive phenomena in the cardiac system.* Advances in experimental medicine and biology ; v. 346. Plenum Press, New York, 1993.

[25] J.Jos A. E. SpaanA. Spaan. Coronary diastolic pressure-flow relation and zero flow pressure explained on the basis of intramyocardial compliance. *Circ Res*, 56(3):293–309, 1985.

[26] J. A. Spaan, A. J. Cornelissen, C. Chan, J. Dankelman, and F. C. Yin. Dynamics of flow, resistance, and intramural vascular volume in canine coronary circulation. *Am J Physiol Heart Circ Physiol*, 278(2):H383–403, 2000.

[27] B. Kaimovitz, Y. Lanir, and G. S. Kassab. Large-scale 3-d geometric reconstruction of the porcine coronary arterial vasculature based on detailed anatomical data. *Ann Biomed Eng*, 33(11):1517–35, 2005.

[28] J. I. Hoffman and J. A. Spaan. Pressure-flow relations in coronary circulation. *Physiol Rev*, 70(2):331–90, 1990.

[29] G. Fibich, N. L. Lanir, and M. Abovsky. Modeling of coronary capillary flow. In S. Sideman and R. Beyar, editors, *Interactive Phenomena in the Cardiac System*, volume 346 of *Advances in Experimental Medicine and Biology*, pages 137–150. Plenum Press, New York, 1993.

[30] P. A. Wieringa. *The Influence of the coronary capillary network on the distribution and control of local blood flow.* Delft, 1985.

[31] R. Beyar, R. Caminker, D. Manor, and S. Sideman. Coronary flow patterns in normal and ischemic hearts: transmyocardial and artery to vein distribution. *Ann Biomed Eng*, 21(4):435–58, 1993.

[32] M. S. Moayeri and G. R. Zendehbudi. Effects of elastic property of the wall on flow characteristics through arterial stenoses. *J Biomech*, 36(4):525–35, 2003.

[33] F. Kajiya, G. A. Klassen, J. A. Spaan, and J. I. E. Hoffmann. *Coronary circulation basic mechanism and clinical relevance.* Springer, Tokyo etc., 1990.

[34] S. Y. Rabbany, J. Y. Kresh, and A. Noordergraaf. Intramyocardial pressure: interaction of myocardial fluid pressure and fiber stress. *Am J Physiol*, 257(2 Pt 2):H357–64, 1989.

[35] N. P. Smith. A computational study of the interaction between coronary blood flow and myocardial mechanics. *Physiol Meas*, 25(4):863–77, 2004.

[36] J. Z. Wang, B. Tie, W. Welkowitz, J. Kostis, and J. Semmlow. Incremental network analogue model of the coronary artery. *Med Biol Eng Comput*, 27(4):416–22, 1989.

[37] D. Zinemanas, R. Beyar, and S. Sideman. Effects of myocardial contraction on coronary blood flow: an integrated model. *Ann Biomed Eng*, 22(6):638–52, 1994.

[38] D. Zinemanas, R. Beyar, and S. Sideman. Relating mechanics, blood flow and mass transport in the cardiac muscle. *International Journal of Heat and Mass Transfer*, 37:191–205, 1994.

[39] A. J. Cornelissen, J. Dankelman, E. VanBavel, and J. A. Spaan. Balance between myogenic, flow-dependent, and metabolic flow control in coronary arterial tree: a model study. *Am J Physiol Heart Circ Physiol*, 282(6):H2224–37, 2002.

[40] E. O. Feigl. Coronary physiology. *Physiol Rev*, 63(1):1–205, 1983.

[41] O.B. Garfein. *Current concepts in cardiovascular physiology.* Academic Press, San Diego ; London, 1990.

[42] T. Komaru, H. Kanatsuka, and K. Shirato. Coronary microcirculation: physiology and pharmacology. *Pharmacol Ther*, 86(3):217–61, 2000.

[43] B. D. Seeley and D. F. Young. Effect of geometry on pressure losses across models of arterial stenoses. *J Biomech*, 9(7):439–48, 1976.

[44] B. Berthier, R. Bouzerar, and C. Legallais. Blood flow patterns in an anatomically realistic coronary vessel: influence of three different reconstruction methods. *J Biomech*, 35(10):1347–56, 2002.

[45] V. Fuster. *Atherothrombosis and coronary artery disease*. Lippincott Williams & Wilkins, Philadelphia, 2nd edition, 2005.

[46] B. Quatember and H. Mühlthaler. Generation of cfd meshes from biplane angiograms: an example of image-based mesh generation and simulation. *Applied Numerical Mathematics*, 46(3-4):379–97, 2003.

[47] E. Shalman, M. Rosenfeld, E. Dgany, and S. Einav. Numerical modeling of the flow in stenosed coronary artery. the relationship between main hemodynamic parameters. *Comput Biol Med*, 32(5):329–44, 2002.

[48] W. Zhang, Y. Liu, and G. S. Kassab. Flow-induced shear strain in intima of porcine coronary arteries. *J Appl Physiol*, 103(2):587–93, 2007.

[49] M. Schrijver and C. Slump. Automatic segmentation of the coronary artery tree in angiographic projections. *Proceedings of BroRISC 2002, November 28-29, 2002, Veldhoven, cNetherlands*, pages 449–464, 2002.

[50] M. Mayr and B. Quatember. Segmentation of the coronary arteries in biplane angiograms based on a differential geometric approach. In F.Pistella Spitaleri and R. M., editors, *MASCOT05 - 5th Meeting on Applied Scientific Computing and Tools / Grid Generation, Approximation, Simulation and Visualization.*, volume 10 of *IMACS Series in Computational and Applied Mathematics*, pages 71–80, Lecce, Italy, 2005.

[51] B. Quatember, M. Mayr, and H. Mühlthaler. Clinical usefulness of a computational grid for diagnosis and planning therapy of coronary artery disease. In J. Volkert, T. Fahringer, D. Kranzlmüller, and W. Schreiner, editors, *1st Austrian Grid Symposium*, volume 210, pages 75–89, Schloss Hagenberg, Austria, 2006. Oesterreichische Computer Gesellschaft (books@ocg.at).

Bernhard Quatember
Innsbruck Medical University, Clinical Department of Radiology II,
Anichstrasse 35, 6020 Innsbruck,
Austria
e-mail: `Bernhard.Quatember@uibk.ac.at`

Martin Mayr
Innsbruck Medical University, Clinical Department of Radiology II,
Anichstrasse 35, 6020 Innsbruck,
Austria
e-mail: `Martin.Mayr@uibk.ac.at`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 195–210

Modelling Vaccine Protocols

Santo Motta, Pier-Luigi Lollini and Francesco Pappalardo

Abstract. Living organisms are natural complex systems where mathematical modelling may play a crucial role, since a model can be built with imperfect knowledge of some related phenomenon and model parameters (initial data, entities, relations between entities) can be adjusted to fit modelling results to experimental measurements. The model can then be used to understand the general behaviour of the phenomenon in different situations, to perform model experiments or simulations, to understand the role of single constituents and relations, to plan new experiments, or to test theoretical assumptions and suggest theory modifications. Modelling can therefore stimulate scientific creativity and produce better theoretical descriptions of the reality. We describe here our efforts to devise models of the immune system, and in particular the competition between immune defences and tumor cells. An agent-based model of the effects of a vaccine designed to prevent mammary carcinoma incidence in transgenic mice was developed. This model faithfully summarises not only the outcome of vaccination experiments, but also the dynamics of immune responses elicited by the vaccine. A genetic algorithm was used to drive the model and predict optimised vaccination schedules, which are currently being tested *in vivo*. The implications of biologic diversity on model development and perspectives to develop natural-scale models of the immune system are also discussed.

Mathematics Subject Classification (2000). Primary 99Z99; Secondary 00A00.

Keywords. Cancer, vaccine, model, simulator.

1. Introduction

A vaccine is an antigenic preparation used to establish immunity to a disease. Vaccines can be prophylactic (e.g., to prevent or ameliorate the effects of a future infection by any natural or “wild” pathogen) or therapeutic (e.g., vaccines against cancer). A vaccine against a particular bacterium or virus is relatively easy to create, since bacteria and viruses are foreign to the body and therefore antigens the immune system can recognise are expressed. Furthermore, there are usually only a few viable variants of a particular virus. However, when viruses like influenza or HIV continually mutate, it is much more difficult to develop appropriate vaccines.

The picture is also different for a cancer vaccine, which is often a process whereby a person's immune system is coaxed into recognising and destroying malignant cells without harming normal cells. Most cancer vaccines in development by pharmaceutical companies are therapeutic and address specific cancer types. A cancer vaccine is generally considered an immunotherapy, because it is not preventive and is only administered after cancerous cells have developed — unlike prophylactic vaccines against diseases such as polio, influenza, and tuberculosis. Moreover, a tumor can contain many different types of cells, each with different cell-surface antigens. Tumor cells are corrupted normal cells and therefore display few if any antigens that are foreign to an individual, which makes it difficult for the immune system to distinguish cancer cells from normal cells.

Cancer Immunotherapy is the use of the immune system to reject cancer. The main premise is that the patient's immune system may be stimulated to attack the malignant tumor cells. This can be either through immunisation of the patient, in which case the patient's immune system is trained to recognise tumor cells as targets to be destroyed, or through the administration of therapeutic antibodies as drugs that recruit the immune system to destroy the tumor cells. However, many kinds of tumor cell that arise as a result of the onset of cancer are more or less tolerated by the patient's immune system, since it responds to environmental factors encountered on the basis of discrimination between self and non-self. These tumor cells are essentially the patient's own cells that grow, divide and spread without proper regulatory control. On the other hand, many tumor cells do display unusual antigens that are either inappropriate for the cell type or its environment (or both), or are normally only present during development. Other tumor cells display cell surface receptors that are rare or absent on the surfaces of healthy cells, and which activate cellular signal transduction pathways that cause unregulated growth and division of such tumor cells.

Nevertheless, despite any recognition difficulty the immune system plays an active role in preventing tumor formation, as is evident from the study of genetically-modified models (GEM) of mice designed to lack immune responses. This simplified scenario shows that modelling the action of a cancer vaccine implies a model of the stimulated immune response — i.e., a model of how the immune system works. Our knowledge of the immune system (e.g., see [16]) is incomplete, but mathematical modelling can provide a better understanding of its underlying principles and organisation, which should ultimately help in the development of new treatments and therapies for various human diseases. The immune system appears to be a distributed system that lacks central control, but which nevertheless performs its complex task in an extremely effective and efficient way. Modelling such a complex system requires the application of knowledge and methodology from disciplines such as applied mathematics, physics and computer science.

In this paper, we review our efforts in approaching this topic, point out major open problems, and pose questions on directions for further investigation. The role of mathematical modelling in biology is discussed in §2. In §3 we describe our model for a cancer immunoprevention vaccine, and our first attempt to understand biological diversity using our simulator. In §4 we draw conclusions and consider future plans.

2. Modelling purposes and scales

2.1. Modelling purposes

A scientific endeavour often begins with the observation of natural phenomena, followed by a classification of the observed phenomena, mostly according to its morphological aspects. Thus one may know the entities taking part in a particular phenomenon but have little or no knowledge of the rules that regulate it, and so first formulate hypotheses and heuristic or qualitative theories to suggest how it can be described and explained. However, a quantitative theory is often needed to describe the observations or experimental results. Mathematical models are quantitative representations of phenomena built up in the framework of a theory using the language of mathematics, used in a broad sense to include approaches based on computer simulations. A mathematical model is appropriate only if a theory uses (and predicts) *measurable quantities* and gives relations between them — i.e., mere qualitative explanations of a phenomenon are not sufficient to construct a model.

The scientific method of course involves the general or qualified acceptance of a theory as long as it continues to explain the observed data and its predictions are verified. Nevertheless, one often faces the problem of how to proceed in scientific research given incomplete and imperfect knowledge. Theories and models are *ipso facto* imperfect representations of reality, and to keep models tractable we typically introduce simplifications and approximations when describing real phenomena. Furthermore, models need to be tested against experimental measurements, and discrepancies often suggest modifications of a theory or its underlying assumptions. Results from models may also suggest new experiments to verify the theory. The theory-model-experiment loop works clearly and efficiently when complex phenomena can be analysed in terms of simple rules and entities. An example is the modelling of complex electric circuits using the entities and rules of simple circuits. However, there are situations where complex phenomena cannot be studied by reduction to simpler ones — e.g., if the rules are not deterministic, or the phenomena have chaotic behaviour, or if collective effects play new and important roles. In the life sciences, it can be extremely difficult to isolate and study the behaviour of single constituents. Even when this is possible, many of the properties of living organisms are due to collective effects (populations) and in most cases these are not deterministic. Moreover, living organisms are the product of evolution. New mechanisms have been built up by nature in order to solve new problems, even though the old mechanism may persist and play a role in special situations. In this sense, redundancy is a feature of living organisms — i.e., a specific function can be analyzed by different parts of the system in order to recover system errors and malfunctioning.

Living organisms are natural complex systems. Modelling may therefore play a crucial role, since models can be built with imperfect knowledge of some phenomenon and the model parameters (initial data, entities, relations between entities) can be adjusted so that modelling results fit experimental measurements. Such models can then be used to understand the general behaviour of the phenomenon in different situations, perform "model experiments" or "simulations" to understand the role of single constituents and relations, to plan new experiments, or to test theoretical assumptions and suggest theory modifications. Modelling can therefore stimulate scientific creativity and produce better theoretical descriptions of the reality.

2.2. Modelling scales

A particular theory or model may describe natural phenomena on some given scale, for there is often a hierarchy of different scales. Choosing the scale may depend on which aspects of the phenomena, from *micro* to *macro*, one intends to represent. This is a well-known feature in physics, but in biology and immunology the definition of scale may be more ambiguous. A basic reference unit can be the cell, irrespective of its physical dimension, when one may define three basic scales — viz. the subcellular scale, the cellular scale and the macroscopic scale. Thus Bellomo *et al.* [4] proposed a classification for tumor evolution and its interaction with the immune system, which we adapt for our purposes as follows:

- the *microscopic* or *subcellular scale* refers to the main activities within the cells or at the cell membrane — e.g., genetic changes, distortion in the cell cycle and loss of apoptosis, expression and transduction of signals between cells, etc.
- the *cellular scale* refers to the main (interactive) activities of each cell, e.g., activation and proliferation of cells
- The *macroscopic scale* refers to phenomena which are typical of continuum systems. For instance: cell migration, convection, diffusion of antibodies.

Phenomena identified at one scale may be related to another scale. For instance, interactions developed at the cellular level are ruled by processes performed at the sub-cellular scale. Indeed, the immune system shows interesting phenomena on each scale, and phenomena on the different scales are related.

Theories and models developed at the microscopic or subcellular scale deal with evolution of the physical and biochemical state of a single cell. The evolution of a cell is regulated by genes contained in its nucleus. Receptors on the cell surface can receive signals transmitted to the cell nucleus, where the cell genes can be activated or repressed. Particular signals can be responsible for identical cell reproduction, or they can induce programmed cell death or apoptosis. Modelling the overall activity of a single cell is a very difficult problem, as many biological details are unclear or unknown. Biologists, mathematicians, physicists, computer scientists and engineers have combined to develop and use mathematical and computer science techniques in modelling sub-cellular phenomena. Many references can be found in PubMed and specialised symposia (e.g. [?]).

At the cellular scale, one is interested in the evolution of a system consisting of a large number of different cells. Cell interactions are regulated by signals emitted and received by cells through complex recognition processes. The connection with the sub-cellular scale is evident, but now one may ignore the details of single cell models and consider their outcome in the large system. This is analogous to what is done in modelling complex circuits, where the electronic component elements are replaced with equivalent circuits. The overall system may be described in a fashion familiar from statistical mechanics and the theory of cellular automata or gases and plasmas, where observable quantities are obtained by suitable moments derived from statistical distributions.

At the macroscopic level, one is interested in describing the dynamical behaviour of observable quantities — in most cases the densities of various entities — using techniques from the framework of continuum phenomenological theories. This is analogous to using Lotka–Volterra equations in population dynamics. Fitting the model parameters to the experimental data is always necessary to validate the model, which may involve ordinary or partial differential equations. Nonlinearity is an intrinsic feature, which leads to some sophisticated mathematical problems.

3. Modelling in immunology

3.1. Modelling the immune system

Models in immunology must take into account some general features of the immune system. The most relevant are uniqueness, distributed detection, imperfect detection and adaptability [10]. *Uniqueness* means that the immune system of each individual is unique and therefore vulnerabilities differ from one system to another. *Distributed detection* indicates that the small and efficient detectors used by the immune system are highly distributed, and not subject to centralised control or coordination. *Imperfect detection* means the immune system does not require the absolute detection of every pathogen, so the system is more flexible in allocating resources. *Anomaly detection* is the property of the immune system by which it can detect and react to pathogens that the body has never encountered before. *Adaptability* (or *learning and memory*) is the ability of the immune system to learn and remember the structures of pathogens, so that future responses to the pathogens can be much faster. These properties result in a system that is scalable, resilient to subversion, robust, very flexible, and which degrades gracefully.

The major problem that the immune system solves is distinguishing between self and non-self. Actually, the success of the immune system is more dependent on its ability to distinguish between harmful non-self and everything else. This is a difficult modelling problem, because the diversity of non-self patterns is much greater than self ones, the environment is highly distributed, the body must continue to function all the time, and resources are scarce. The immune system solves all this by using a multi-layered architecture of barriers — viz. physical (the skin), physiological (e.g., pH values), and the cells and molecules of the innate and acquired immune response.

Systemic models of immune responses have mainly been devoted to collective effects of various immune system constituents. These models do not study single cells or single molecules, but focus on cell interactions and collective behaviour in the initiation, control, and mounting of immune responses. Inside the scale framework, these models focus on cellular and macroscopic levels. The panorama of immune system models is quite broad. Nevertheless, all of these models are based on two biological theories underpinning our understanding of the immune system — viz. the clonal selection theory [5] and idiotypic network theory [11–13]. Nowadays, immunologists consider these two independent theories as mutually complementary and consistent. However, while clonal selection theory is believed to be the fundamental theory for understanding our modern knowledge of the immune system, the idiotypic network theory is believed to be correct as far as the existence of anti-idiotypic reactions is concerned but it is probably not relevant to controlling the immune response. Most macroscopic level models, also referred to as continuous models, have been formulated using the framework of both immunological theories [24, 25]. The cellular level models, also referred to as discrete models, are mostly based on the idiotypic network theory.

The main task of the immune system is to perform a pattern recognition, using cellular receptors to recognise target antigens. The binding mechanism, mostly unknown in detail, is based on different physical effects such as short range non-covalent interactions, hydrogen binding, van der Waals interactions, etc. [25]. A cellular receptor can recognise its target epitope if their surfaces have regions of extensive complementarity similar to the key and lock mechanism. Perelson and Oster [26] considered the constellation of features to be important in determining binding among molecules the *generalised shape* of the molecules.

Assuming this shape can be described by an η parameters, then a point in an η-dimensional space (*shape space*) specifies the generalised shape of a receptor-binding region, so they estimated that to be complete the receptor repertoire should satisfy the following conditions: (i) each receptor can recognise a set of related epitopes, each of which differs slightly in shape; (ii) the repertoire size is of the order of 10^6 or larger; and (iii) at least a subset of the repertoire size is distributed randomly throughout the shape space [25]. Farmer *et al.* [8] introduced binary strings to represent shapes of receptors and epitopes, which enabled the use of numerous readily available string matching algorithms that determine the degree of complementarity between strings. Discrete models of the immune system widely use this representation, to describe interactions between cell receptors and antigens. Continuous models use *affinity functions*, which globally represent the interactions between the cell population and the antigen population, and crucially determines the behaviour of the model.

3.2. Modelling tumor vaccines

The investigation of tumor immunity has led to many clinical attempts at curing human tumors (immunotherapy). Once a therapeutic agent has demonstrated its efficacy, it can be approved by regulatory agencies for routine use. An evaluation of the preclinical results of vaccines in mouse models shows a clear dichotomy between their therapeutic and prophylactic uses. In most instances, vaccination before the challenge (prophylactic vaccination) prevents tumor growth, whereas vaccination after the challenge (therapeutic vaccination) is much less effective.

Tumors are caused by a combination of exogenous and endogenous factors. Cancer immunoprevention is based on the use of immunological approaches to prevent solid tumors, rather than to cure cancer. This is mostly important in tumors caused by endogenous carcinogens, where cancer cells are continuously formed from corrupted normal cells. Cancer immunoprevention vaccines are based (like all vaccines) on components which give the immune system the necessary information to recognise tumor cells as harmful. Consequently, the cancer vaccine must be administered for an entire lifetime — i.e., the immune response induced by the vaccine must be maintained in a host for his entire life. Further, although a typical vaccine may not eliminate all tumor cells, it can stabilise them to a safe level. Moreover, a vaccine that is effective for a large population is very seldom optimal for a single individual, although efforts in this direction started a few years ago (e.g., see [6]). With regard to translation of cancer immunopreventive approaches to humans, it is desirable to minimise the number of vaccinations in a personalised schedule. In this paper, we report results from a first approach in the search for an optimal personalised schedule of a cancer immunoprevention vaccine.

The Triplex vaccine [7,14,19] was designed to improve the efficacy of existing immunopreventive treatments. A standard approach in oncology was adopted — viz. combining multiple immune signals in the same vaccine. The Triplex vaccine combines the target antigen with two "adjuvant" stimuli, interleukine 12 (IL-12) and allogeneic histocompatibility molecules (MHC). The main purpose of IL-12 is to enhance antigen presentation and helper T cell (Th) activation in response to the antigen. Allogeneic MHC molecules stimulate multiple T cell clones, and cause a broad production of immunostimulatory cytokines that amplify immune responses. A complete prevention of mammary carcinogenesis with the Triplex vaccine was obtained when vaccination cycles started at age 6 weeks of age and continued for the entire length of the experiment of about one year (Chronic vaccination) [7]. The major unresolved issue with the Triplex vaccine is whether or not the Chronic schedule is the minimal set of vaccinations to provide complete long-term protection from mammary carcinoma. Shorter vaccination protocols failed to prevent cancer — but between the Chronic and shorter protocols, there is a large set of schedules that might yield complete protection with significantly fewer vaccinations than the Chronic schedule. A large set of experiments to investigate this, each lasting one year, would be a feat to discourage any biological team from the pursuit of an experimental solution *in vivo*. In our modelling approach, we first developed a vaccine computational model that specifically addressed mammary cancer for a vaccine previously studied and tested *in vivo* by the cancer immunologists group at the University of Bologna, and then used the model to search for a better schedule than the Chronic one.

3.3. The model

To describe the cancer - immune system competition one needs to include all the entities (cells, molecules, adjuvants, etc.) that biologists recognise as relevant. In our case, the choice of entities was driven by the experimental data on the Triplex vaccine, where the relevant entities (the cells or molecules) have mechanical and biological states — viz. position, lifetime, internal states and specificity. Position and lifetime apply to all, but internal states only to cellular entities, and specificity to both cellular and molecular entities.

The Catania Mouse Model (CMM) described previously [18, 20] was implemented in a simulator called SimTriplex using a Lattice Boltzmann-like approach, including the entities (cells and molecules) of the adaptive and natural immune system, the Cancer and the vaccine. The immune system entities are B Cells (B), Antibody Secreting Plasma Cells (PLB), T-helper lymphocytes (TH), T-cytotoxic lymphocytes (TC), Macrophages (MP), Dendritic Cells (DC), Interleukin-2 (IL-2), Immunoglobulins (IgG), Danger Signal (D), Major Histocompatibility Complex Class I (MHCI), Major Histocompatibility Complex Class II (MHCII) Immunocomplexes (IC), Natural Killer Cells (NK). Cancer cells and Vaccine components are: Cancer Cells (CC), Tumor Associated Antigens (Ag), Vaccine Cells (VC), and Interleukin-12 (IL-12). All of the various classes of immune functional activity, phagocytosis, immune activation, opsonisation, infection and cytotoxicity are described using probability functions and translated into computational rules.

An interaction between two entities is a complex stochastic event, which may end with a state change of one or both entities. Interactions can be *specific* or *non-specific*. Specific interactions need a *recognition phase* between the two entities (e.g., B ↔ TAA). Recognition is based on the Hamming distance and affinity function, and is eventually enhanced by adjuvants. We refer to *positive interaction* when this first phase occurs successfully. Non-specific interactions do not have a recognition phase (e.g., DC ↔ TAA). When two entities that may interact lie in the same lattice site, they interact with a probabilistic law. Both specific and non-specific interactions are stochastically determined using a probability function, which depends upon different parameters computed via random number generators. Changing the seed of the random number generator yields a different sequence of probabilistic events. This simulates biological differences between individuals who share the same event probabilities. An overall scheme of the interactions included in the CMM is shown in Fig. 1, which shows how the vaccine interacts with the different components of the immune system and elicits cytotoxic (left side) and antibody (right side) responses that kill tumor cells.

After appropriate tuning of the model, the *in silico* simulations were able to reproduce the *in vivo* experiments using two independent sets of 100 different

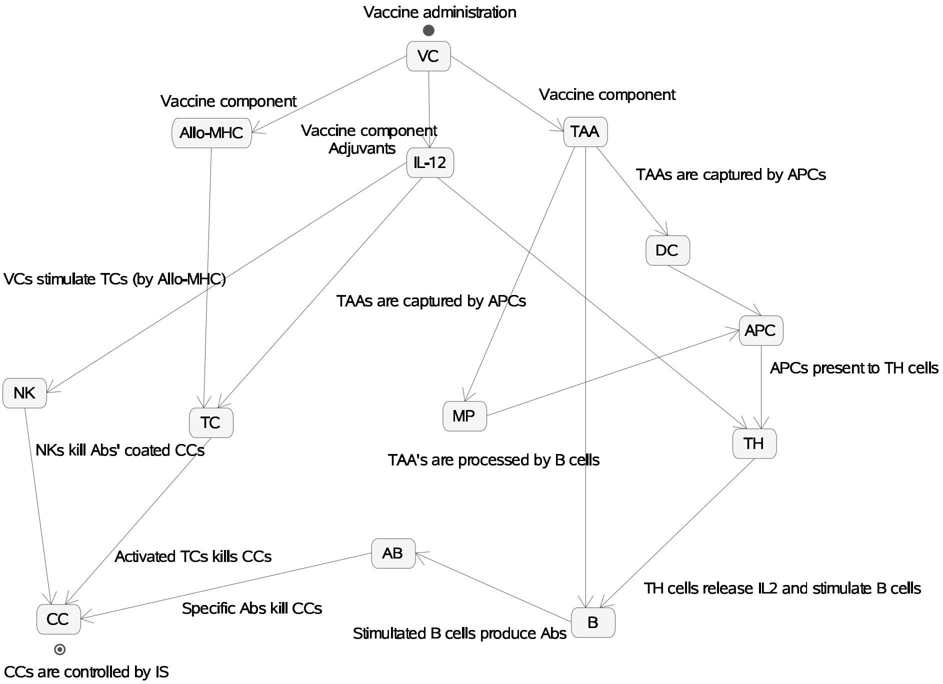


FIGURE 1. Interactions included in CMM model and simulator

virtual mice [20]. This result was achieved using a bitstring representation of length $l = 12$. The repertoire we represented was then $2^{12} = 4096$, which is very poor compared with the natural scale repertoire of $10^{16} \div 10^{18}$.

3.4. Search for an optimal schedule

When a newly designed vaccine is ready to be administered for the first time *in vivo*, whether to mice or to humans, the schedule is designed empirically — using a combination of immunological knowledge, vaccinological experience from previous endeavours, and practical constraints. In subsequent trials, the schedule of vaccinations is then refined on the basis of the protection elicited in the first batch of subjects and their immunological responses — e.g., kinetics of antibody titers, cell mediated response, etc. The problem of defining optimal schedules is particularly acute in cancer immunopreventive approaches like the Triplex vaccine, which requires a sequence of vaccine administrations to keep a high level of protective immunity against the continuing generation of cancer cells over very long periods, ideally for the entire lifetime of the host.

In searching for an optimal schedule, we tried different strategies. The first was "trial and error". We set successively repeating cycles of injections at different stages of virtual mouse age, and the simulator was used to determine the

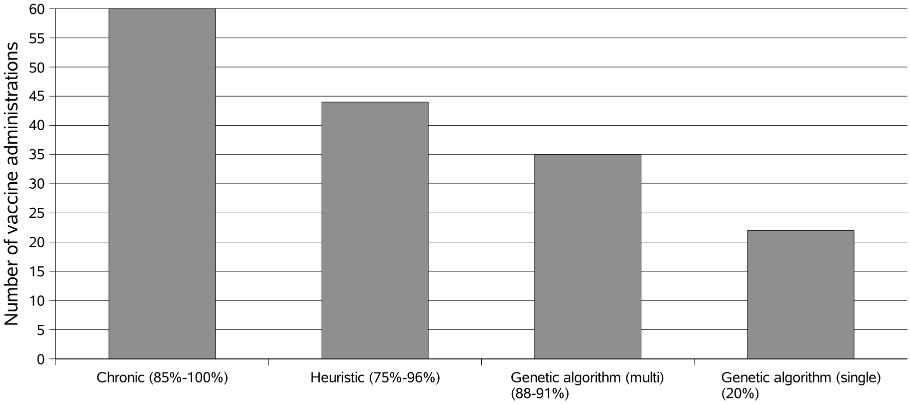


FIGURE 2. Progression in reducing the number of vaccine administration. The percentage of tumor-free mice at the end of the experiment is indicated in parenthesis.

survival of vaccinated mice. In this way, we found an effective schedule of only 44 vaccinations — i.e., 27% less than the standard Chronic protocol [18]. A second search strategy was based on genetic algorithms [17]. Attempts at using an unconstrained genetic algorithm on a single mouse [21] led to the conclusion that an effective schedule for that mouse does not protect against solid tumor formation in a large (∼83%) set of mice, due to biological diversity. (All biological experiments are affected by natural immunological variability, resulting from subtle individual variations in the generation of the immunological repertoire and interactions with environmental variables [7].) We then concluded that a genetic search should simultaneously take different simulated individuals into account, and consequently we were able to find a 35 injections schedule [15]. Figure 2 shows the number of vaccine administrations for different schedules, together with the percentage of survival of the individuals in the trial sets.

3.5. Searching for a personalised protocol

As already mentioned, a property of living organisms is biological diversity. This diversity originates from fundamental constituents of the organism like DNA sequences, and changes in the organism due to interactions with its environment. Consequently, each organism reacts in a different manner to an external stimulus like a drug or a vaccine. Personalised medicine has recently attracted the interest of many researchers. Most of these studies try to optimise a drug's efficacy by identifying the best dose and schedule in drug administration [2, 3]. Minimisation of drug toxicity is also an important goal.

A vaccination schedule is a series of vaccine administrations at different times. The ideal administration times and the number of administrations are peculiar to each individual organism, and the efficacy depends on both the time of administration and the number of administrations. In a protocol for the entire population, efficacy can be achieved only by inserting extra administrations, which somehow balance *errors* in administration times. Efficacy of the treatment for a single mouse can be achieved with roughly 1/3 less administrations than in the Chronic case (cf. Fig. 2). In translational research this implies that a human patient could receive only 1/3 less of the standard protocol, drastically reducing toxicity.

Our model mimics biological diversity using probability functions in the repertoire production and interaction events. A uniform probability density can be represented by a sequence of uniformly distributed random numbers. For a given random number generator, the sequence of random numbers is uniquely determined by the initial seed. By setting a seed for the repertoire and a seed for the interaction events, the model represents a single mouse and a specific history of its evolution. To mimic a population, we ran the simulator many times with different seeds. On the other hand, experiments on mice show that mice respond differently to the vaccine. We tried to understand the underlying reasons that make mice different, using numerical experiments. As the model reproduces the *in vivo* experiments, we assumed it would reproduce the behaviour of a population of mice just as well.

Then we tried to understand if environmental events are relevant in changing the mouse response to the vaccine. In [23] we randomly chose a mouse from our sample of 100 mice. We ran the genetic search for optimal protocol for this mouse, obtaining a 22 injections schedule. We used this protocol on the set of 100 mice, and found that 27% of the mice were tumor-free (TF) while the remaining 73% were not (NTF). We then proceeded to force all the mice in the sample to have the same interactions with the environmental variables, by setting the environmental seed equal for all mice. As a result, we obtained a set of TF mice of 30%. This result led to the conclusion that the environment was not critical. The result, which we plan to confirm with other numerical experiments, is not surprising because we are dealing with an endogenous tumor.

The difference should lie in the immunological repertoire. To highlight this, we considered two different mice — one belonging to the TF mice, and the other belonging to the NTF mice. Numerical results (cf. Fig. 3) show that the dynamic of specific B (against tumor associated antigen) and helper T lymphocytes (against peptide/major histocompatibility class II complex) have clear peaks in the case of the TF mouse, but the results are almost flat for the NTF mouse.

However, the difference was less evident for specific cytotoxic T lymphocytes, so we considered the cumulative number of B and T lymphocytes as functions of time — i.e., respectively

$$B_{cumulative}(t) = \int_0^t B(\tau) d\tau$$

and

$$T_{cumulative}(t) = \int_0^t T(\tau) d\tau.$$

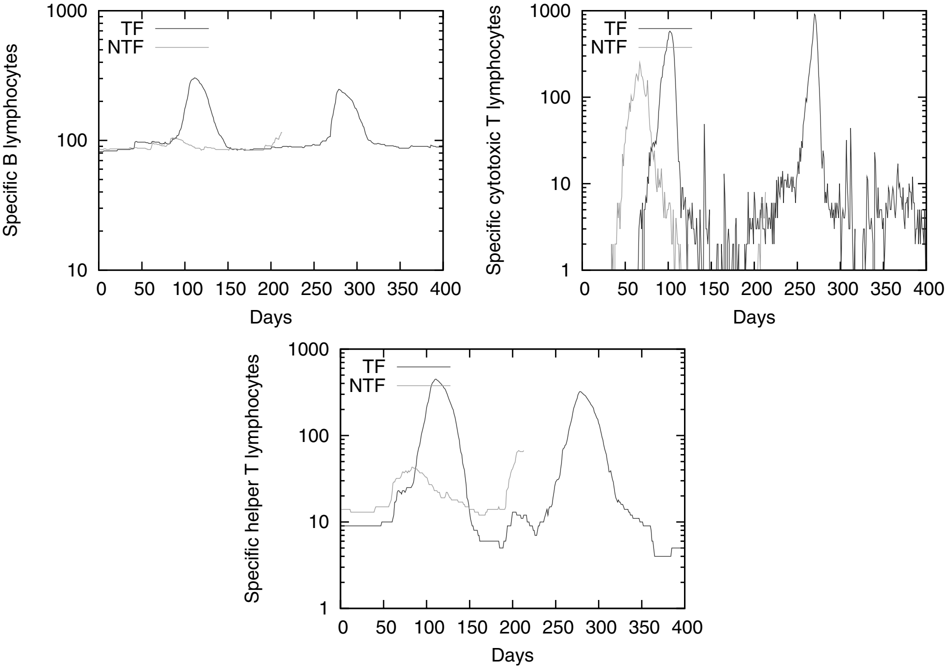


FIGURE 3. Specific B (against tumor associated antigen), specific cytotoxic T lymphocytes (against peptide/major histocompatibility class I complex) and specific helper T lymphocytes (against peptide/major histocompatibility class II complex) for TF and NTF mouse.

The plots in Fig. 4 show no initial significant differences in the repertoire, and the TF mouse shows evident antibody response later that is not present in the case of the NTF mouse.

These very preliminary investigations leave open the question of whether the individual response is related to the initial repertoire, and clearly suggest that the present version of our simulator is unable to detect any meaningful difference between TF and NTF mice. We believe there are two major reasons for this — viz. i) the diversity of modeled repertoire is too small with respect to the natural one, and ii) the number of entities considered in our model is not sufficient to allow a realistic representation of the expressed repertoire. Thus the model does not generate sufficient statistical data to detect any significant signal-to-noise ratio.

Extension of the repertoire to a more natural scale faces nontrivial computational problems. As already mentioned, the repertoire was modelled using a binary string of length l and interactions driven by the Hamming distance between the binding site of the two interacting entities. In the present version of the simulator, computation of the Hamming distance uses a pre-computed look-up table of 2^l entries. With $l = 12$ this occupies roughly 16 KB, which is easily allocated in the

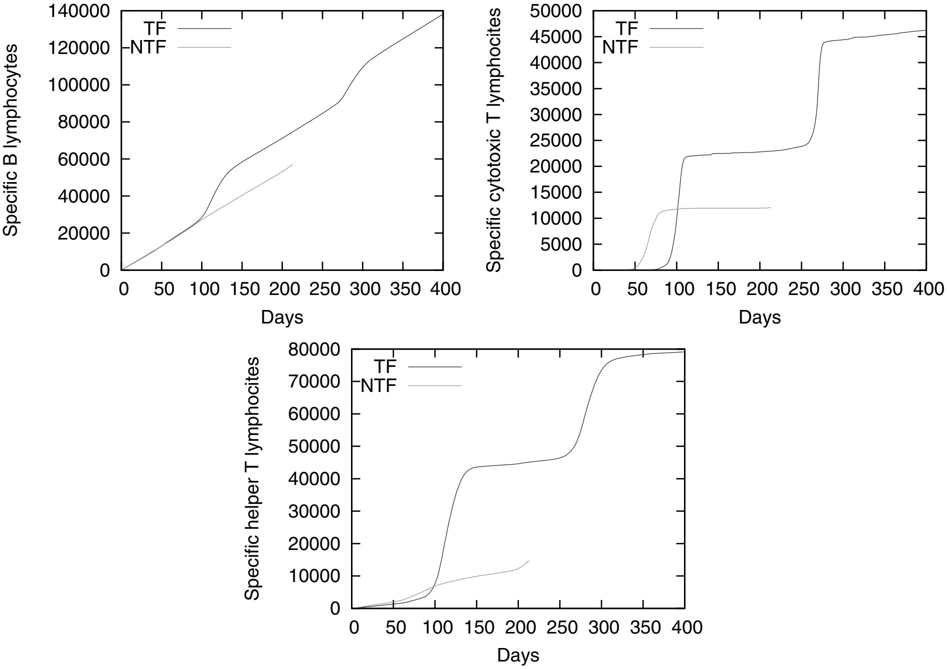


FIGURE 4. Cumulative behaviour for specific B lymphocytes (against tumor associated antigen), specific cytotoxic T lymphocytes (against peptide/major histocompatibility class I complex) and specific helper T lymphocytes (against peptide/major histocompatibility class II complex) for TF and NTF mouse.

cache-memory of any platform. However, a natural scale repertoire has a diversity of roughly $10^{16} \sim 2^{48}$, so the look-up table would have to be of the order of $4 \cdot 10^9$ KB, which cannot be handled in any present or near-future computer platform. The look-up table would need to be accessed from an external storage, leading to an unacceptable decline in the performance of the simulator, so a different approach to compute the Hamming distance that keeps the competitive performance of the look-up table is needed.

In [22] we have shown that an alternative, but still efficient, method relies on *binary magic numbers* [9]. The best bit counting method takes only 12 operations and does not depend on bitstring length. This method avoids the memory and potential cache misses of the look-up table method. The counts of bits set in the bytes is done in parallel, and the sum total of the bits set in the bytes is computed by multiplying and shifting right bits, using binary magic numbers. Magic numbers have unusual or special properties in certain calculations. With binary magic numbers, it is possible to write algorithms that are typically faster by a factor of $N/log_2(N)$ than more obvious ones. The binary magic numbers come in a sequence and a recursive template, where the N-th number is itself an infinite binary number from right to left of 2^N ones followed by 2^N zeroes, followed by 2^N ones, and so on. Often one just uses these patterns, but occasionally also the inverse (complement) reverse pattern, and we believe that this approach should solve the problem of raising diversity to a natural scale without losing computational performance. Nevertheless, the computational solution shall require parallel platforms to run the simulator which implies re-coding the simulator using parallelisation libraries. This work is in progress.

4. Conclusions

The immune system is a natural complex system. Modelling such a system is a challenge requiring multidisciplinary contributions. Mathematical and computational methods play an important role. We discussed a modelling investigation in the field of Artificial Immunity — viz. activation of the immune response induced by a vaccine, and a first attempt to understand the immunological difference between different individuals. This problem is still under investigation.

Even if our model can reproduce experimental results, it is very naive. An important further step would be to include models at different scales, from the subcellular scale to organs and functionalities. This goal may not be achieved at the present stage of knowledge, but it is possible to use different modelling tools to create a modelling environment to help provide an integrated approach. Considerable effort is presently being devoted to achieve this goal in a European Community funded project (*ImmunoGRid: The Virtual Human Immune System*), and to show that mathematical and computational modelling can provide valuable benefits in biology and immunology.

Acknowledgment

This work was supported under the EC contract FP6-2004-IST-4, No. 028069 (ImmunoGrid).

FP and SM acknowledge partial support from University of Catania research grant and MIUR PRIN 2006.

PLL acknowledges financial support from the University of Bologna, the Department of Experimental Pathology (“Pallotti” fund), MIUR and the Italian Association for Cancer Research (AIRC).

FP and SM would like to thank A. Torrisi of the University of Catania Library Service for his help in finding a copy of Freed’s paper.

References

[1] 2nd International Symposium on Computational Cell Biology (2003).
http : //www.nrcam.uchc.edu/2nd_symposium/main.html, Cited 16 January 2007.
http : //www.nrcam.uchc.edu/news/symposium.html, Cited 16 January 2007.

[2] Z. Agur (2006). *Biomathematics in the development of personalised medicine in oncology.* Future Oncol. **2(1)**, 39–42.

[3] Z. Agur, R. Hassin, S. Levy. *Optimizing chemotherapy scheduling using local search heuristics.* Operations Research **54(5)** (2006).

[4] N. Bellomo, L. Preziosi, *Modelling and mathematical problems related to tumor evolution and its interaction with the Immune System.* Math. Comp. Model. **32** (2000), 413–452.

[5] F. Burnet, *T*he Clonal Selection Theory of Acquired Immunity. Vanderbilt University, Nashville TN 1959.

[6] S. Croci, G. Nicoletti, L. Landuzzi, C. De Giovanni, A. Astolfi, C. Marini, E. Di Carlo, P. Musiani, G. Forni, P. Nanni, P.-L. Lollini. *Immunological Prevention of a Multigene Cancer Syndrome.* Cancer Research **64**, (2004) 8428–8434.

[7] C. De Giovanni, G. Nicoletti, L. Landuzzi, A. Astolfi, S. Croci, A. Comes, S. Ferrini, R. Meazza, M. Iezzi, E. Di Carlo, P. Musiani, F. Cavallo, P. Nanni, P.-L. Lollini, *Immunoprevention of HER-2/neu transgenic mammary carcinoma through an interleukin 12-engineered allogeneic cell vaccine.* Cancer Research **64(11)**, (2004) 4001–4009.

[8] J. D. Farmer, N. Packard, A. S. Perelson, *The immune system, adaption and machine learning.* Phisica D, **22** (1986) 187–204.

[9] E. E. Freed, *Binary Magic Numbers—Some Applications and Algorithms.* Dr Dobb's Journal **8:4** (1983), 24–37.

[10] S. A. Hofmeyr, *An Overview of the Immune System*, (1997) *http://www.cs.unm.edu/˜immsec/html-imm/immune-system.html.* Cited January 21st, 2007.

[11] N. K. Jerne, *The Immune System.* Scientific American **229(1)** (1973), 52–60.

[12] N. K. Jerne, *Towards a Network Theory of The Immune System.* Ann. Immunol. (Ist. Pasteur) **125C** (1974), 373–389.

[13] J. R. Josephson and S. G. Josephson, *Abductive Inference*, Cambridge University Press, 1994

[14] P.-L. Lollini, C. De Giovanni, L. Landuzzi, G. Nicoletti, F. Frabetti, F. Cavallo, M. Giovarelli, G. Forni, A. Modica, A. Modesti, P. Musiani, P. Nanni, *Transduction of genes coding for a histocompatibility (MHC) antigen and for its physiological inducer interferon-gamma in the same cell: efficient MHC expression and inhibition of tumor and metastasis growth.* Hum. Gene Ther. **6(6)** (1995), 743–752.

[15] P.-L. Lollini, S. Motta, F. Pappalardo, *Discovery of cancer vaccination protocols with a genetic algorithm driving an agent based simulator.* BMC Bioinformatics **7:352** (2006), doi:10.1186/1471-2105-7-352.

[16] P.-L. Lollini, S. Motta, F. Pappalardo, *Modeling tumor immonology.* Mathematical Models and Methods in Applied Sciences, **16:supp01** (2006), 1091–1124.

[17] M. Mitchell, *An introduction to generic algorithms.* MIT Press, 1996.

[18] S. Motta, P.-L. Lollini, F. Castiglione, F. Pappalardo, *Modelling Vaccination Schedules for a Cancer Immunoprevention Vaccine.* Immunome Research **1(5)** (2005), doi:10.1186/1745-7580-1-5.

[19] P. Nanni, G. Nicoletti, C. De Giovanni, L. Landuzzi, E. Di Carlo, F. Cavallo, S. M. Pupa, I. Rossi, M. P. Colombo, C. Ricci, A. Astolfi, P. Musiani, G. Forni, P.-L. Lollini, *Combined allogeneic tumor cell vaccination and systemic interleukin 12 prevents mammary carcinogenesis in HER-2/neu transgenic mice.* J. Exp. Med. **194(9)** (2001), 1195–1205.

[20] F. Pappalardo, P.-L. Lollini, F. Castiglione, S. Motta, *Modelling and Simulation of Cancer Immunoprevention vaccine.* Bioinformatics **21(12)** (2005), 2891–2897.

[21] F. Pappalardo, E. Mastriani, P.-L. Lollini, S. Motta, *Genetic Algorithm against Cancer.* Lectures Notes in Computer Science **3849** (2006), 223–228.

[22] F. Pappalardo, C. Calonaci, P.-L. Lollini, E. Mastriani, M. Pennisi, E. Rossi, S. Motta, *Computational effort and natural scale immune system repertoire* Preprint, 2007.

[23] M. Pennisi, E. Mastriani, F. Pappalardo, P.-L. Lollini, S. Motta, *Toward a personalized schedule with Triplex vaccine* Preprint, 2007.

[24] A. S. Perelson, *T*heoretical Immunology, Part One & Two. SFI Studies in the Sciences of Complexity, Addison Wesley, 1988.

[25] A. S. Perelson, G. Weisbuch, *Immunology for physicists.* Rev. Mod. Phys. **69** (1997), 1219–1267.

[26] A. S. Perelson, G. F. Oster, *Theoretical studies on clonal selection: Minimal antibody repertoire size and reliability of self-nonself discrimination.* J. Theor. Biol. **81** (1979), 645–670.

Santo Motta
Dipartimento di Mathematica e Informatica
V.le A. Doria, 6
95125 Catania
Italy
e-mail: motta@dmi.unict.it

Pier-Luigi Lollini
Laboratory of Immunology and Biology of Metastasis
Cancer Research Section, Department of Experimental Pathology
Viale Filopanti 22, I-40126 Bologna
Italy
e-mail: pierluigi.lollini@unibo.it

Francesco Pappalardo
Dipartimento di Mathematica e Informatica
V.le A. Doria, 6
95125 Catania
Italy
e-mail: francesco@dmi.unict.it

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 211–227

Modelling the Response of Intracranial Pressure to Microgravity Environments

William D. Lakin and Scott A. Stevens

Abstract. A majority of astronauts experience symptoms of headache, vomiting, nausea, lethargy, and gastric discomfort during the first few hours or days after entering a microgravity environment. It has been hypothesised that some of these symtoms are related to the development of benign intracranial hypertension as a result of the cephalic fluid shifts and relative venous congestion that occur in microgravity. This hypothesis is tested here using a mathematical model of lumped-parameter type that embeds the intracranial system in whole-body physiology. In addition to considering microgravity environments, this model is used to examine the response of intracranial pressures to head-down tilt (HDT), a ground-based experimental procedure often used to simulate the cardiovascular effects of microgravity. Predicted pressures in these simulations include those in the cerebral vasculature, ventricular and extra-ventricular cerebrospinal fluid (CSF), and the brain tissue extracellular fluid. Various cardiovascular stimuli associated with microgravity, including changes in arterial pressure, central venous pressure, and blood colloid osmotic pressure, are considered both individually and in concert. Small alterations of the blood-brain barrier in space due to factors such as gravitational unloading and increased exposure to radiation are also allowed. Simulation results predict that in a healthy individual the upward fluid shifts and changes in central venous pressure in microgravity cannot, by themselves, produce a large elevation in ICP so long as the blood-brain barrier remains intact. Indeed, in this case the simulations suggest that ICP in microgravity is significantly less than that in long-term HDT, and may even be less than that in the supine position on Earth. However, simulations predict that ICP can increase significantly if, combined with a drop in blood colloid osmotic pressure, there is even a slight reduction in the integrity of the blood-brain barrier. These results suggest that in some otherwise healthy individuals, microgravity environments may elevate ICP to levels associated with benign intracranial hypertension, producing symptoms that can adversely affect crew performance.

Mathematics Subject Classification (2000). Primary 93A30, 76Z05; Secondary 92C10 , 92C50.

Keywords. Lumped-parameter mathematical model, Intracranial pressure, Microgravity, Blood-brain barrier.

1. Introduction

Lumped-parameter models represent an attractive method for examining pressure dynamics involving complicated human physiology. In this modelling approach, the physiological system is subdivided into a number of linked, interacting subunits termed "compartments". In general, each compartment will contain a single physical constituent, such as blood, cerebrospinal fluid (CSF), or tissue and interstitial fluid. However, depending on the model's complexity, a given constituent may appear in more than one compartment of the model. Dynamics in each compartment is specified by lumped time-dependent functions giving compartmental pressures, while incremental changes in flows and compartmental volumes are obtained by associating resistance and compliance parameters with adjacent compartments. In particular, interaction between adjacent subunits is assumed to take place at the interfaces of the model's compartments. Spatial resolution is limited by the number of defined compartments, but models of this type with even a relatively small number of compartments often produce excellent agreement with clinical data.

Lumped-parameter models have been used to study intracranial pressure for more than 200 years. However, with few exceptions, previous models have adopted restrictions known as the "Kellie–Monro Doctrine" to reduce complexity. The Kellie-Monro framework considers the intracranial system to be completely enclosed within the intracranial vault, which is assumed to be rigid. A specified inflow of blood to the intracranial arteries provides a forcing for the system, and outflow from the Jugular Bulb is assumed to instantaneously equate to this inflow. These restrictions yield a closed system with constant total volume. Strictly intracranial models have produced a number of important results that illuminate the mechanisms of intracranial pressure adjustments in situations involving both normal and pathophysiology. However, the ability of these closed-system models to incorporate the influence of important extracranial factors on intracranial pressure dynamics is clearly limited. For example, the important buffering effects of the spinal CSF space on intracranial pressure cannot be directly included.

During the first few hours or days of space flight, a collection of symptoms — including headache, vomiting, nausea, lethargy and gastric discomfort — often affects a majority of crew members. Although evidence suggests that some of these symptoms are a form of motion sickness associated with the functioning of the vestibular and sensory systems during exposure to microgravity environments, other observed symptoms differ in significant ways from terrestrial manifestations of motion sickness (TMS). While TMS is roughly coincident with the initiation of motion, symptoms in microgravity are delayed. They involve active, rather than passive, head motions and do not involve cold sweats, salivation, facial pallor, or bowel urgency. Also, unlike TMS, symptoms in microgravity are usually associated with headache, and vomiting tends to be sudden without initial retching or nausea.

It has been hypothesised [6] that elevated intracranial pressure (ICP) may play a role in the development of some of the above symptoms in microgravity.

Indeed, several symptoms, such as headache and sudden vomiting, strongly resemble those associated with benign intracranial hypertension. Further, the fact that the onset of these symptoms occurs early in a flight suggests that they may be related to the well-documented fluid shift that begins at launch and continues for approximately 10 hours into flight. During this period, between 1500ml and 2000ml of fluid may shift from the lower to the upper body, with 90% of this shift taking place within the first two and a half hours [22]. Central venous pressure also decreases in microgravity, although the basis for this finding is measured data from one astronaut over a 9-hour period [1].

This paper employs a mathematical model of lumped-parameter type that has been developed [20] to study the effects of microgravity and its ground-based clinical analogues on intracranial pressure (ICP). The Kellie–Monro Doctrine has been revoked in this model. By embedding the intracranial system in whole-body physiology, as opposed to confining it within the cranial vault, the present model allows consistent inclusion of the shift of cephalic fluid from the lower to the upper body observed in microgravity.

The cardiovascular effects of microgravity are often simulated by ground-based head down tilt (HDT) procedures. Clinical data exists for intracranial pressure (ICP) during short-term head down tilt [13], and comparisons between this data and the results of simulations have been used to validate the present model. All of the participants exposed to 8 hours of the 5 degree HDT procedures in [4] reported headache and other mild symptoms. However, little corresponding data exists for ICP's during long-term exposure to either HDT or microgravity. The headache and nausea suffered by astronauts in microgravity are not reported by those exposed to extended 5 or 6 degree HDT in [14, 15], where a shift of cephalic fluid similar to that in microgravity is speculated to occur. There is thus some question about the extent to which long-term HDT effectively simulates the response of ICP to microgravity.

Although the present simulations provide insights into the possible cardiovascular effects of long-term HDT, the primary focus of this modelling effort is a study of causal relationships between microgravity effects and intracranial pressures. Stimuli considered will include microgravity-induced changes in arterial pressure, central venous pressure, and blood colloid osmotic pressure. From this analysis, levels of intracranial pressure may be quantified with respect to potential changes in the cardiovascular system due to microgravity. Results of this study suggest that the upward fluid shifts in microgravity and associated changes in the cardiovascular system cannot, by themselves, elevate intracranial pressure to levels associated with benign intracranial hypertension. ICP is found to be unaffected by the expected changes in arterial pressure and to change in parallel with central venous pressure. A moderate increase is predicted due to the expected drop in colloid osmotic pressure in microgravity. However, even with all of these stimuli combined, ICP in microgravity is predicted to be less than that predicted for long-term HDT, and it may even be less than experienced on Earth in the supine position. Consequently, if benign intracranial hypertension does develop in space, the cause must be a factor not normally included in simulations assuming healthy human physiology.

The integrity of the blood-brain barrier, which inhibits the movement of large molecules into the brain from the cerebral capillaries, is related to the tightness of the junctions between adjacent endothelial cells in the walls of the cerebral capillary bed. Models of intracranial pressure dynamics usually consider deficits in the integrity of the blood brain barrier only in simulations involving pathology or trauma. However, even for healthy individuals, there are factors in microgravity, such as gravitational unloading of the capillary walls and increased exposure to radiation in space beyond the shield provided by the Earth's atmosphere, which may influence the tightness of these junctions. The present simulations indicate that ICP can increase significantly if, combined with a drop in blood colloid osmotic pressure, there is even a slight reduction in the integrity of the blood-brain barrier. These results suggest that in some otherwise healthy individuals microgravity environments may elevate ICP and produce symptoms associated with benign intracranial hypertension.

2. The lumped-parameter model

The lumped-parameter model used for this study is a simplified variant of the extensive whole-body model introduced by Lakin *et al.* [9] to study intracranial pressure dynamics. These simplifications allow steady-state solutions for the pressure fields to be obtained algebraically in closed form, thus preserving the accuracy of the model over the full expected range of pressures. In the present model, the intracranial portion of the larger model in [9] is preserved nearly intact. Additionally, filtration from the intracranial capillaries is now modelled with a Starling–Landis equation as opposed to the traditional hydrodynamic version of Ohm's law used in electrical circuit analogies [12]. This allows changes in colloid osmotic pressure to affect capillary filtration and absorption.

Consistent with previous models of this type [7, 18], the intracranial region is divided into interacting subunits termed "compartments." As depicted in Figure 1, the intracranial region is divided into six compartments. Three compartments are vascular: intracranial arteries (I); capillaries (C), including the choroid plexus; and the venous sinus (S). Two compartments involve cerebrospinal fluid (CSF): ventricular CSF (F); and extraventricular CSF (T). The latter compartment (T) includes both the subarachnoid CSF and CSF in the spinal theca. It thus extends beyond the intracranial region and provides a bridge between intracranial and whole-body physiology. The brain compartment (B) represents brain tissue and interstitial fluid. The model contains three additional compartments modelling strictly extra-cranial physiology: central arteries (A); central veins (V); and the thoracic space (Y). Pressures in each compartment are given in mmHg and denoted by a P, with a subscript indicating the compartment. For example, P_F represents the spatially-averaged (lumped) ventricular CSF pressure. Fluid flow or filtration between compartments is given in ml/min and denoted by a Q with an ordered-pair subscript indicating the direction of flow. For example, Q_{IC} represents blood flow from the intracranial arteries to the capillary bed, and Q_{CB} represents the fluid filtration from the capillaries into the brain tissue.

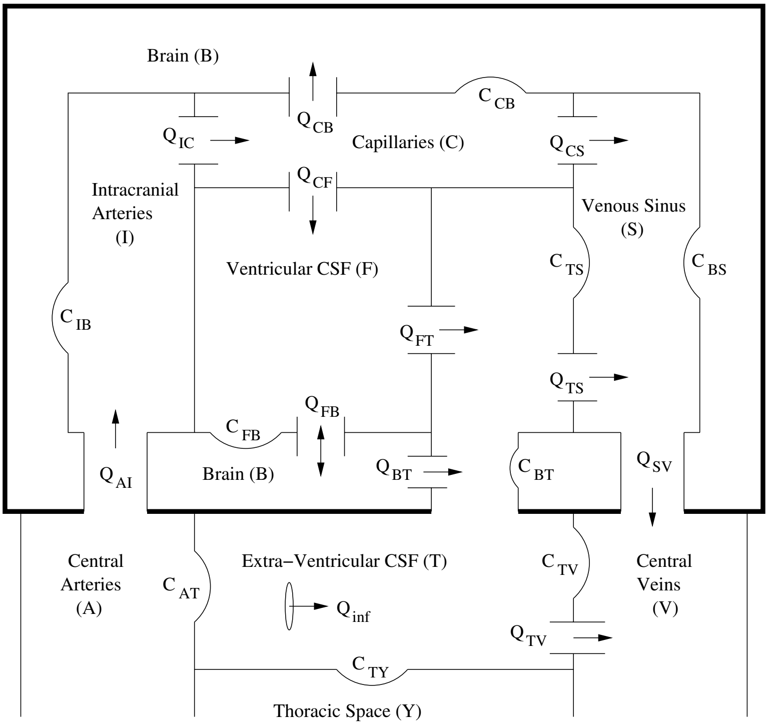


FIGURE 1. The lumped-parameter model. The dark line represents the rigid cranial wall, Q_{ij} represents fluid flow from compartment i to compartment j, arrows indicate the customary direction of flow, Q_{inf} represents an infusion rate of CSF, and C_{ij} represents a distensible surface between compartments i and j.

Assumptions

The following basic assumptions lead to the time-dependent differential equations that describe the pressure dynamics of this system.

- All fluids are considered incompressible and isothermal.
- The regulation of cerebral blood flow (Q_{IC}) and CSF production by the choroid plexus (Q_{CF}) over a full range of intracranial pressures is described in [9]. For the pressure ranges considered here, these regulation mechanisms remain robust, so for simplicity constant rates will be assumed for these two flows.

- Fluid filtration across the blood-brain barrier (Q_{CB}) is modelled by the Starling–Landis equation
$$Q_{CB} = K_{CB}\ [(P_C - P_B) - \sigma_{CB}(\pi_C - \pi_B)]\ , \tag{1}$$
where P_C is the capillary pressure, P_B is the brain interstitial fluid pressure, π_C is the blood colloid osmotic pressure, π_B is the colloid osmotic pressure of the brain interstitial fluid, K_{CB} is the filtration coefficient and σ_{CB} is the reflection coefficient. The osmolality of the interstitial brain tissue fluid and the blood plasma are assumed to be equal [16], so the only osmotic forces considered in the model of this filtration are those due to differences in protein concentrations.
- All other flows are related to pressure differences by the hydrodynamic version of Ohm's law
$$Q_{ij} = \frac{P_i - P_j}{R_{ij}} = Z_{ij}(P_i - P_j)\,, \tag{2}$$
where Q_{ij} is the flow from compartment i to compartment j, P_i and P_j are the spatially-averaged pressures of compartments i and j respectively, R_{ij} is the lumped resistance ($R_{ij} = -R_{ji}$), and Z_{ij} is the fluidity (inverse of R_{ij}). Equation (2) is altered to accommodate position changes by taking
$$\begin{aligned} Q_{SV} &= Z_{SV}\,(P_S - P_V + G_{SV}\sin\theta)\,, & (3)\\ Q_{AI} &= Z_{AI}\,(P_A - P_I - G_{AI}\sin\theta)\,, & (4)\end{aligned}$$
where θ is the angle of head tilt with up being positive, and G represents the gravity-induced hydrostatic pressure exerted by the column of fluid between the respective compartments. These equations represent a simplified form of Bernoulli's equation subject to the conditions of this model [5]. The forces due to a change in the gravitational direction are applied only between central and intracranial vascular compartments where the vertical column of fluid between compartments is significant during a change in position. This is in agreement with the viewpoint that alterations in CSF pressure due to position change are primarily induced by the resulting change in intracranial blood pressures [3].
- The deformation of the membrane between adjacent compartments is a function of the change in pressure difference between these compartments, so
$$\frac{dV_{ij}}{dt} = C_{ij}\,\frac{d}{dt}\,[P_i - P_j] = C_{ij}\,\left(\frac{dP_i}{dt} - \frac{dP_j}{dt}\right)\,, \tag{5}$$
where V_{ij} denotes the volume of the *cup* formed at the interface of compartments i and j, and $C_{ij} = C_{ji}$ denotes the local compliance between the two compartments [19].

Governing equations

Applying the law of mass conservation in the five strictly intracranial compartments (I,C,F,S,B) and the bridging compartment (T) of the present model results in a set of six differential equations. For example, the differential equation for the CSF (F) compartment is

$$Q_{CF} - Z_{FB}\,(P_F - P_B) - Z_{FT}\,(P_F - P_T) = C_{FB}\left(\frac{dP_F}{dt} - \frac{dP_B}{dt}\right). \tag{6}$$

Treating the pressures in these six compartments as the dependent variables, the resulting system of equations in matrix form is

$$\mathbf{C}\frac{d\mathbf{P}}{dt} + \mathbf{ZP} = \mathbf{Q}, \tag{7}$$

where the vectors **P** and **Q** are

$$\mathbf{P} = \begin{bmatrix} P_I \\ P_C \\ P_S \\ P_F \\ P_B \\ P_T \end{bmatrix} \text{ and } \mathbf{Q} = \begin{bmatrix} Z_{AI}\,(P_A - G_{AI}\sin\theta) - Q_{IC} \\ Q_{IC} - Q_{CF} + K_{CB}\,\sigma_{CB}(\pi_C - \pi_B) \\ Z_{SV}\,(P_V - G_{SV}\sin\theta) \\ Q_{CF} \\ -K_{CB}\,\sigma_{CB}(\pi_C - \pi_B) \\ C_{AT}\frac{dP_A}{dt} + C_{TV}\frac{dP_V}{dt} + C_{TY}\frac{dP_Y}{dt} + Q_{inf} + Z_{TV}P_V \end{bmatrix} \tag{8}$$

and the fluidity matrix (**Z**) and compliance matrix (**C**) are

$$\mathbf{Z} = \begin{bmatrix} Z_{AI} & 0 & 0 & 0 & 0 & 0 \\ 0 & K_{CB} + Z_{CS} & -Z_{CS} & 0 & -K_{CB} & 0 \\ 0 & -Z_{CS} & Z_{CS,SV,TS} & 0 & 0 & -Z_{TS} \\ 0 & 0 & 0 & Z_{FB,FT} & -Z_{FB} & -Z_{FT} \\ 0 & -K_{CB} & 0 & -Z_{FB} & K_{CB} + Z_{BT,FB} & -Z_{BT} \\ 0 & 0 & -Z_{TS} & -Z_{FT} & -Z_{BT} & Z_{BT,FT,TS,TV} \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} C_{IB} & 0 & 0 & 0 & -C_{IB} & 0 \\ 0 & C_{CB} & 0 & 0 & -C_{CB} & 0 \\ 0 & 0 & C_{BS,TS} & 0 & -C_{BS} & -C_{TS} \\ 0 & 0 & 0 & C_{FB} & -C_{FB} & 0 \\ -C_{IB} & -C_{CB} & -C_{BS} & -C_{FB} & C_{BS,BT,CB,FB,IB} & -C_{BT} \\ 0 & 0 & -C_{TS} & 0 & -C_{BT} & C_{AT,BT,TS,TV,TY} \end{bmatrix}.$$

Repeated subscripts in any entry in these matrices denotes summation — e.g., $Z_{BT,FB} = Z_{BT} + Z_{FB}$. Equation (7) may appear to be linear, but it is nonlinear if even one entry in the matrices **Z** or **C** depends on a component in the pressure vector **P**.

If the oscillatory effects of the forcing terms in **Q** are subtracted, the solution of equation (7) is a set of time-dependent pressures averaged over each cardiac cycle. However, as shown in [20], properties of the matrices **C** and **Z** imply that a solution **P** of equation (7) always tends to a unique steady state solution $\mathbf{P}^*$. For this steady state, time derivatives in (7) are identically zero and hence

$$\mathbf{P}^* = \mathbf{Z}^{-1}\mathbf{Q}^*, \tag{9}$$

where $\mathbf{Q}^*$ denotes the matrix obtained from **Q** in (8) by setting all derivative terms to zero and replacing P_A, P_V, π_C, and π_B by ${P_A}^*$, ${P_V}^*$, π_C^*, and π_B^* respectively. Thus in the steady-state the set of coupled differential equations is replaced by a set of coupled algebraic equations. Analysis of these steady state solutions is the principal focus of this paper. All of the simulations performed here involve the solution of equation (9). Because of the complexity of $\mathbf{Z^{-1}}$, closed form solutions were obtained with the aid of the mathematical software package Mathematica.

Base-state calibrations

Before analyzing changes from a base state due to various stimuli, it is necessary to approximate normal mean values for all dependent and independent variables as well as obtain scale values for model parameters such as fluidity, filtration, and reflection coefficients. Whenever possible, these starting values have been obtained from available clinical data. However, in the case of other variables and parameters, such as the filtration coefficient K_{CB} and the base state pressures $\overline{P}_T$ and $\overline{P}_S$, it is necessary to estimate base and scale values from model calibration simulations by achieving consistency between model predictions and clinically observed steady-state results of constant-rate CSF infusions. Constant-rate CSF infusion procedures are simulated by incorporating an infusion term Q_{inf} into equation (9). This calibration process is presented in full detail in [20]. Initial and calibrated values used in the numerical simulations are given in Table 1.

3. Simulation methods

3.1. Head-down tilt simulations

Short term HDT is simulated by introducing a negative angle θ in equation (9). Additionally, clinically observed changes in central vascular pressures are incorporated into P_A^* and P_V^* in this equation. These results are then compared to the clinical observations of Katkov and Chestukhin [8] for venous sinus pressure, and the observations of Murthy *et al.* [13] for intracranial pressure. The values of G_{SV} and G_{AI} representing the gravity-induced hydrostatic force exerted by the column of fluid between the central and intracranial compartments are based on the distance between the right atrium and the base of the brain being 28.8 cm [8]. This results in $G_{SV} = G_{AI} = 22.232$ mmHg. In clinical studies, a steady-state response was measured after several minutes. Therefore, blood colloid osmotic pressure is assumed to remain unchanged during the current simulations. Furthermore, the jugular vein pressures provided by Katkov and Chestukhin [8] are considered to be indicative of the venous-sinus compartment pressure as the pathway between these compartments should remain unrestricted during supine and head down tilts.

During extended HDT, blood colloid osmotic pressure drops [15], which in turn alters the normal forces involved in fluid filtration at the cerebral capillary level. Thus, long term HDT is simulated in a manner similar to short-term HDT, except that blood colloid osmotic pressure drops by 3.3 mmHg after 4 hours of 6 degree HDT [15]. This is incorporated into the model by setting $\pi_C^* = \overline{\pi}_C - 3.3$ in equation (9). In this case, no clinical data exists for the intracranial response.

Table 1. Initial and calibrated values used in the numerical simulations

symbol	starting value	units	description
$\overline{P}_A$	92.0	mmHg	central artery pressure
$\overline{P}_I$	82.0	mmHg	intracranial artery pressure
$\overline{P}_C$	34.1546	mmHg	capillary pressure
$\overline{P}_S$	7.82	mmHg	venous sinus pressure
$\overline{P}_V$	5.4	mmHg	central vein pressure
$\overline{P}_F$	11.2	mmHg	ventricular CSF pressure
$\overline{P}_B$	11.2	mmHg	brain pressure
$\overline{P}_T$	11.0	mmHg	extra-ventricular CSF pressure
$\overline{Q}_{AI}$	1035.0	ml/min	cerebral blood flow (arteries)
$\overline{Q}_{IC}$	1035.0	ml/min	cerebral blood flow (capillaries)
$\overline{Q}_F$	.4278	ml/min	total CSF formation rate
$\overline{Q}_{CF}$	0.2995	ml/min	CSF formation from choroid plexus
$\overline{Q}_{BT}$	0.1283	ml/min	other CSF formation
$\overline{Q}_{CB}$	0.1283	ml/min	filtration across blood brain barrier
$\overline{Q}_{FB}$	0	ml/min	net flow between ventricle and brain
$\overline{Q}_{FT}$	0.2995	ml/min	transmantle CSF flow
$\overline{Q}_{CS}$	1034.5712	ml/min	blood flow from capillaries to sinuses
$\overline{Q}_{TS}$	0.3209	ml/min	CSF absorption into the venous sinuses
$\overline{Q}_{TV}$	0.1069	ml/min	extracranial CSF absorption
$\overline{\sigma}_{CB}$	1	none	blood-brain barrier reflection coefficient
$\overline{K}_{CB}$	0.0665	(ml/min)/mmHg	blood-brain barrier filtration coefficient
$\overline{Z}_{FB}$	66.50	(ml/min)/mmHg	fluidity between ventricular CSF and brain
$\overline{Z}_{ij}$	$\overline{Q}_{ij}/(\overline{P}_i - \overline{P}_j)$	(ml/min)/mmHg	all other fluidities

3.2. Microgravity simulations

Since there is a greater fluid shift away from the dependent limbs in microgravity than in HDT [21], microgravity simulations are performed by altering blood plasma colloid osmotic pressures to a greater extent than that observed during head-down tilt. If the blood colloid osmotic pressure drops 3.3 mmHg [15] with an 800 ml reduction in dependent limb volume [14] during HDT, it is estimated from the data of [4, 14] that an 1800 ml reduction in dependent limb volume [21] during microgravity should result in a plasma colloid osmotic pressure drop of no more than 6.3 mmHg. The limited amount of data available for humans in microgravity suggests that mean central artery pressure remains unchanged (4 astronauts) [2] and that central venous pressure drops to about 0 mmHg (one astronaut) [1]. These changes are incorporated into the model by setting $P_V^* = 0$, $P_A^* = \overline{P}_A$ and $\pi_C^* = \overline{\pi}_C - 6.3$, in equation (9). Again, in this case no clinical data exists for the intracranial response.

3.3. The blood-brain barrier in microgravity

There is no data available for the response of a healthy blood-brain barrier to immersion in a microgravity environment. The integrity of the blood-brain barrier is related to the tightness of the junctions between adjacent endothelial cells in the walls of the intracranial capillary vessels. Several components contribute to maintaining the tightness of these junctions on Earth. A major role is certainly played by adhesion between the endothelial cells, and this component will not be affected by alterations in gravity. However, gravitational unloading of body tissues and fluids, one of the most pervasive changes caused by a microgravity environment, may have an ability to alter the integrity of the blood-brain barrier. Exposure to higher levels of radiation in space beyond the shield of the Earth's atmosphere may also affect the normal volumes of the endothelial cells in a way that will reduce tightness.

Gravitational unloading

A measure of tightness may be obtained by examining the extent of the overlap region and the space between the endothelial cells in the walls of the cerebral capillary vessels. Due to the rigidity of the capillary basement membrane, the capillary volume remains nearly constant, and hence the extent of the overlap region will also remain nearly constant, even in the face of increased capillary pressure. A key component of the tightness of this junction is therefore the space between overlapping cells. This space will be termed the junction gap. In normal gravity, closure of the junction gap is maintained through the joint action of the interior capillary pressure and the external interstitial fluid pressure, which augment adhesion of the endothelial cells. In terms of pressures in the model, the pressure ϕ_T that acts to keep the gap closed can be written as the sum $\phi_T = P_B + P_C$, where P_B is the interstitial fluid pressure of the brain and P_C is the intracranial capillary pressure.

Since the brain is mostly fluid, it is capable of transmitting hydrostatic pressure to the capillaries. Thus, in normal gravity, hydrostatic forces act to augment ϕ_T. This augmentation is depicted in the top illustration of Figure 2. However, in microgravity no such hydrostatic pressure is present and so this augmentation is removed. This, in turn, reduces ϕ_T and, as depicted by the bottom illustration in Figure 2, the tightness of the junctions may be reduced. This reasoning is consistent with observations that for general capillary-tissue barriers in the rest of the body much more fluid exits the leg tissues in actual microgravity conditions as opposed to either the supine or head-down tilt (HDT) positions in ground-based experiments [22].

Radiation effects

Beyond Low Earth Orbit, the protection of the Earth's atmosphere against radiation is no longer available, and protective mechanisms such as increased shielding are necessary in order to protect crew members in space against adverse radiation effects. Even with current countermeasures, considerable concern still remains about radiation health issues.

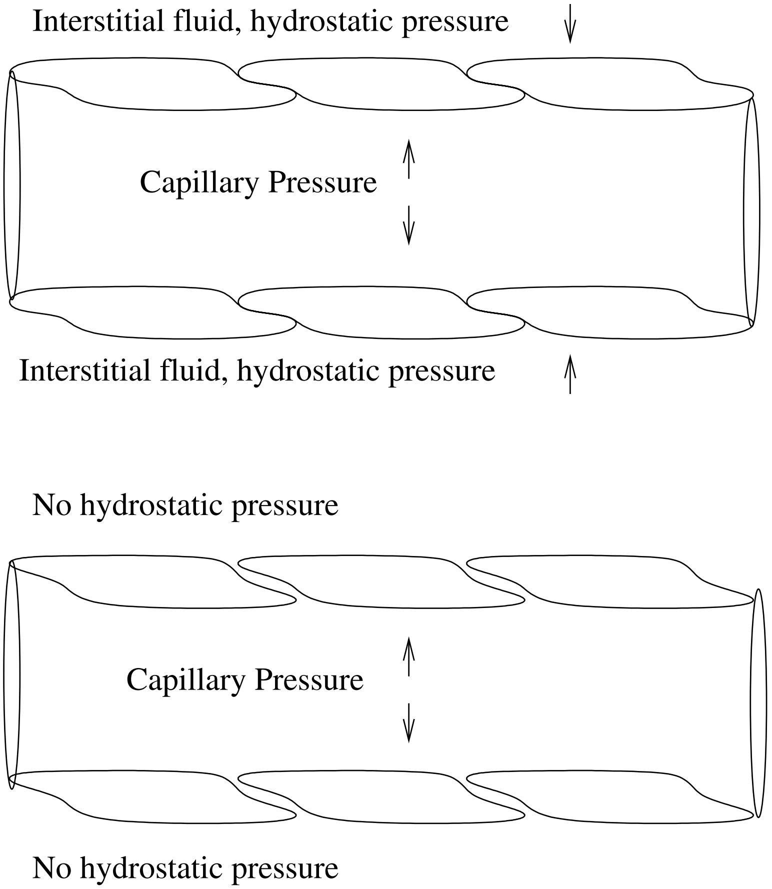


FIGURE 2. The upper illustration represents the capillary endothelial cell arrangement with gravity induced hydrostatic pressures augmenting cell adhesion. The lower illustration represents the possible reduction in tightness of the endothelial cell junctions in microgravity when the augmenting force is eliminated.

Recent experiments on Earth by Leszczynski *et al.* [10, 11] involving cell phone radiation demonstrate the potential effect that exposure to even small amounts of radiation in space can have on the blood-brain barrier. They reported that the mobile phone radiation activated non-thermal transient changes in the protein expression levels of hsp27 and p38MAPK in human endothelial cells. It is hypothesised in [10] that activation of hsp27 may cause an increase in blood-brain barrier permeability through stabilisation of endothelial cell stress fibres. Increased protein activity may even cause the endothelial cells themselves to shrink – lessening their volume, widening the junction gap, and reducing the overlap region.

Consequently, radiation exposure in space appears capable of adversely impacting the integrity of the blood brain barrier.

Modelling the integrity of the blood-brain barrier

If gravitational unloading, exposure to increased radiation beyond Low Earth Orbit, or some other feature of a microgravity environment induces slight changes in the blood-brain barrier, with a Starling–Landis model for the flow Q_{CB} the filtration coefficient K_{CB} and the reflection coefficient σ_{CB} in (1) can be adjusted to reflect these changes. In particular, a decrease in the tightness of the blood-brain barrier due to a change from the Earth-bound supine position to microgravity may be modelled by an increase in the filtration coefficient and a decrease in the reflection coefficient.

In performing the simulations, equation (9) is solved for the steady-state pressures P_F^*, P_B^*, and P_S^* in terms of P_A^*, P_V^*, π_C^*, and the tilt angle θ. These are described in terms of changes from the base state by

$$\begin{aligned} P_F^* - \overline{P}_F = \Delta P_F &= (P_V^* - P_V) - 0.37\,(\pi_C^* - \overline{\pi}_C) - 0.90\,G_{SV}\ \sin(\theta)\,, && (10)\\ P_B^* - \overline{P}_B = \Delta P_B &= (P_V^* - P_V) - 0.37\,(\pi_C^* - \overline{\pi}_C) - 0.90\,G_{SV}\ \sin(\theta)\,, && (11)\\ P_S^* - \overline{P}_S = \Delta P_S &= (P_V^* - P_V) - G_{SV}\sin(\theta)\,. && (12) \end{aligned}$$

These equations form the basis for both HDT ($\theta \neq 0$) and microgravity ($\theta = 0$) simulation results. It should be noted that due to the strict regulation of cerebral blood flow in this model, the pressure changes above are unaffected by changes in central artery pressure. In the more elaborate model in [9], where large changes in central artery pressure are expected, this may not be the case. However, since central artery pressure remains relatively constant in short term HDT [8, 17], long-term HDT [14] and microgravity [2], there is no need to incorporate such large changes into these simulations.

4. Simulation results

Validation of the mathematical model is necessary before it can consistently be used to simulate the effects of long-term HDT and microgravity environments where clinical data is unavailable for comparison. In the present work, model validation is provided by a comparison of the model's predictions for short-term HDT and clinical data.

4.1. Short-term head-down tilt

The results of short-term HDT simulations as well as clinical results are given in Table 2, taken from [20], where changes from the base state are displayed. Here, $\Delta(P_S - P_V) = (P_S^* - P_V^*) - (\overline{P}_S - \overline{P}_V)$. This approach is adopted to accommodate differences in the base state values between the model and clinical data. Also in this table, ΔP_{ICP} represents both $P_F^* - \overline{P}_F$ and $P_B^* - \overline{P}_B$, as these differences are identical in short-term head-down tilt simulations. In the row corresponding

TABLE 2. Summarised data from Katkov and Chestukhin [8] and Murthy [13] and model results. Pressures are in mmHg and angles are in degrees.

	$\theta = -6$	$\theta = -10$	$\theta = -15$	$\theta = -30$	$\theta = -75$
Ref. [8] Data $\Delta(P_S - P_V)$	–	3.1	–	11	20.9
Model $\Delta(P_S - P_V)$	–	3.86	–	11.11	21.47
Ref. [13] Data ΔP_{ICP}	3.3	–	6.1	–	–
Model ΔP_{ICP}	2.10 to 3.70	–	5.18 to 7.78	–	–

to the model's ΔP_{ICP}, a range is given that represents possible values, depending on how central venous pressure changes during the HDT procedure. Since central venous pressure was not measured by Murthy [13], a range of values for central venous pressure was based on the data from [8]. This resulted in the range for ΔP_{ICP} presented in the table.

The agreement displayed in Table 2 between measured data and model predictions validates the present lumped-parameter model as a vehicle for studying HDT and microgravity. Agreement between model simulations and clinical data further indicates that maintaining model resistances constant is a valid assumption for predicting steady-state responses in the supine and HDT positions.

4.2. Long-term head-down tilt

The results of long term head-down tilt simulations for intracranial pressures can be derived explicitly from equations (10) and (11). Specifically, ICP's increase in parallel with central venous pressure, and increase approximately 0.37 mmHg for each 1 mmHg drop in blood colloid osmotic pressure. Thus if central venous pressure increases by 1.6 mmHg [8, 14] and blood colloid osmotic pressure drops by 3.3 mmHg [15] during extended 6^o HDT, it can be expected that ICP will increase by about 4.9 mmHg.

4.3. Microgravity environment with blood-brain barrier fully intact

The results of microgravity simulations for intracranial pressures in this case may also be explicitly derived from equations (10) and (11), where θ is now zero. If the blood colloid osmotic pressure drops by 6.3 mmHg, ICP's increase by 2.3 mmHg in addition to the changes in central venous pressure. Therefore, if central venous pressure drops to zero during microgravity [1] then the ICP's in microgravity should fall to aproximately 3 mmHg below the base-state value in the supine position on Earth. Consequently, so long as the blood-brain barrier remains fully intact, even with all stimuli active, the model simulations do not predict that ICP will be elevated by the cephalic fluid shift in microgravity, much less approach levels associated with benign intracranial hypertension.

4.4. Microgravity environment with blood-brain barrier effects

If small alterations in the integrity of the blood-brain barrier due to factors in microgravity are allowed, results of the present simulations are shown in Figure 3, which gives level curves in terms of the parameters that characterise the tightness of the blood-brain barrier. Normal ICP is defined as a resting value below 15 mmHg, and abnormal ICP is considered to be a value greater than 18.35 mmHg [23]. Figure 3 shows that for a halving of the reflection coefficient and a doubling of the filtration coefficient, the simulation predicts a brain pressure of 19 mmHg. In absolute terms, these are small changes in the two coefficients. Even so, simulation results predict a significantly elevated ICP that is within the symptomatic range for benign intracranial hypertension.

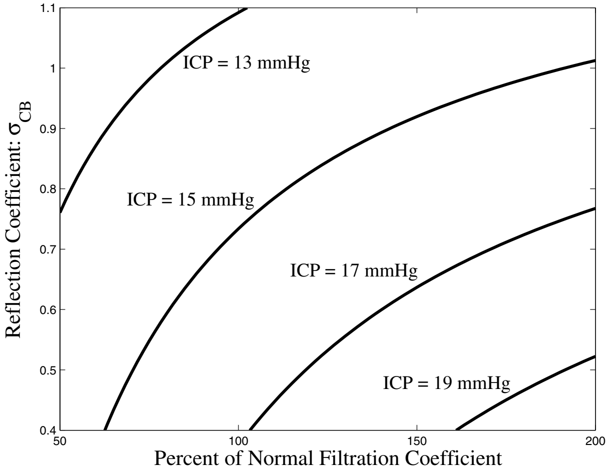


FIGURE 3. Level curves for brain pressure in the filtration coefficient-reflection coefficient plane.

5. Conclusions

The model developed in this paper is designed to accurately reflect the steady-state pressures of the intracranial system in response to various stimuli associated with microgravity and its ground-based clinical analogues. Emphasis has been placed on steady-state pressures as it can be shown that convergence to such a steady-state is guaranteed, regardless of the value or nonlinear nature of the compliances [20]. Currently, this guaranteed convergence has only been proven when the resistances, or fluidities, remain constant. Therefore this assumption is made for all fluidities except the fluidity between the intracranial arteries and capillaries. In this exception, the fluidity is inversely proportional to the pressure difference between these compartments, resulting in a constant cerebral blood flow. All other fluidities, as well as the capillary filtration coefficient, are assumed constant. Comparisons to clinical CSF infusion tests validate this assumption at the capillary/venule level, and comparisons to clinical HDT procedures validate this assumption at the venous sinus/jugular level.

The primary concern of this work is to test the hypothesis [22, 6] that the headaches and nausea experienced by astronauts in space are caused by induced benign intracranial hypertension. Since there is a greater fluid shift away from the dependent limbs in microgravity than in HDT [21], it is to be expected that the resulting decrease in blood colloid osmotic pressure will be greater than that observed in HDT. Calculations suggest that in this case, blood colloid osmotic pressure should not drop by more than 6.3 mmHg. If central venous pressure remains unchanged this results in an increase of ICP by about 2.3 mmHg. However, if central venous pressure drops to approximately zero [1], this reduces ICP to approximately 3 mmHg below the Earth-bound supine position. The simulations therefore suggest that, without some other cardiovascular stimuli in addition to the effects of both the cephalic fluid shift and the relative venous congestion due to the lack of a gravity assist in the venous return, it is probable that ICP in microgravity is significantly less than that in HDT and may even be less than that in the supine position on Earth. The sensitivity analyses detailed in [20] shows that this conclusion is valid over a wide range of values for model parameters that require indirect estimation due to a lack of clinical data. Furthermore, this conclusion is independent of the numerical values or nonlinear nature of the compliance terms. Comparisons to clinical data further suggest that the assumed linear relationship between pressure and flow is valid in both the supine and HDT position.

The model simulations that lead to the above conclusions for changes in ICP assumed that an otherwise healthy blood-brain barrier remains intact in microgravity. Factors in microgravity such as gravitational unloading and radiation effects appear capable of affecting the integrity of the blood-brain barrier. Changes in the integrity of the blood-brain barrier are included in the present model by adjusting the filtration and reflection coefficients in the Starling–Landis equation (1). The present simulations indicate that intracranial pressure can increase significantly if, combined with a drop in blood colloid osmotic pressure, there is even a slight reduction in the integrity of the blood-brain barrier. Indeed, the simulations predict that changes in the filtration and reflection coefficients that are small in absolute terms can produce a significantly elevated ICP that is within the symptomatic range for benign intracranial hypertension. Thus although the dramatic upward fluid shifts in microgravity are not predicted by themselves to elevate intracranial pressure to symptomatic levels [20], the present results predict that some symptoms associated with benign intracranial hypertension may be produced if some aspect of microgravity slightly affects the tightness of the blood-brain barrier.

No data is available for the response of a healthy blood-brain barrier to immersion in a microgravity environment. The degree to which various mechanisms affect the tightness of the blood-brain barrier in microgravity can be expected to vary among otherwise healthy individuals. As seen in Figure 3 lesser reductions of the tightness of the blood-brain barrier are predicted to produce ICPs that, while elevated, fall below the abnormal range. This may help explain why some astronauts experience symptoms related to elevated ICP while others remain symptom free.

Acknowledgment

Portions of this work were supported by NASA under cooperative agreement NCC5-581.

References

[1] J.C. Buckey, F.A. Gaffney, L.D. Lane, B.D. Levine, D.E. Watenpaugh, and C.G. Blomqvist. Central venous pressure in space. *N. Engl. J. Med.*, 328(25):1853–1854, 1993.

[2] J.F. Cox and K. Tahvanainen. Influence of microgravity on astronaut's sympathetic and vagal responses to valsalva's manoeuvre. *Journal of Physiology*, 538(1):309–320, 2002.

[3] R. A. Fishman. *Cerebrospinal Fluid in Diseases of the Nervous System.* W.B. Saunders Company, Philadelphia, PA, second edition, 1992.

[4] A.R. Hargens. Fluid shifts in vascular and extravascular spaces during and after simulated weightlessness. *Med. Sci. Sprots Exerc.*, 15(5):421–427, 1983.

[5] D. Jaron, T.W. Moore, and J. Bai. Cardiovascular response to acceleration stress: a computer simulation. *Proceeding of the IEEE*, 76(6):700–707, 1988.

[6] T. Jennings. Space adaptation syndrome is caused by elevated intracranial pressure. *Med. Hypothesis*, 32(4):289–291, August 1990.

[7] Z. Karni, J. Bear, S. Sorek, and Z. Pinczewki. A quasi-steady state compartmental model of intracranial fluid dynamics. *Med. Biol. Engng. Comput.*, 25:167–172, 1987.

[8] V.E. Katkov and V.V. Chestukhin. Blood pressures and oxygenation in different cardiovascular compartments of a normal man during postural exposures. *Aviat Space Environ Med*, 51(11):1234–1242, 1980.

[9] W.D. Lakin, S.A. Stevens, B.I. Tranmer, and P.L. Penar. A whole-body mathematical model for intracranial pressure dynamics. *J. Math. Biol.*, 46:347–383, 2003.

[10] D. Leszczynski, S. Joenvaara, J. Reivinen, and R. Kuokka. Non-thermal activation of the hsp27/p38mapk stress pathway by mobile phone radiation in human endothelial cells: Molecular mechanism for cancer- and blood-brain barrier-related effects. *Differentiation*, 70:120–129, 2002.

[11] D. Leszczynski, R. Nylund, S. Joenvaara, and J. Reivinen. Applicability of discovery science approach to determine biological effects of mobile phone radiation. *Proteomics*, 4:426–431, 2004.

[12] A. Marmarou, K. Shulman, and R.M. Rosende. A nonlinear analysis of the cerebrospinal fluid system and intracranial pressure dynamics. *J. Neurosurg.*, 48:332–344, 1978.

[13] G. Murthy, J. Marchbanks, D. E. Watepaugh, J. U. Meyer, N. E. Eliashberg, and A. R. Hargens. Increased intracranial pressure in humans during simulated microgravity. *The Physiologist*, 35(1):S184–S185, 1992.

[14] J. V. Nixon, R. G. Murray, C. Bryant, R. L. Johnson, J. H. Mitchell, O. B. Holland, C Gomez-Sanchez, P Vergne-Marini, and CG Blomqvist. Early cardiovascular adaptation to simulated zero gravity. *J. Appl. Physiol.: Respirat. Environ. Exercise Physiol.*, 46(3):541–548, 1979.

[15] S. E. Parazynski, A. R. Hargens, B. Tucker, M. Aratow, J. Styf, and A. Crenshaw. Transcapillary fluid shifts in tissues of the head and neck during and after simulated microgravity. *J. Appl. Physiol.*, 71(6):2469–2475, 1991.

[16] SI Rapoport. *Blood-brain barrier in physiology and medicine.* Raven Press, New York, NY, 1976.

[17] J. J. Smith, C. V. Hughes, M. J. Ptacin, J. A. Barney, F. E. Tristani, and T. J. Ebert. The effect of age on hemodynamic response to graded postural stress in normal men. *Journal of Gerontology*, 42(4):406–411, 1987.

[18] S. Sorek, J. Bear, and Z. Karni. A non-steady compartmental flow model of the cerebrovascular system. *J. Biomechanics*, 21:695–704, 1988.

[19] S. A. Stevens and W. D. Lakin. Local compliance effects on the global csf pressure-volume relationship in models of intracranial pressure dynamics. *Mathematical and Computer Modelling of Dynamical Systems*, 6(4):445–465, 2001.

[20] S. A. Stevens, W. D. Lakin, and P. L. Penar. Modelling steady-state intracranial pressures in supine, head-down tilt, and microgravity conditions. *Aviat Space Environ Med*, 76:329-338, 2005.

[21] W. E. Thornton, T. P. Moore, and S. L. Pool. Fluid shifts in weightlessness. *Aviat. Space Environ. Med.*, 58:A86–A90, 1987.

[22] W. E. Thornton, T. P. Moore, S. L. Pool, and J. Vanderploeg. Clinical characterization and etiology of space motion sickness. *Aviat. Space Environ. Med.*, 58:A1–A8, 1987.

[23] M. T. Torbey, R. G. Geocadin, A. Y. Razumovsky, D. Rigamonti, and M. A. Williams. Utility of csf pressure monitering to identify idiopathic intracranial hypertension without papilledema in patients with chronic daily headache. *Cephalalgia*, 24:495–502, 2004.

William D. Lakin
Department of Mathematics and Statistics
University of Vermont
Burlington, VT
USA
e-mail: `wlakin@together.net`

Scott A. Stevens
Division of Information Technology and Sciences
Champlain College
Burlington, VT
USA
e-mail: `sas56@psu.edu`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 229–234
© 2008 Birkhäuser Verlag Basel/Switzerland

"Noisy Oncology": Some Caveats in using Gaussian Noise in Mathematical Models of Chemotherapy

Alberto d'Onofrio

Abstract. This article discusses some paradoxical results that arise when modelling uncertainties in models of anti-tumor chemotherapies using Gaussian noise. The effects of intrinsic and environmental perturbations and uncertainties on the dynamics of tumor growth and anti-tumor chemotherapy delivered via continuous infusion are considered.

Mathematics Subject Classification (2000). Primary 60H10; Secondary 92C50.

Keywords. Tumor, Chemotherapy, Stochastic differential equations, Stochastic bifurcations.

1. Introduction

Initially, the growth of a tumor is dominated by uncontrolled mythosis [1], which is a phase of exponential growth. Then as the size of the tumor increases, the nutrients available become insufficient to satisfy all of the cells, so they must compete for nutrients and the tumor growth is no longer exponential but tends to plateau — i.e., the growth curve approaches a horizontal asymptote. In practice, only an *in vitro* tumor nears this final steady state, since *in vivo* the host (e.g., a human patient) unfortunately dies well beforehand. Many mathematical models of tumor growth involve an ordinary differential equation of form

$$x' = f(x)x \ ,$$

where $x(t)$ is the biomass of the tumor and the prime denotes differentiation with respect to the time t, the function $f(x)$ is approximately constant for small x, and then has derivative $f'(x) < 0$ for larger x until $x = \hat{x} > 0$ (say), when the growth stops (i.e., $f(\hat{x}) = 0$). One of the most prominent and robust of this family of models is the generalised logistic equation, where

$$f(x) = q - rx^{\nu}, \nu > 0.$$

After diagnosis, patients may undergo various therapies, including surgery as a leading option. However, it cannot be guaranteed that a tumor has been totally removed, and in order to kill metastases the patient may undergo chemotherapy. In some cases, a chemotherapy is carried out beforehand, to reduce the size of the tumor prior to its surgical removal. An anti-tumor chemotherapy may be modelled by the ordinary differential equation

$$x' = x(q - rx^{\nu}) - g(t)x,$$

where $g(t) > 0$ is the profile of the drug concentration. One way to deliver the therapy is continuous infusion of the drug, in order to partially reduce drug-related major side effects — i.e., $g(t)$ is approximately constant, so the chemotherapy model becomes

$$\begin{aligned} x' &= qx - rx^{1+\nu} - cx, \\ x(0) &= x_0 > 0\ . \end{aligned} \tag{1.1}$$

Let us suppose that the chemotherapy proceeds for a very long time, so that we are interested in the asymptotic solution behaviour.

The constancy of $g(t)$ is of course only approximate in reality, and a classical way to represent the variability of $g(t)$ is to include a white noise perturbation function $\xi(t)$ of known standard error σ. This is considered in § 2, where σ is treated as a stochastic bifurcation parameter. However, in § 3 some potential problems with imposing a Gaussian perturbation in tumor models are stressed, which therefore must be complemented by some biological caveats. Some alternative ways to model uncertainties in mathematical oncology are discussed in the concluding remarks.

2. Deterministic and stochastic modelling of anti-tumor chemotherapy delivered with continuous infusion

It is easy to verify that the deterministic model (1.1) has the following properties:

- if $0 \le c < q$, then $x(t) \to x_e(c) = ((q-c)/r)^{1/\nu}$, so the tumor is not eradicated; and
- if $c \ge q$, then $x' \le -rx^{1+\nu} \Rightarrow x(t) \to 0^+$, so the tumor is eradicated.

However, it is important to stress that $x_e(c)$ is not normally compatible with actual human life experience (even when $x_e(c)$ is quite small), because of the insurgency of tumor-correlated phenomena — principally tumor diffusion processes that lead to the birth of metastases.

To model the unknown variations of the therapy infusion and also the basic tumor growth rate q, let us include a stochastic term $\sigma\xi(t)x$ in the model (1.1) to obtain

$$\begin{aligned} x' &= qx - rx^{1+\nu} - cx + \sigma\xi(t)x, \\ x(0) &= x_0\ , \end{aligned}$$

where $\xi(t)$ is a white noise. In particular, let us consider the stochastic model

$$\begin{aligned} dx &= (qx - rx^{1+\nu} - cx)dt + \sigma x dW, \\ x(0) &= x_0, \end{aligned}$$

where $W(t)$ is a Wiener process. To the best of my knowledge, this is the first stochastic model of CI chemotherapy in tumors, and both mathematical and biological aspects of the results are considered in this section.

As is well known, the evolution of the probability density of a random variable $x(t)$ is determined by the Fokker–Planck equation, which in our context has the form

$$\frac{\partial \rho}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}\left(x^2\rho\right) - \frac{\partial}{\partial x}\left(\left(ax - rx^{1+\nu}\right)\rho\right) .$$

At steady state, if $\dfrac{\sigma^2}{2} \geq (q-c)$, then $\rho(x,t) \to \delta(x)$,

so if there is eradication (i.e., if $q-c \leq 0$) then the presence of noise cannot change the outcome of the therapy in the deterministic case (since $\sigma > 0$). Furthermore and more interestingly:

Remark 2.1. *If $q-c>0$, a noise with a sufficiently high σ can induce the tumor eradication with unitary probability.*

On the other hand, if $\sigma^2/2 < (q-c)$, the steady state equation

$$\frac{\sigma^2}{2}\frac{d^2}{dx^2}\left(x^2\rho\right) = \frac{d}{dx}\left((((q-c))x - rx^{1+\nu}\right)\rho)$$

has solutions of the form

$$\rho(x) = Cx^{2/\sigma^2(q-c)-2}\exp\left(-\frac{2r}{\sigma^2\nu}x^\nu\right).$$

Thus

- if $\sigma^2 \in ((q-c), 2(q-c)]$, then $\rho(x)$ is decreasing and unbounded; and
- if $\sigma^2 \leq q-c$, then $\rho(x)$ is bounded and it has a non-null maximum.

Summarising, there are two stochastic bifurcations — viz,

- at $\sigma^2 = 2(q-c)$ there is a transition between eradication with probability 1 (i.e., $\rho(x)=\delta(x)$) to the regimen 'small tumor highly probable'; and
- at $\sigma^2 = q-c$ there is a transition between the regimen 'small tumor highly probable' to the regimen 'small sizes are unlikely'.

3. Gaussian noise modelling in population growth models

The eradication result envisaged in Remark 2.1 is quite paradoxical, and it may be biologically questionable — since it is likely that a real "tumor + chemotherapy" system may be subject to significant variations in the tumor proliferation rate, and to a lesser extent in the instantaneous delivery of the drug. Thus it is likely that there may be rather high values of the parameter σ, and the tumor eradication should be reached more often than actually observed. It is therefore worthwhile to investigate why this happens, in a general setting. For this purpose, let us consider a population of whatever kind (tumor cells, animals etc.) growing with a birth rate β and a death rate μ, so the population grows according as

$$x' = \beta x - \mu x \; .$$

Suppose the growth is not constant, due to many small and partially unknown external environmental factors, and that one considers modelling the sum of all the external influences as a white noise: i.e.,

$$\begin{aligned} &< \xi(t)\xi(t+\theta) >= \delta(\theta), \\ &x' = \beta\left(1+\tfrac{\sigma}{\beta}\xi(t)\right)x - \mu x. \end{aligned} \tag{3.1}$$

A first and worthy objection, that equation (3.1) represents an approximation of a stochastic process with an integer space state, is neglected here. Our original model involves an ordinary differential equation, and it was considered acceptable to approximate the discrete state space by a real continuum. Thus let us proceed by representing (3.1) as the stochastic differential equation

$$dX = \beta X dt + \sigma X dW - \mu X dt,$$

which has the analytical solution

$$x(t) = x(0)\exp\left(\left(\beta - \mu - \frac{\sigma^2}{2}\right)t + \int_0^t dW\right) \; .$$

As a consequence, we obtain the surprising result that

$$\beta - \mu - \frac{\sigma^2}{2} < 0 \Rightarrow X(t) \to 0 \; .$$

However, this elegant formal approach has a hidden pitfall — viz. since the Gaussian noise is unbounded, the perturbed birth rate may become negative, which is of course biologically unrealistic. In mathematical terms:

$$Prob\,(\beta dt + \sigma dW < 0) > 0.$$

Another major drawback is that the birth rate may become too big, which is equally unrealistic. Similar problems may arise by adopting Gaussian perturbations of many positive parameters in other biological models. Some intriguing biological aspects related to the use of the Stratonovich approach in modelling population dynamics may also be of interest [2].

4. Discussion

Bounded noise is necessary in population models, as has already been stressed by several authors (cf. [4] and references therein). Non-stochastic models using a fuzzy approach have also been proposed [3] as an alternative to stochastic models in population dynamics [4] and in immuno-oncology [5]. The author has recently proposed a fuzzy oncological model of CI therapy [6], which produces some results that are significantly different from those discussed in § 2 — viz.

- if $q - c \geq 0$, there cannot be the noise-induced tumor eradication; and
- if $q - c < 0$, there can be tumor escape from the therapy-induced eradication only if the "amplitude" of the fuzzy noise exceeds a threshold value.

However, although one may define a bounded fuzzy noise, the fuzzy approach is not completely satisfactory given that the theory of fuzzy systems and fuzzy differential equations is relatively unexplored. For example:

- the membership function of fuzzy theory does not describe the statistical properties of fuzzy variables; and
- the theory of fuzzy bifurcations is incomplete and not yet well-founded.

In brief, in the mathematical description of parameter perturbations in oncological models, it appears that:

- the use of Gaussian noise may lead to biologically paradoxical results;
- adopting the fuzzy noise approach leads to results that are biologically more robust, but the underlying theory is immature; and
- a better approach may be to use non-Gaussian bounded stochastic noise, but this does not allow the use of the Ito or Stratonovitch calculus [2].

In conclusion, a significant improvement in modelling tumor growth might require the development of a more complete theory of differential equations with bounded noise perturbations.

Acknowledgment

I wish to thank an anonymous referee and the editors of this book for numerous useful suggestions that helped me to significantly improve this article.

This work has been done in the framework of the integrated project "Advancing clinical-genomic trials on cancer ACGT". The ACGT project (FPG-2005-IST-026996) is partly funded by the EC and the author is grateful for this support.

References

[1] M. Peckham, H. M. Pinedo and U. Veronesi (Editors), *Oxford Textbook of Oncology*, 2nd edition, Oxford University Press, 2001.

[2] C. A. Braumann, *Harvesting in a Random Environment: Ito or Stratonovich calculus?*, J. of Theoretical Biology **244** (2007), 424-432.

[3] L. Zadeh, *Fuzzy Sets*, Information and Control **8** (1965), 338-353.

[4] V. Krivan and G. Colombo, *A Non-stochastic Approach for Modeling Uncertainty in Population Dynamics*, Bulletin of Mathematical Biology **60** (1998), 721-751.

[5] K. K. Majumdar and D. D. Majumder, *Fuzzy Differential Inclusions in Atmospheric and Medical Cybernetics*, IEEE Transactions on Systems, Man and Cybernetics **34** (2004), 877-887.

[6] A. d'Onofrio, *Fuzzy Oncology*, Applied Mathematics Letters (in press, electronic version available at the webpage http://dx.doi.org/10.1016/j.aml.2007.05.019).

Alberto d’Onofrio
Division of Epidemiology and Biostatistics
European Institute of Oncology
Via Ripamonti, 435
Milano
Italy
e-mail: alberto.donofrio@ieo.it

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 235–248

Phylogenetic Analysis, Split Systems and Boolean Functions

Andreas Dress

Abstract. In phylogenetic analysis, *split systems* have been investigated extensively over the last twenty or thirty years. In particular, the following *inverse problem* has found much attention in this field: Given a finite set X, let $\mathcal{S}(X)$ denote the set consisting of all *splits* of X, i.e., all 2-element subsets $\{A,B\}$ of the power set $\mathcal{P}(X)$ of X for which $A \cup B = X$ and $A \cap B = \emptyset$ holds. Associate, to any $\mathbb{R}_{\geq 0}$-*weighted split system* Σ — i.e., to any map Σ from $\mathcal{S}(X)$ into the set $\mathbb{R}_{\geq 0}$ of non-negative real numbers — the metric

$$D_\Sigma : X \times X \to \mathbb{R}_{\geq 0} : (x,y) \mapsto \sum_{\{A,B\} \in \mathcal{S}(X):\, x \in A,\, y \in B} \Sigma(\{A,B\}).$$

Then given any metric D defined on X, one wants to find such a map Σ from $\mathcal{S}(X)$ into $\mathbb{R}_{\geq 0}$ such that D_Σ, at least approximately, coincides with D and such that, in addition, the support of Σ has certain desirable properties.

Here we re-interpret this task in the context of a rather naturally defined injective map $\mathcal{D}_\bullet$ from the $\mathbb{R}$-vectorspace of all $\mathbb{R}$-weighted split systems into the $\mathbb{R}$-vectorspace $\mathcal{B}(X|\mathbb{R})$ of $\mathbb{R}$-*valued Boolean functions* defined on X (i.e., the $\mathbb{R}$-vectorspace consisting of all maps from the power set $\mathcal{P}(X)$ of X into $\mathbb{R}$) that associates, to any given $\mathbb{R}$-weighted split system $\Sigma : \mathcal{S}(X) \to \mathbb{R}$, the map $\Sigma_\bullet \in \mathcal{B}(X|\mathbb{R})$ that maps any subset A of X onto the sum $\Sigma_\bullet(A) := \sum_{A' \in \mathcal{P}(X-A)} \Sigma(\{A \cup A', X - (A \cup A')\})$. Note that $D_\Sigma(x,y)$ apparently coincides, for all $x, y \in X$, with the difference between the sum $|\Sigma| := \sum_{S \in \mathcal{S}(X)} \Sigma(S)$ over all values of Σ and the value $\Sigma_\bullet(\{x,y\})$ that $\Sigma_\bullet$ attains at the subset $\{x,y\}$ of X.

More specifically, we show that there exist two canonically defined $\mathbb{R}$-linear involutions τ and ρ of $\mathcal{B}(X|\mathbb{R})$ (i.e. $\mathbb{R}$-linear endomorphisms of $\mathcal{B}(X|\mathbb{R})$ for which $\tau^2 = \rho^2 = \mathrm{Id}_{\mathcal{B}(X|\mathbb{R})}$ holds) that are mutually *anti-adjoint* (relative to the canonical inner product defined on $\mathcal{B}(X|\mathbb{R})$); and have the property that the fixed-point space $\mathcal{B}(X|\mathbb{R})^\tau$ of τ coincides with the set of all Boolean functions Φ of the form $\Phi = \Sigma_\bullet$ for some (necessarily unique!) map $\Sigma \in \mathcal{S}(X)$, as well as with the set of all Boolean functions of the form $\Phi = \Pi + \tau(\Pi)$ for some appropriate Boolean function $\Pi \in \mathcal{B}(X|\mathbb{R})$, and we discuss possible applications of these observations in the context of phylogenetic reconstruction.

Mathematics Subject Classification (2000). 05C05, 92D15.

Keywords. Phylogenetic Analysis, Splits, Split Systems, Boolean Functions.

1. Introduction: set systems in phylogenetic analysis

Given a collection X of species, a *clade* C in X is a subset of X that consists of all species in X that are offspring of a single ancestral species, while none of the species in the complement $X - C$ of C have evolved from this ancestral species. In other words, denoting the last common ancestor of all species in an arbitrary subset C of X by $\mathtt{lca}(C)$, a subset C of X is a clade if and only if none of the species in $X - C$ is a descendant of $\mathtt{lca}(C)$.

One of the most basic tasks in phylogenetic analysis is, given a set X as above, to identify the collection of all clades in X. Yet, as **Charles Darwin** put it in his treatise **The descent of man, and selection in relation to sex**: *As we have no record of the lines of descent, the pedigree can be discovered only by the degrees of resemblance between the beings which are to be classed.* That is, all that we commonly can rely on to identify the collection of all clades in X is information about how distinct, or how similar, the present-day species are that make up the set X.

Consequently, a standard assumption in phylogenetic analysis is that, together with a finite set X of species or, more generally, of any kind of taxonomic units (for short, *taxa*), we are given a metric D defined on X that quantifies that *degree of resemblance between* the taxa contained in X. In other words, one assumes that one is given a map $D : X \times X \to \mathbb{R} : (x, y) \mapsto D(x, y)$ from $X \times X$ into the set $\mathbb{R}$ of real numbers for which $D(x, x) = 0$ and $D(x, y) \leq D(x, z) + D(y, z)$ holds for any three taxa x, y, z under consideration[1]. And the task one has to address can then be described as that of designing methods for deriving, from these data, a *phylogenetic X-tree* $T = T(D)$ that — at least approximately — represents the map D. That is, one has to find a finite edge-weighted and X-labeled tree $T = (V, E, \ell; \varphi)$ consisting of

- a vertex set V,
- an edge set $E \subseteq \binom{V}{2}$,
- a *weight map* $\ell : E \to \mathbb{R}_{>0}$ from E into the set $\mathbb{R}_{>0}$ of positive real numbers,
- and a *labeling map* $\varphi : X \to V$ whose image $\varphi(X)$ contains — at least — all vertices in V of degree 1 or 2

such that the distance $D(x, y)$ of any two taxa x, y in X coincides — at least approximately — with the length $\ell_T\big(\varphi(x), \varphi(y)\big)$ of the unique path in T from $\varphi(x)$ to $\varphi(y)$ relative to ℓ (cf. for example [18] for a thorough discussion of this

The author was supported by the Chinese Academy of Sciences and the German BMBF; and further support from the Royal Brunei Airlines, which enabled the author to contribute to the International Conference on Mathematical Modelling and Computation held at the University of Brunei Darussalam during 5–8 June 2006 in conjunction with the 20th anniversary celebration of the foundation of that university, is also gratefully acknowledged. The author is particularly grateful for many helpful comments by the editors and two unusually attentive and thorough reviewers, who helped to improve the text drastically.

[1] Note that, putting $y := x$, this implies that $0 \leq 2D(x, z)$ and hence $0 \leq D(x, z)$ holds for all $x, z \in X$; and putting $z := x$, that $D(x, y) \leq D(y, x)$ and hence $D(x, y) = D(y, x)$ holds for all $x, y \in X$ — thus implying the standard inequalities required for D being a (pseudo-)metric.

concept that was, and still is, one of the focal points for all conceptual development in computational phylogenetics).

Remarkably, denoting the set consisting of all *splits* of X — i.e., the set of all 2-element subsets $\{A,B\}$ of the power set $\mathcal{P}(X)$ of X for which $A \cup B = X$ and $A \cap B = \emptyset$ holds — by $\mathcal{S}(X)$, this task is simply equivalent to finding a map Σ from $\mathcal{S}(X)$ into the set $\mathbb{R}_{\geq 0}$ of non-negative real numbers such that:

(i) the distance $D(x,y)$ of x and y coincides — again at least approximately — with the sum

$$\Sigma(x:y) := \sum_{\{A,B\}\in\mathcal{S}(X):\, x\in A,\, y\in B} \Sigma(\{A,B\});$$

(ii) $\Sigma(\{\emptyset, X\}) = 0$ holds; and

(iii) any two splits in the support

$$\mathrm{supp}(\Sigma) := \big\{\{A,B\} \in \mathcal{S}(X) : \Sigma(\{A,B\}) \neq 0\big\}$$

of Σ are *compatible* — i.e., one of the four intersections $A\cap A'$, $A\cap B'$, $B\cap A'$, $B \cap B'$ is empty for any two splits $\{A,B\}$ and $\{A',B'\}$ in $\mathrm{supp}(\Sigma)$.

This fact, also discussed *in extenso* in [18], was probably folklore already in the mid-twentieth century , in one or the other disguise. It was stated explicitly, more or less just as stated above, by Buneman around 1970 (cf. for instance [7]); and it has also been another one of those fundamental insights on which much further development of computational phylogenetics was based.

Let us now recall the following simple facts:

(i) Any metric D defined on a set X — or, more generally, any symmetric map $D: X \times X \to \mathbb{R}$ — can be viewed as a real-valued map defined on the subset $\mathcal{P}_{\leq 2}(X)$ of $\mathcal{P}(X)$ consisting of all non-empty subsets of X of cardinality at most 2.

(ii) Pursuing this point of view and noting that

$$\begin{aligned}\Sigma(x:y) &= \sum_{\{A,B\}\in\mathcal{S}(X):\, x\in A,\, y\in B} \Sigma(\{A,B\}) \\ &= \sum_{\{A,B\}\in\mathcal{S}(X):\, \{x,y\}\not\subseteq A,\, \{x,y\}\not\subseteq B} \Sigma(\{A,B\})\end{aligned}$$

holds for all $x,y \in X$, it was suggested in recent investigations of *phylogenetic diversity* and related issues [1,6,9–17,19] to associate, to any map Σ from $\mathcal{S}(X)$ into $\mathbb{R}$, the $\mathbb{R}$-linear map $\Sigma^\bullet$ from $\mathcal{P}(X)$ into $\mathbb{R}$ that attains, on a given subset A_0 of X, the value

$$\Sigma^\bullet(A_0) := \sum_{\{A,B\}\in\mathcal{S}(X):\, A_0\not\subseteq A,\, A_0\not\subseteq B} \Sigma(\{A,B\}),$$

this way introducing an $\mathbb{R}$-linear map

$$\mathcal{D}^\bullet : \mathcal{S}(X|\mathbb{R}) \to \mathcal{B}(X|\,\mathbb{R}) : \Sigma \mapsto \Sigma^\bullet$$

from the $\mathbb{R}$-vectorspace

$$\mathcal{S}(X|\mathbb{R}) := \mathbb{R}^{\mathcal{S}(X)}$$

of all $\mathbb{R}$-*weighted split systems* (i.e., all maps from $\mathcal{S}(X)$ into $\mathbb{R}$) into the $\mathbb{R}$-vectorspace

$$\mathcal{B}(X|\,\mathbb{R}) := \mathbb{R}^{\mathcal{P}(X)}$$

of all $\mathbb{R}$-*valued Boolean functions* defined on X (i.e., all maps from $\mathcal{P}(X)$ into $\mathbb{R}$).

(iii) Furthermore, it is easily verified (e.g., see [12]) that, putting

$$|\Sigma| := \sum_{S\in\mathcal{S}(X)} \Sigma(S)$$

for every $\Sigma \in \mathcal{S}(X|\,\mathbb{R})$ and restricting the map $\mathcal{D}^\bullet$ to the subspace

$$\mathcal{S}_0(X|\mathbb{R}) := \{\Sigma \in \mathcal{S}(X|\,\mathbb{R}) : |\Sigma| = 0\}$$

of $\mathcal{S}(X|\mathbb{R})$, one obtains an **injective** map

$$\mathcal{D}^0 := \mathcal{D}^\bullet|_{\mathcal{S}_0(X|\mathbb{R})}$$

from $\mathcal{S}_0(X|\mathbb{R})$ into $\mathcal{B}(X|\,\mathbb{R})$.

(iv) Thus given any symmetric map $D : X \times X \to \mathbb{R}$, viewing this map D as a map from $\mathcal{P}_{\leq 2}(X)$ into $\mathbb{R}$ implies that the inverse problem of finding a map Σ from $\mathcal{S}(X)$ into $\mathbb{R}$ such that

(i) $D(\{x, y\})$ coincides, for all x, y in X, with $\Sigma(x : y)$

(ii) and certain additional desirable properties are also satisfied by Σ

can be rephrased as asking for an extension of the map D to a map $\overline{D}$ in $\mathcal{B}(X|\,\mathbb{R})$ that (i) is contained in the image $\mathcal{D}^\bullet\big(\mathcal{S}_0(X|\mathbb{R})\big)$ of the subspace $\mathcal{S}_0(X|\mathbb{R})$ of $\mathcal{S}(X|\,\mathbb{R})$ relative to the map $\mathcal{D}^\bullet$, and (ii) satisfies in addition certain desirable properties — an observation that motivated searching for the characterisations of that image $\mathcal{D}^\bullet\big(\mathcal{S}_0(X|\mathbb{R})\big)$ communicated in [12], and the resulting consequences communicated here.

(v) Finally, associating to any $\mathbb{R}$-weighted split system $\Sigma \in \mathcal{S}(X|\,\mathbb{R})$ the map $\Sigma_\bullet \in \mathcal{B}(X|\,\mathbb{R})$ from $\mathcal{P}(X)$ into $\mathbb{R}$ that attains, on a given subset A_0 of X, the value

$$\Sigma_\bullet(A_0) := \sum_{A\in\mathcal{P}(X-A_0)} \Sigma(\{A_0 \cup A, X - (A_0 \cup A)\}),$$

we obtain an injective map

$$\mathcal{D}_\bullet : \mathcal{S}(X|\,\mathbb{R}) \to \mathcal{B}(X|\,\mathbb{R}) : \Sigma \mapsto \Sigma_\bullet$$

for which

$$\Sigma^\bullet(A) + \Sigma_\bullet(A) = |\Sigma|$$

holds for every $\Sigma \in \mathcal{S}(X|\,\mathbb{R})$ and every non-empty subset A of X (while $\Sigma^\bullet$ and $\Sigma_\bullet$ attain the value 0 and $2|\Sigma|$ respectively, on the empty subset of X provided X itself is not the empty set).

Thus from a formal point of view, studying the map $\mathcal{D}_\bullet$ is essentially just as good as studying the map $\mathcal{D}^\bullet$ while, from the point of view of mathematical simplicity, studying the map $\mathcal{D}_\bullet$ is often clearly preferable to studying the map $\mathcal{D}^\bullet$. In particular, the image $\mathcal{D}^\bullet\big(\mathcal{S}_0(X|\mathbb{R})\big)$ of the subspace $\mathcal{S}_0(X|\mathbb{R})$ of $\mathcal{S}(X|\,\mathbb{R})$ relative to the map $\mathcal{D}^\bullet$ coincides with the image $\mathcal{D}_\bullet\big(\mathcal{S}_0(X|\mathbb{R})\big)$ of that subspace relative to $\mathcal{D}_\bullet$ while, as shown in [12], the image $\mathcal{D}_\bullet\big(\mathcal{S}(X|\,\mathbb{R})\big)$ of $\mathcal{S}(X|\,\mathbb{R})$ relative to $\mathcal{D}_\bullet$ consists of all maps $\Phi \in \mathcal{B}(X|\,\mathbb{R})$ for which

$$\Phi(A) = \sum_{A' \in \mathcal{P}(A)} (-1)^{|A'|}\, \Phi(A')$$

holds for all $A \subseteq X$, and the image $\mathcal{D}_\bullet\big(\mathcal{S}_0(X|\mathbb{R})\big)$ of $\mathcal{S}_0(X|\mathbb{R})$ relative to $\mathcal{D}_\bullet$ — and, hence, also the image $\mathcal{D}^\bullet\big(\mathcal{S}_0(X|\mathbb{R})\big)$ of $\mathcal{S}_0(X|\mathbb{R})$ relative to $\mathcal{D}^\bullet$ — consists of all such maps $\Phi \in \mathcal{B}(X|\,\mathbb{R})$ that, in addition, vanish on every one-element subset of X (and therefore also on the empty set).

These observations prompted the investigations communicated below. While they may be of some interest in their own right in view of their intriguing simplicity, they also imply for instance that, associating to any map $\Psi \in \mathcal{B}(X|\,\mathbb{R})$ from $\mathcal{P}(X)$ into $\mathbb{R}$ the linear form

$$\Psi^* : \mathcal{S}(X|\,\mathbb{R}) \to \mathbb{R} : \Sigma \mapsto \sum_{A \in \mathcal{P}(X)} \sum_{A' \in \mathcal{P}(A)} \Sigma\big(\{A, X-A\}\big)\, \Psi(A'),$$

one has $\Psi^* = 0$ for some map $\Psi \in \mathcal{B}(X|\,\mathbb{R})$ if and only if

$$\Psi(A) = (-1)^{1+|A|} \sum_{A' :\, A \subseteq A' \subseteq X} \Psi(A')$$

holds for every subset A of X, and that every $\mathbb{R}$-linear form defined on $\mathcal{S}(X|\,\mathbb{R})$ is of the form Ψ^* for some map $\Psi \in \mathcal{B}(X|\,\mathbb{R})$ from $\mathcal{P}(X)$ into $\mathbb{R}$.

Further, they imply that there exists a canonical one-to-one correspondence between collections $\mathcal{S} \subseteq \mathcal{S}(X)$ of X-splits and those collections $\mathbf{p} \subseteq \mathcal{P}(X)$ of subsets of X for which, for every subset A of X, the number $\#\{A' \in \mathbf{p} : A' \subsetneq A\}$ of proper subsets of A that are elements of $\mathbf{p}$ is even. And that this, in turn, holds if and only if — dually — $\mathbf{p} \cap \mathbf{p}'$ is of even cardinality for all collections $\mathbf{p}' \subseteq \mathcal{P}(X)$ of subsets of X for which, for every subset A of X, the number $\#\{A' \in \mathbf{p}' : A \subsetneq A'\}$ of subsets in $\mathbf{p}'$ properly containing A is even.

Thus if $\mathbf{p}$ is such a collection of subsets of X and X is not empty, $\mathbf{p}$ cannot contain the empty set as this is the only proper subset of any one-element subset $\{a\}$; and it either contains all or no one-element subset of X, because the cardinality of $\{A' : A' \subsetneq \{a,b\}, A' \in \mathbf{p}\} = \{\{a\},\{b\}\} \cup \mathbf{p}$ must be even for any two distinct elements $a, b \in X$.

More specifically, associating to each split $S = \{A, B\}$ of X the set system $\mathbf{p}(S) := \{A_0 \subseteq X : \emptyset \neq A_0 \subseteq A\} \cup \{A_0 \subseteq X : \emptyset \neq A_0 \subseteq B\}$ yields a collection $\{\mathbf{p}(S) : S \in \mathcal{S}(X)\}$ of subsets of $\mathcal{P}(X)$ such that, given a collection $\mathbf{p} \subseteq \mathcal{P}(X)$ of subsets of X, the following assertions are equivalent:

(i) $\mathbf{p}$ is a symmetric difference

$$\Delta(\mathbf{p}(S) : S \in \mathcal{S}) := \{A \subseteq X : \#\{S \in \mathcal{S} : A \in \mathbf{p}(S)\} \text{ is odd}\}$$

of set systems of the form $\mathbf{p}(S)$, where S runs through all splits in some system $\mathcal{S} = \mathcal{S}(\mathbf{p}) \subseteq \mathcal{S}(X)$ of X-splits;

(ii) the number $\#\{A' \in \mathbf{p} : A' \subsetneq A\}$ of proper subsets of any given subset A of X that are elements of $\mathbf{p}$ is even,

in which case the set $\mathcal{S} = \mathcal{S}(\mathbf{p})$ is uniquely determined by $\mathbf{p}$.

It is hoped that these observations may help to clarify some of the arguments developed in [9–12] and related papers.

2. Two mutually anti-adjoint canonical involutions defined on the space of Boolean functions

In this section, for a field K and a finite set X of cardinality $n > 0$, we introduce two canonical K-linear involutions defined on the K-vectorspace

$$\mathcal{B}(X|\, K) := K^{\mathcal{P}(X)}$$

consisting of all K-valued Boolean functions defined on X — i.e., all maps from the power set $\mathcal{P}(X)$ of X into the field K.

The first is the involution

$$\tau = \tau_{(X|K)} : \mathcal{B}(X|\, K) \to \mathcal{B}(X|\, K) : \Pi \to \widehat{\Pi}$$

from $\mathcal{B}(X|\, K)$ onto itself, which maps any map $\Pi \in \mathcal{B}(X|\, K)$ onto the map $\widehat{\Pi}$ that associates, to any given subset A of X, the element

$$\widehat{\Pi}(A) := \sum_{A' \in \mathcal{P}(A)} (-1)^{|A'|}\, \Pi(A').$$

The map τ is evidently a K-linear endomorphism of $\mathcal{B}(X|\, K)$, and it is also an automorphism of $\mathcal{B}(X|\, K)$ of order 2 — i.e., one has

$$\tau^2 = \mathrm{Id}_{\mathcal{B}(X|\, K)}.$$

Indeed, we have

Lemma 2.1. *The identity*

$$\sum_{A' :\, A'' \subseteq A' \subseteq A} (-1)^{|A'|} = (-1)^{|A|}\, \delta_{A,A''} = (-1)^{|A''|}\, \delta_{A,A''}$$

holds for all $A, A'' \subseteq X$ with $A'' \subseteq A$.

Consequently, given any map $\Pi \in \mathcal{B}(X|\,K)$ and any subset A of X, we have

$$\begin{aligned}\widehat{(\widehat{\Pi})}(A) &= \sum_{A'\in\mathcal{P}(A)} (-1)^{|A'|}\,\widehat{\Pi}(A')\\ &= \sum_{A'\in\mathcal{P}(A)} (-1)^{|A'|} \sum_{A''\in\mathcal{P}(A')} (-1)^{|A''|}\,\Pi(A'')\\ &= \sum_{A''\in\mathcal{P}(A)} (-1)^{|A''|}\,\Pi(A'') \sum_{A'\colon A''\subseteq A'\subseteq A} (-1)^{|A'|}\\ &= \sum_{A''\in\mathcal{P}(A)} (-1)^{|A''|}\,\Pi(A'')\,(-1)^{|A''|}\,\delta_{A,A''}\\ &= \Pi(A).\end{aligned}$$

The second involution we introduce is the map

$$\rho = \rho_{(X|K)} : \mathcal{B}(X|\,K) \to \mathcal{B}(X|\,K) : \Pi \to \overline{\Pi}$$

from $\mathcal{B}(X|\,K)$ onto itself that maps any map $\Pi \in \mathcal{B}(X|\,K)$ onto the map $\overline{\Pi}$ that associates, to any given subset A_0 of X, the element

$$\overline{\Pi}(A) := (-1)^{1+|A|} \sum_{A'\colon A\subseteq A'\subseteq X} \Pi(A') = -(-1)^{|A|} \sum_{A'\colon A\subseteq A'\subseteq X} \Pi(A').$$

This is also evidently a K-linear endomorphism of $\mathcal{B}(X|\,K)$, which satisfies the identity

$$\rho^2 = \mathrm{Id}_{\mathcal{B}(X|\,K)}$$

in view of the fact that

$$\begin{aligned}\overline{(\overline{\Pi})}(A) &= (-1)^{1+|A|} \sum_{A'\colon A\subseteq A'\subseteq X} \overline{\Pi}(A')\\ &= (-1)^{1+|A|} \sum_{A'\colon A\subseteq A'\subseteq X} (-1)^{1+|A'|} \sum_{A''\colon A'\subseteq A''\subseteq X} \Pi(A'')\\ &= (-1)^{|A|} \sum_{A''\colon A\subseteq A''\subseteq X} \Pi(A'') \sum_{A'\colon A\subseteq A'\subseteq A''} (-1)^{|A'|}\\ &= (-1)^{|A|} \sum_{A''\colon A\subseteq A''\subseteq X} \Pi(A'')\,(-1)^{|A|}\delta_{A,A''}\\ &= \Pi(A)\end{aligned}$$

holds for every map $\Pi \in \mathcal{B}(X|\,K)$ and every subset A of X.

In the next section, we use the fact that these two involutions are *anti-adjoint* relative to the canonical inner product defined on $\mathcal{B}(X|\,K)$ that associates, to any two maps $\Phi, \Psi \in \mathcal{B}(X|\,K)$, the sum

$$\langle \Phi \mid \Psi \rangle := \sum_{A\in\mathcal{P}(X)} \Phi(A)\,\Psi(A).$$

Indeed, one has $\langle \widehat{\Phi} \mid \Psi \rangle = -\langle \Phi \mid \overline{\Psi} \rangle$ for all $\Phi, \Psi \in \mathcal{B}(X|\,K)$ as, given any two maps $\Phi, \Psi \in \mathcal{B}(X|\,K)$, we have

$$\begin{aligned}
\langle \widehat{\Phi} \mid \Psi \rangle &= \sum_{A \in \mathcal{P}(X)} \widehat{\Phi}(A)\, \Psi(A) \\
&= \sum_{A \in \mathcal{P}(X)} \sum_{A' \in \mathcal{P}(A)} (-1)^{|A'|}\, \Phi(A')\, \Psi(A) \\
&= \sum_{A' \in \mathcal{P}(X)} \sum_{A\colon A' \subseteq A \subseteq X} (-1)^{|A'|}\, \Phi(A')\, \Psi(A) \\
&= \sum_{A' \in \mathcal{P}(X)} \Phi(A')\, (-1)^{|A'|} \sum_{A\colon A' \subseteq A \subseteq X} \Psi(A) \\
&= \sum_{A' \in \mathcal{P}(X)} -\Phi(A')\, \overline{\Psi}(A') \\
&= -\langle \Phi \mid \overline{\Psi} \rangle.
\end{aligned}$$

3. The fixed-point spaces of τ and ρ

Now let $\mathcal{B}(X|\,K)^\tau$ denote the subspace of $\mathcal{B}(X|\,K)$ consisting of all fixed points $\Phi \in \mathcal{B}(X|\,K)$ of the involution $\tau = \tau_{(X|K)}$, and let $\mathcal{B}(X|\,K)^\rho$ denote the subspace of $\mathcal{B}(X|\,K)$ consisting of all fixed points $\Psi \in \mathcal{B}(X|\,K)$ of the involution $\rho = \rho_{(X|K)}$ — i.e., put

$$\begin{aligned}
\mathcal{B}(X|\,K)^\tau &:= \big\{\Phi \in \mathcal{B}(X|\,K) : \widehat{\Phi} = \Phi\big\} \\
&= \Big\{\Phi \in \mathcal{B}(X|\,K) : A_0 \subseteq X \;\Rightarrow\; \Phi(A_0) = \sum_{A \in \mathcal{P}(A_0)} (-1)^{|A|}\, \Phi(A)\Big\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{B}(X|\,K)^\rho &:= \big\{\Psi \in \mathcal{B}(X|\,K) : \overline{\Psi} = \Psi\big\} \\
&= \Big\{\Psi \in \mathcal{B}(X|\,K) : A_0 \subseteq X \;\Rightarrow\; \Psi(A_0) = -(-1)^{|A_0|} \sum_{A\colon A_0 \subseteq A \subseteq X} \Psi(A)\Big\}.
\end{aligned}$$

Of course, the fact that τ and ρ are anti-adjoint relative to the canonical inner product defined on $\mathcal{B}(X|\,K)$ implies that, given any Φ in $\mathcal{B}(X|\,K)^\tau$ and Ψ in $\mathcal{B}(X|\,K)^\rho$, we have

$$\langle \Phi \mid \Psi \rangle = \langle \widehat{\Phi} \mid \Psi \rangle = -\langle \Phi \mid \overline{\Psi} \rangle = -\langle \Phi \mid \Psi \rangle$$

and therefore

$$2\,\langle \Phi \mid \Psi \rangle = 0.$$

However, we can do a bit better — i.e., we can establish the following

Theorem 3.1. *Given X and K as above, the two subspaces $\mathcal{B}(X|\,K)^\tau$ and $\mathcal{B}(X|\,K)^\rho$ of $\mathcal{B}(X|\,K)$ are mutually orthogonal complements relative to the canonical inner product defined on $\mathcal{B}(X|\,K)$, every ρ-invariant element in $\mathcal{B}(X|\,K)^\rho$ is a ρ-*trace —

i.e., it is of the form $\Pi+\overline{\Pi}$ *for some* $\Pi\in\mathcal{B}(X|\,K)$*; and every* τ*-invariant element in* $\mathcal{B}(X|\,K)^\tau$ *is a* τ-trace *— i.e., it is of the form* $\Pi+\widehat{\Pi}$*, for some* $\Pi\in\mathcal{B}(X|\,K)$*.*

Proof: The fact that

$$\langle\Phi\,|\,\Pi+\overline{\Pi}\rangle=\langle\Phi\,|\,\Pi\rangle+\langle\Phi\,|\,\overline{\Pi}\rangle=\langle\Phi|\,\Pi\rangle-\langle\widehat{\Phi}\,|\,\Pi\rangle=\langle\Phi-\widehat{\Phi}|\,\Pi\rangle$$

and

$$\langle\Pi+\widehat{\Pi}\,|\,\Psi\rangle=\langle\Pi\,|\,\Psi\rangle+\langle\widehat{\Pi}\,|\,\Psi\rangle=\langle\Pi|\,\Psi\rangle-\langle\Pi\,|\,\overline{\Psi}\rangle=\langle\Pi|\,\Psi-\overline{\Psi}\rangle$$

holds for all $\Phi,\Psi,\Pi\in\mathcal{B}(X|\,K)$ implies that the subspace

$$\mathcal{B}(X|\,\mathbb{R})_\rho:=\{\,\Pi+\overline{\Pi}:\Pi\in\mathcal{B}(X|\,K)\}$$

of $\mathcal{B}(X|\,K)^\rho$ coincides with the orthogonal complement of $\mathcal{B}(X|\,K)^\tau$, and that the subspace

$$\mathcal{B}(X|\,\mathbb{R})_\tau:=\{\,\Pi+\widehat{\Pi}:\Pi\in\mathcal{B}(X|\,K)\}$$

of $\mathcal{B}(X|\,K)^\tau$ coincides with the orthogonal complement of $\mathcal{B}(X|\,K)^\rho$. In particular,

$$\dim_K\mathcal{B}(X|\,K)^\tau+\dim_K\mathcal{B}(X|\,\mathbb{R})_\rho=\dim_K\mathcal{B}(X|\,K)^\rho+\dim_K\mathcal{B}(X|\,\mathbb{R})_\tau$$

$$=\dim_K\mathcal{B}(X|\,K)=2^n\ \le\ \dim_K\mathcal{B}(X|\,K)^\rho+\dim_K\mathcal{B}(X|\,K)^\tau$$

must hold.

Furthermore, given any arbitrary but fixed element $z\in X$, every τ-invariant map $\Phi\in\mathcal{B}(X|\,K)^\tau$ is easily seen to be completely determined by its values on the subsets of X containing z, and every ρ-invariant map $\Psi\in\mathcal{B}(X|\,K)^\rho$ by its values on the subsets of X not containing z. Indeed, $\Phi\in\mathcal{B}(X|\,K)^\tau$, $A\in\mathcal{P}(X)$ and $z\notin A$ implies that (with $B+z:=B\cup\{z\}$ for all $B\subseteq X-\{z\}$),

$$\begin{aligned}\Phi(A+z)&=\widehat{\Phi}(A+z)\\&=\sum_{A'\subseteq A+z}(-1)^{|A'|}\Phi(A')\\&=\sum_{A'\subseteq A}(-1)^{|A'|}\big(\Phi(A')-\Phi(A'+z)\big)\\&=\widehat{\Phi}(A)-\sum_{A'\subseteq A}(-1)^{|A'|}\Phi(A'+z)\\&=\Phi(A)-\sum_{A'\subseteq A}(-1)^{|A'|}\Phi(A'+z),\end{aligned}$$

and therefore

$$\Phi(A)=\Phi(A+z)+\sum_{A'\subseteq A}(-1)^{|A'|}\Phi(A'+z).$$

Further, $\Psi \in \mathcal{B}(X|\,K)^{\rho}$, $A \in \mathcal{P}(X)$ and $z \in A$ implies (with $B - z := B - \{z\}$ for any $B \subseteq X$ with $z \in B$)

$$\begin{aligned}\Psi(A-z) &= \overline{\Psi}(A-z)\\ &= (-1)^{|A|}\sum_{A-z\subseteq A'\subseteq X}\Psi(A')\\ &= (-1)^{|A|}\sum_{A\subseteq A'\subseteq X}\Psi(A') + (-1)^{|A|}\sum_{A-z\subseteq A'\subseteq X-z}\Psi(A')\\ &= \overline{\Psi}(A) + (-1)^{|A|}\sum_{A-z\subseteq A'\subseteq X-z}\Psi(A')\\ &= \Psi(A) + (-1)^{|A|}\sum_{A-z\subseteq A'\subseteq X-z}\Psi(A'),\end{aligned}$$

and therefore

$$\Psi(A) = \Psi(A-z) - (-1)^{|A|}\sum_{A-z\subseteq A'\subseteq X-z}\Psi(A').$$

Thus we must also have $\dim_K(\mathcal{B}(X|\,K)^{\tau}), \dim_K(\mathcal{B}(X|\,K)^{\rho}) \leq 2^{n-1}$, implying that

$$\mathcal{B}(X|\,\mathbb{R})_{\tau} = \mathcal{B}(X|\,K)^{\tau}, \mathcal{B}(X|\,\mathbb{R})_{\rho} = \mathcal{B}(X|\,K)^{\rho},$$

and $\dim_K(\mathcal{B}(X|\,K)^{\tau}) = \dim_K(\mathcal{B}(X|\,K)^{\rho}) = 2^{n-1}$ must hold, and that $\mathcal{B}(X|\,K)^{\tau}$ and $\mathcal{B}(X|\,K)^{\rho}$ are mutually orthogonal complements, as claimed.

4. Discussion: some consequences and some special cases

(1) Note first that the above results imply that, to find a map $\Sigma \in \mathcal{S}(X|\,\mathbb{R})$ for which $D(x,y) = \Sigma(x:y)$ holds, for all $x, y \in X$, for some given metric D, we may just as well try to extend D, considered as a map from $\mathcal{P}_{\leq 2}(X)$ to $\mathbb{R}$, to a suitable map $\Pi \in \mathcal{B}(X|\,\mathbb{R})^{\tau} = \mathcal{D}_{\bullet}\big(\mathcal{S}(X|\,\mathbb{R})\big)$ that is actually contained in the image of $\mathcal{S}_0(X|\mathbb{R})$ relative to $\mathcal{D}_{\bullet}$, and then consider the pre-image $\Sigma \in \mathcal{S}_0(X|\mathbb{R})$ of Σ relative to the map $\mathcal{D}^{\bullet}$. Actually, one can find any such extension Π by

(i) first extending D to an arbitrary map $\Pi' \in \mathcal{P}(X|\mathbb{R})$ that vanishes on the empty set and on all one-element subsets,

(ii) and then putting $\Pi := \frac{\Pi'+\tau(\Pi')}{2}$.

Indeed, noting that, for every extension $\Pi' \in \mathcal{P}(X|\mathbb{R})$ of D that vanishes on the empty set and on all one-element subsets, also $\tau(\Pi')$ is such an extension of D, we see that also $\Pi = \frac{\Pi'+\tau(\Pi')}{2}$ is necessarily such an extension and that, in addition, this extension must coincide with Π' in case Π' was already contained in $\mathcal{B}(X|\,\mathbb{R})^{\tau}$. We suggest exploring this approach towards dealing with the fundamental inverse problem in computational phylogenetics in future research.

(2) It is also worth noting that

$$\mathcal{B}(X|\,K)^{\rho} \cap \mathcal{B}(X|\,K)^{\tau} = \{0\}$$

holds, of course, for every formally real field K and hence, as we deal with linear equations over $\mathbb{Z}$ only, for every field K of characteristic 0. Therefore, also

$$\mathcal{B}(X|\,K) = \mathcal{B}(X|\,K)^{\rho} \oplus \mathcal{B}(X|\,K)^{\tau}$$

must hold for every field K of characteristic 0 in view of Theorem 3.1.

However, this does not necessarily hold for fields of finite characteristic: E.g., it is easily checked that the map

$$\Phi : \mathcal{P}(1,2) \to \mathbb{F}_2 : A \mapsto |A| \mod 2$$

is contained in $\mathcal{B}(X|\,\mathbb{F}_2)^{\rho} \cap \mathcal{B}(X|\,\mathbb{F}_2)^{\tau}$. And it can also be checked easily that the map

$$\Phi : \mathcal{P}(1,2,3) \to \mathbb{F}_5 : A \mapsto \begin{cases} 1 \text{ if } |A| = 0, \\ -2 \text{ if } |A| = 1, \\ -1 \text{ if } |A| = 2, \\ 2 \text{ if } |A| = 3, \end{cases}$$

is contained in $\mathcal{B}(X|\,\mathbb{F}_5)^{\rho} \cap \mathcal{B}(X|\,\mathbb{F}_5)^{\tau}$.

It might be of interest to determine, for every field K, the K-dimension of the intersection $\mathcal{B}(X|\,K)^{\rho} \cap \mathcal{B}(X|\,K)^{\tau}$ in terms of (i) the cardinality of X and (ii) the characteristic of K.

(3) It might also be of interest to determine explicit formulae for decomposing, for a given field K of characteristic 0, any given map $\Pi \in \mathcal{B}(X|\,K)$ into an orthogonal sum $\Pi = \Pi' + \Pi''$ of a map $\Pi' \in \mathcal{B}(X|\,K)^{\tau}$ and a map $\Pi'' \in \mathcal{B}(X|\,K)^{\rho}$.

(4) Finally, specialising to the case $K := \mathbb{F}_2$, recall first that

(i) associating, to each map Π in $\mathcal{B}(X|\,\mathbb{F}_2)$, its support

$$\operatorname{supp}(\Pi) := \{A \subseteq X : \Pi(A) \neq 0\}$$

yields a bijection between $\mathcal{B}(X|\,\mathbb{F}_2)$ and the set $\mathcal{P}\big(\mathcal{P}(X)\big)$ of all collections $\mathbf{p} \subset \mathcal{P}(X)$ of subsets of X,

(ii) the sum $\Pi + \Pi'$ of two such maps $\Pi, \Pi' \in \mathcal{B}(X|\,\mathbb{F}_2)$ corresponds to the symmetric difference $\mathbf{p}\Delta\mathbf{p}' := (\mathbf{p} - \mathbf{p}') \cup (\mathbf{p}' - \mathbf{p})$ of the associated collections $\mathbf{p} := \operatorname{supp}(\Pi)$ and $\mathbf{p}' := \operatorname{supp}(\Pi')$, i.e., one has

$$\operatorname{supp}(\Pi + \Pi') = \operatorname{supp}(\Pi)\ \Delta\ \operatorname{supp}(\Pi')$$

for all $\Pi, \Pi' \in \mathcal{B}(X|\,\mathbb{F}_2)$,

(iii) the collections $\mathbf{p}$ of subsets of X corresponding to the maps in $\mathcal{B}(X|\,\mathbb{F}_2)^{\tau}$ are, in consequence, those collections $\mathbf{p} \subseteq \mathcal{P}(X)$ of subsets of X for which, for every subset A of X, the number $\#\{A' \in \mathbf{p} : A' \subsetneq A\}$ of proper subsets of A that are elements of $\mathbf{p}$ is even,

(iv) the collections $\mathbf{p}'$ of subsets of X corresponding to the maps in $\mathcal{B}(X|\,\mathbb{F}_2)^{\rho}$ are those collections $\mathbf{p}' \subseteq \mathcal{P}(X)$ of subsets of X for which, for every subset A of X, the number $\#\{A' \in \mathbf{p}' : A \subsetneq A'\}$ of subsets in $\mathbf{p}'$ properly containing A is even,

(v) and the inner product $\langle\Pi|\Pi'\rangle$ of two maps $\Pi, \Pi' \in \mathcal{B}(X|\,\mathbb{F}_2)$ coincides with the parity of the intersection $\mathbf{p} \cap \mathbf{p}'$ of the corresponding two set systems $\mathbf{p} := \mathrm{supp}(\Pi)$ and $\mathbf{p}' := \mathrm{supp}(\Pi')$.

Thus, using what we have established above in conjunction with the fact, established in [12], that $\mathcal{B}(X|\,K)^{\tau}$ coincides, for every field K, with the image of the K-vectorspace $\mathcal{S}(X|K) := K^{\mathcal{S}(X)}$ consisting of all maps from $\mathcal{S}(X)$ into K relative to the map

$$\mathcal{D}_{\bullet}^{K} : \mathcal{S}(X|K) \to \mathcal{B}(X|\,K) : \Sigma \mapsto \big(\Sigma_{\bullet}^{K} : \mathcal{P}(X) \to K\big)$$

where, in analogy to the case $K{:=}\,\mathbb{R}$ discussed above, $\Sigma_{\bullet}^{K}$ denotes the map from $\mathcal{P}(X)$ into K that maps every subset A_0 of X onto the sum

$$\Sigma_{\bullet}^{K}(A_0) := \sum_{A \in \mathcal{P}(X - A_0)} \Sigma(\{A_0 \cup A, X - (A_0 \cup A)\}),$$

we see that the following holds:

Corollary 4.1. *There exists a canonical one-to-one correspondence between*

(i) *collections $\mathcal{S} \subseteq \mathcal{S}(X)$ of X-splits,*

(ii) *collections $\mathbf{p} \subseteq \mathcal{P}(X)$ of subsets of X for which, for every subset A of X, the number $\#\{A' \in \mathbf{p} : A' \subsetneq A\}$ of proper subsets of A that are elements of $\mathbf{p}$ is even,*

(iii) *collections $\mathbf{p} \subseteq \mathcal{P}(X)$ of subsets of X for which $\mathbf{p} \cap \mathbf{p}'$ is of even cardinality for all collections $\mathbf{p}' \subseteq \mathcal{P}(X)$ of subsets of X for which, for every subset A of X, the number $\#\{A' \in \mathbf{p}' : A \subsetneq A'\}$ of subsets in $\mathbf{p}'$ properly containing A is even.*

More specifically, associating to each split $S = \{A, B\}$ of X the set system $\mathbf{p}(S) := \{A_0 \subseteq X : \emptyset \neq A_0 \subseteq A\} \cup \{A_0 \subseteq X : \emptyset \neq A_0 \subseteq B\}$ yields a family $\big(\mathbf{p}(S)\big)_{S \in \mathcal{S}(X)}$ of subsets of $\mathcal{P}(X)$ such that a collection $\mathbf{p} \subseteq \mathcal{P}(X)$ of subsets of X is a symmetric difference

$$\Delta(\mathbf{p}(S) : S \in \mathcal{S}) := \{A \subseteq X : \#\{S \in \mathcal{S} : A \in \mathbf{p}(S)\} \textit{ is odd}\}$$

of set systems of the form $\mathbf{p}(S)$ for some system $\mathcal{S} = \mathcal{S}(\mathbf{p}) \subseteq \mathcal{S}(X)$ of X-splits if and only if, for any given subset A of X, the number $\#\{A' \in \mathbf{p} : A' \subsetneq A\}$ of elements in $\mathbf{p}$ that are proper subsets of A is even, in which case the set $\mathcal{S} = \mathcal{S}(\mathbf{p})$ is uniquely determined by $\mathbf{p}$.

It could be of interest to characterise those set systems $\mathbf{p} \subseteq \mathcal{P}(X)$ that correspond to compatible, cyclic, or weakly compatible split systems (as defined in [7] and [3], respectively, see also [2,4,5,8]).

References

[1] Backelin, J. and Linusson, S. (2006). Parity splits of X-trees, *Annals of Combinatorics*, **10**, 1-18.

[2] Bandelt, H.-J. (1990). Recognition of tree metrics, *SIAM J. Disc. Math.*, **3**, 1-6.

[3] Bandelt, H.-J. and Dress, A. W. M. (1992). A canonical split decomposition theory for metrics on a finite set, *Adv. Math.*, **92**, 47-105.

[4] Bandelt, H.-J. and Dress, A. W. M. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data *Molecular Phylogenetics and Evolution*, **1**, 242-252.

[5] Bandelt, H.-J. and Steel, M. A. (1995). Symmetric matrices representable by weighted trees over a cancellative abelian monoid, *SIAM Journal on Discrete Mathematics*, **8**, 517—525.

[6] Barker, G. M. (2002). Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation, *Biol. J. Linnean Soc.* **76**, 165—194.

[7] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, 387–395. Edinburgh University Press, Edinburgh.

[8] Dress, A. W. M., Huber, K., and Moulton, V. (2007). Some uses of the Farris Transform in Mathematics and Phylogenetics — A Review, *Annals of Combinatorics*, **11**, 1–37.

[9] Dress, A. W. M. and Steel, M. A. (2007). Phylogenetic diversity over an abelian group, Annals of Combinatorics, **11**, 143–160.

[10] Dress, A. W. M. (2005). Split decomposition over an abelian group, manuscript, Shanghai.

[11] Dress, A. W. M. (2006). Even set systems, manuscript, Shanghai.

[12] Dress, A. W. M. (2006). A note on group-valued split and set systems, manuscript, Shanghai.

[13] Evans, S. N. and Speed, T. P. (1993). Invariants of some probability models used in phylogenetic inference. *Annals of Statistics*, **21**, 355—377.

[14] Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**, 1—10.

[15] Joly, S. and Le Calvé, G. (1995). Three-way distances. *Journal of Classification* **12**, 191—205.

[16] Heiser, W. J. and Bennani, M. (1997). Triadic distance models: aximomatization and least squares representation. *Journal of Mathematical Psychology* **41**, 189—206.

[17] Pachter, L. and Speyer, D. (2004). Reconstructing trees from subtree weights. *Applied Mathematics Letters* **17(6)**, 615—621.

[18] Semple, C. and Steel, M. A. (2003). Phylogenetics. Oxford University Press.

[19] Steel, M. A. (2005). Phylogenetic diversity and the greedy algorithm. *Systematic Biology*, **54**, 527–529.

Andreas Dress
Department for Combinatorics and Geometry
CAS-MPG Partner Institute for Computational Biology
Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai, China
and Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22–26, D 04103 Leipzig, Germany
e-mail: dress@sibs.ac.cn and dress@mis.mpg.de

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 249–267

Exponential Convergence Analysis of DCNNs having Unbounded Activations and Inhibitory Self-Connections

Sannay Mohamad

Abstract. We investigate the exponential convergence characteristics of an equilibrium state of a delayed cellular neural network (DCNN) whose state variables are governed by a system of nonlinear integrodifferential equations with delays distributed continuously over unbounded intervals. The network is designed in such a way that the self-connections are inhibitory and instantaneous, and the activation functions are globally Lipschitz continuous and they are not necessarily bounded and differentiable. While monotonicity is generally not required for the activation functions, it is however needed for the activations of the inhibitory and instantaneous self-connections. By applying a Young inequality to an appropriate form of Lyapunov functionals, we establish the exponential convergence of the network towards a unique equilibrium state under a set of easily verifiable and delay independent sufficient conditions. It is shown that the restriction holding between the neural parameter values can be relaxed by the presence of the inhibitory and instantaneous self-connections and the corresponding monotonically increasing activation functions. The global exponential stability results obtained in this article will improve and extend the existing results which have been published in the literature on neural networks.

Mathematics Subject Classification (2000). Primary 34K20; Secondary 92B20.

Keywords. Cellular neural networks; Distributed delays; Inhibitory self-connections; Equilibria; Lyapunov functionals; Global exponential stability.

Results in this paper were presented at the International Conference on Mathematical Modelling and Computation held at the University of Brunei Darussalam during 5–8 June 2006, in conjunction with the 20th anniversary celebration of the foundation of the university.

1. Introduction

The potential applications of a delayed cellular neural network (DCNN) given by

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = -a_i x_i(t) + \sum_{j=1}^{m} b_{ij} f_j(x_j(t)) + \sum_{j=1}^{m} c_{ij} f_j(x_j(t-\tau_{ij})) + I_i \tag{1.1}$$

for solving a wide range of problems, such as occur in various scientific disciplines [28, 33, 39, 43, 44], have motivated many researchers to perform fundamental studies such as investigating the existence and uniqueness and examining the stability characteristics of an equilibrium state of the network [1, 2, 8, 10, 11, 14, 17, 18, 19, 21, 23, 27, 30, 40, 46, 45]. The outcomes of these studies led to a series of modifications to the original design of the network developed by Chua and Yang [15, 16] and Roska and Chua [41]. Among them we include the removal of the monotonicity and differentiability properties of the activation functions $f_i(\cdot)$, and waiving the symmetry arrangement within the connection weight matrices $[b_{ij}]_{m\times m}$ and $[c_{ij}]_{m\times m}$. Such modifications have been found to improve the computational performance as well as widening the scope of applicability of the network.

In view of its promising capability for solving problem areas which require fast computation in real time, a number of researchers have investigated the exponential convergence of the network (1.1) [4, 5, 6, 7, 9, 12, 13, 22, 29, 31, 34, 36, 37, 38, 48, 53, 54, 51, 56, 57]. Most of the results obtained relied on the construction (or the modification) of the Lyapunov functional given by

$$V(\boldsymbol{x}(t)) = \sum_{i=1}^{m} \alpha_i \Big(\mathrm{e}^{\mu t}|x_i(t) - x_i^*| + \sum_{j=1}^{m} |c_{ij}| L_j \mathrm{e}^{\mu\tau_{ij}} \int_{t-\tau_{ij}}^{t} \mathrm{e}^{\mu s}|x_i(s) - x_i^*|\mathrm{d}s\Big). \tag{1.2}$$

This functional is one of the many variations of the original functional developed by Gopalsamy [26]. By using (1.2) one can, through an appropriate analysis, extract the following sufficient condition (see for instance, Cao [9])

$$a_i > \sum_{j=1}^{m} \frac{\alpha_j}{\alpha_i}(|b_{ji}| + |c_{ji}|)L_i \tag{1.3}$$

under which the exponential convergence of the DCNN (1.1) is guaranteed. Accompanying the condition (1.3) are also the boundedness and the Lipschitz property of the activation functions $f_j(\cdot)$. We remark that monotonicity and differentiability of the activation functions were not assumed in the DCNN (1.1). There are other results found in the articles cited above which have been obtained by modifying the form of the functional $V(\cdot)$. These conditions have been found to be more relaxed than the condition (1.3) in terms of restricting the parametric values of a_i.

Although the results include various aspects of connectivity within the network, they are not adequate to capture a possible and useful relaxation, especially when the design of the network has inhibitory and instantaneous self-connections, namely, $b_{ii}, c_{ii} \leqslant 0$ and $\tau_{ii} = 0$. Under this proposed set-up the DCNN (1.1) can be envisaged by the system given by

$$\begin{aligned}\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} &= -a_i x_i(t) + (b_{ii}+c_{ii})g_i(x_i(t)) + \sum_{j=1,j\neq i}^{m} b_{ij} f_j(x_j(t)) \\ &\quad + \sum_{j=1,j\neq i}^{m} c_{ij} f_j(x_j(t-\tau_{ij})) + I_i.\end{aligned} \tag{1.4}$$

We have found in writing this article that the sufficient condition holding the neural parametric values of the DCNN (1.4) is less restrictive due to the presence of the inhibitory and instantaneous self-connections $(b_{ii}+c_{ii})g_i(x_i(\cdot))$. Moreover, such relaxation is further enhanced by employment of the activation functions $g_i(\cdot)$ which are strictly monotonic.

The other aspect that we have included in this article is the removal of the boundedness property of the activation functions. Such removal would allow one to apply the DCNN (1.4) for solving optimization problems in the presence of constraints of a more general type. One may refer to the articles [18, 20] for some discussions on the need to use unbounded activation functions such as the diode-like exponential-type functions that suit the constraint requirement of an optimization problem. One ought to be cautious, however, that the problem of analysing the existence of a unique equilibrium point of the network, when the activation functions are unbounded, may not be a straightforward case. It may even be possible that the network does not have an equilibrium state. It thus becomes our main intention to dedicate Section 3 of the article to an investigation of the existence of a unique equilibrium state of the DCNN (1.4) when the activation functions are unbounded. The sufficient conditions obtained in this section will be carried over to Section 4 for establishing the exponential convergence of the neural states toward the unique equilibrium state.

2. Model formulation

It is always assumed in the formulation of the DCNN (1.4) that the time delays are discrete so as to allow the processing and transfer of signals among the cellular units to occur. In the actual implementation of the network for practical purposes, however, it is always difficult to determine the values of these delays. One of the better alternatives is to assume the propagation of signals being distributed over a certain duration of time in a manner in which the distant past has less influence compared to the recent behaviour of the neural state. The duration over which the past effects affect the current state can extend over a finite or an infinite interval. By incorporating this type of time delay into the processing part of the network architecture of (1.4), the model becomes

$$\begin{aligned}\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} &= -a_i x_i(t) + (b_{ii}+c_{ii})g_i(x_i(t)) + \sum_{j=1,j\neq i}^{m} b_{ij} f_j(x_j(t)) \\ &\quad + \sum_{j=1,j\neq i}^{m} c_{ij} f_j\Big(\int_{-\infty}^{t} K_{ij}(t-s)x_j(s)\mathrm{d}s\Big) + I_i;\end{aligned}$$

for convenience we put it in the form

$$\begin{aligned}\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} &= -a_i x_i(t) + (b_{ii}+c_{ii})g_i(x_i(t)) + \sum_{j=1,j\neq i}^{m} b_{ij} f_j(x_j(t)) \\ &\quad + \sum_{j=1,j\neq i}^{m} c_{ij} f_j\Big(\int_0^\infty K_{ij}(s) x_j(t-s)\mathrm{d}s\Big) + I_i\end{aligned} \tag{2.1}$$

for $i \in \mathcal{I} = \{1, 2, \ldots, m\}$, $t > 0$. The state $x_i(t)$ of the cell i at time t is measured in terms of its voltage, whose rate of change is influenced by the current received from the inhibitory and instantaneous self-connectivity $(b_{ii}+c_{ii})g_i(x_i(t))$, by an external input current source I_i, and also by an input current received from a neighbouring cell j through the nonlinear or piecewise linear activation function $f_j(\cdot)$. The connection strengths between cells i and j at times t and $t-s$ are parameterized respectively by the constants b_{ij} and c_{ij}. The parameter $K_{ij}(\cdot)$ corresponds to the delay kernel that controls the past effect received from the cell j, which influences the recent neural state behaviour of the cell i. The parameter a_i denotes the rate at which the cell i resets its potential to its resting state when isolated from other cells and inputs.

In studying the DCNN (2.1), the neural parameters are assumed to satisfy

$$\begin{aligned}&a_i > 0, \quad b_{ii}, c_{ii} \leqslant 0, \quad I_i \in \mathbb{R} \quad \text{for } i \in \mathcal{I}, \\ &b_{ij}, c_{ij} \in \mathbb{R} \quad \text{for } i, j \in \mathcal{I}, \quad i \neq j.\end{aligned} \tag{2.2}$$

The activation functions $g_i(\cdot)$ and $f_i(\cdot)$ with $g_i(0) = f_i(0) = 0$ are assumed to be globally Lipschitzian, in the sense that there exist positive constants κ_i, $\overline{\kappa}_i$ and L_i for which

$$\kappa_i \leqslant \frac{g_i(u)-g_i(v)}{u-v} \leqslant \overline{\kappa}_i, \quad |f_i(u)-f_i(v)| \leqslant L_i|u-v| \tag{2.3}$$

for all $u, v \in \mathbb{R}$. We remark that (2.3) does not necessarily mean that both functions are bounded and differentiable. The functions $g_i(\cdot)$ are monotonically increasing while $f_i(\cdot)$ are not necessarily so. We have found in this article and also in [4, 35] that the monotonicity property of $g_i(\cdot)$ is needed in the attempt to ease the restriction imposed on the parameter a_i. In the analysis below, we assume that $|g_i(u)|, |f_i(u)| \to \infty$ as $|u| \to \infty$.

The delay kernels $K_{ij} : [0,\infty) \to [0,\infty)$ in (2.1) denote continuous functions and they are assumed to satisfy

$$\int_0^\infty K_{ij}(s)\mathrm{d}s = 1 \quad \text{and} \quad \int_0^\infty K_{ij}(s)\mathrm{e}^{\mu s}\mathrm{d}s < \infty \tag{2.4}$$

for some positive constant μ. A typical class of the delay kernels is given by $K_{ij}(s) := K_r(s) = \frac{s^r}{r!}\gamma_{ij}^{r+1}\mathrm{e}^{-\gamma_{ij}s}$ for $s \in [0,\infty]$, where $\gamma_{ij} \in (0,\infty)$, $r \in \{0,1,2,\ldots\}$. These kernels have been used by a number of authors [25, 38, 47, 51, 55] in various stability investigations and applications of neural networks with distributed delays. One observes that $K_r(\cdot) \to \delta(\cdot)$ as $r \to \infty$, where $\delta(\cdot)$ denotes a Dirac delta function. We can use this to extract the delay terms in the DCNN (1.4) from those of (2.1). For instance,

$$\int_{-\infty}^{t} K(t-\tau)x(\tau)\mathrm{d}\tau = \int_{0}^{\infty} K(\tau)x(t-\tau)\mathrm{d}\tau$$

and

$$f\Big(\int_{-\infty}^{t} K(t-\tau)x_j(\tau)\mathrm{d}\tau\Big) \to f(x(t-\tau)) \quad \text{as } K(t-\tau) \to \delta(t-\tau).$$

It is for this reason one claims that the DCNN (1.4) denotes a special case of the DCNN (2.1).

We supplement the DCNN (2.1) with an initial condition of the form

$$x_i(s) = \phi_i(s) \quad \text{for } i \in \mathcal{I}, \quad s \in [-\infty, 0], \tag{2.5}$$

where each function $\phi_i(\cdot)$ is bounded and continuous on $(-\infty, 0]$. A neural state vector of (2.1) is denoted by $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_m(t))^{\mathrm{T}}$ for $t > 0$, where each component $x_i(\cdot)$ satisfies (2.1) and (2.5).

In this study, we associate a vector $\boldsymbol{u} \in \mathbb{R}^m$ with the general Euclidean norm $\|\boldsymbol{u}\|_p = (\sum_{i=1}^m |u_i|^p)^{1/p}$, where $p > 1$. And, we use a Young inequality (see for instance Beckenbach [3]) which states that for r, s satisfying $r > 1$ and $\frac{1}{r} + \frac{1}{s} = 1$, we have

$$uv \leqslant \frac{u^r}{r} + \frac{v^s}{s} \quad \text{for any } u, v > 0. \tag{2.6}$$

We refer to the articles [4, 12, 52] for the use of the inequality (2.6) in the stability investigations of delayed cellular neural networks.

3. Existence and uniqueness theorem

We investigate the existence of a unique equilibrium state $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_m^*)^{\mathrm{T}}$ of the DCNN (2.1), the existence of which is an important prerequisite for the global exponential stability of the DCNN (2.1). By letting $x_i(t) = x_i^*$ in (2.1) and applying (2.4), one can derive the algebraic system

$$-a_i x_i^* + (b_{ii} + c_{ii}) g_i(x_i^*) + \sum_{j=1, j\neq i}^{m} (b_{ij} + c_{ij}) f_j(x_j^*) + I_i = 0 \tag{3.1}$$

for $i \in \mathcal{I}$ that governs the components of $\boldsymbol{x}^*$.

In the following theorem, we prove the existence of a unique equilibrium state $\boldsymbol{x}^*$ under the sufficient condition (3.2) given below. The method of proof is based on constructing a continuous mapping on $\mathbb{R}^m$. If the mapping is a homeomorphism on $\mathbb{R}^m$, then the existence of a unique fixed point of the mapping can be guaranteed. This technique has been used elsewhere in the literature on neural networks [12, 18, 20, 51] wherein the usual boundedness condition of the activation functions has been removed. We remark that unbounded activations in neural networks have also been considered by Gopalsamy [24] who used a contraction mapping principle for establishing the uniqueness of the equilibrium state.

Theorem 3.1. *Let $p > 1$ be a real number, and let the assumptions* (2.2)–(2.4) *hold. Suppose the condition*

$$\begin{aligned} a_i - (b_{ii} + c_{ii})\kappa_i > \frac{p-1}{p} \sum_{j=1,j\neq i}^{m} \Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}} + |c_{ij}|^{\sigma_{ij}\frac{p}{p-1}} \Big) L_j^{\gamma_{ij}\frac{p}{p-1}} \\ + \frac{1}{p} \sum_{j=1,j\neq i}^{m} \frac{\alpha_j}{\alpha_i} \Big(|b_{ji}|^{\delta^*_{ji}p} + |c_{ji}|^{\sigma^*_{ji}p} \Big) L_i^{\gamma^*_{ji}p} \end{aligned} \tag{3.2}$$

is satisfied, where the α_i denote positive numbers, and δ_{ij}, δ^*_{ij}, σ_{ij}, σ^*_{ij}, γ_{ij}, γ^*_{ij} denote real numbers that satisfy $\delta_{ij} + \delta^*_{ij} = 1$, $\sigma_{ij} + \sigma^*_{ij} = 1$ and $\gamma_{ij} + \gamma^*_{ij} = 1$ for $i \neq j$. Then there exists a unique equilibrium state $\boldsymbol{x}^*$ of the DCNN (2.1).

Proof. We construct a map $\boldsymbol{h}(\boldsymbol{u}) \in C^0(\mathbb{R}^m, \mathbb{R}^m)$ defined by $\boldsymbol{h}(\boldsymbol{u}) = (h_1(\boldsymbol{u}), h_2(\boldsymbol{u}), \ldots, h_m(\boldsymbol{u}))^{\mathrm{T}}$ where

$$h_i(\boldsymbol{u}) = -a_i u_i + (b_{ii} + c_{ii}) g_i(u_i) + \sum_{j=1,j\neq i}^{m} (b_{ij} + c_{ij}) f_j(u_j) + I_i \tag{3.3}$$

for $u_i \in \mathbb{R}$. In order for the map $\boldsymbol{h}$ to be a homeomorphism on $\mathbb{R}^m$, one has to show that the mapping is injective on $\mathbb{R}^m$ and it satisfies $\|\boldsymbol{h}(\boldsymbol{u})\|_p \to \infty$ as $\|\boldsymbol{u}\|_p \to \infty$.

In the following, we show that the map $\boldsymbol{h}$ is injective on $\mathbb{R}^m$, namely, $\boldsymbol{h}(\boldsymbol{u}) = \boldsymbol{h}(\boldsymbol{v})$ implies $\boldsymbol{u} = \boldsymbol{v}$ for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$. We have

$$\begin{aligned} &- a_i u_i + (b_{ii} + c_{ii}) g_i(u_i) + \sum_{j=1,j\neq i}^{m} (b_{ij} + c_{ij}) f_j(u_j) + I_i \\ &= -a_i v_i + (b_{ii} + c_{ii}) g_i(v_i) + \sum_{j=1,j\neq i}^{m} (b_{ij} + c_{ij}) f_j(v_j) + I_i. \end{aligned}$$

Noting that $a_i - (b_{ii} + c_{ii})\kappa_i > 0$ and applying the assumption (2.3) to the above, we obtain

$$\begin{aligned} a_i(u_i - v_i) &= (b_{ii} + c_{ii})[g_i(u_i) - g_i(v_i)] + \sum_{j=1,j\neq i}^{m} (b_{ij} + c_{ij})[f_j(u_j) - f_j(v_j)] \\ &\leqslant (b_{ii} + c_{ii})\kappa_i(u_i - v_i) + \sum_{j=1,j\neq i}^{m} (b_{ij} + c_{ij})[f_j(u_j) - f_j(v_j)] \end{aligned}$$

which then gives

$$[a_i - (b_{ii} + c_{ii})\kappa_i]|u_i - v_i| \leqslant \sum_{j=1,j\neq i}^{m} (|b_{ij}| + |c_{ij}|) L_j |u_j - v_j|.$$

Now,

$$
\begin{aligned}
&\sum_{i=1}^{m}\alpha_i[a_i-(b_{ii}+c_{ii})\kappa_i]|u_i-v_i|^p\\
&\leqslant\sum_{i=1}^{m}\alpha_i\sum_{j=1,j\neq i}^{m}(|b_{ij}|+|c_{ij}|)L_j|u_i-v_i|^{p-1}|u_j-v_j|\\
&=\sum_{i=1}^{m}\alpha_i\sum_{j=1,j\neq i}^{m}(|b_{ij}|^{\delta_{ij}}L_j^{\gamma_{ij}}|u_i-v_i|^{p-1})(|b_{ij}|^{\delta^*_{ij}}L_j^{\gamma^*_{ij}}|u_j-v_j|)\\
&\quad+\sum_{i=1}^{m}\alpha_i\sum_{j=1,j\neq i}^{m}(|c_{ij}|^{\sigma_{ij}}L_j^{\gamma_{ij}}|u_i-v_i|^{p-1})(|c_{ij}|^{\sigma^*_{ij}}L_j^{\gamma^*_{ij}}|u_j-v_j|),
\end{aligned}
$$

where $\delta_{ij}+\delta^*_{ij}=1$, $\gamma_{ij}+\gamma^*_{ij}=1$ and $\sigma_{ij}+\sigma^*_{ij}=1$. By using the Young inequality (2.6) in the above, we obtain

$$
\begin{aligned}
&\sum_{i=1}^{m}\alpha_i[a_i-(b_{ii}+c_{ii})\kappa_i]|u_i-v_i|^p\\
&\leqslant\sum_{i=1}^{m}\alpha_i\sum_{j=1,j\neq i}^{m}\Big(\frac{p-1}{p}|b_{ij}|^{\delta_{ij}\frac{p}{p-1}}L_j^{\gamma_{ij}\frac{p}{p-1}}|u_i-v_i|^p+\frac{1}{p}|b_{ij}|^{\delta^*_{ij}p}L_j^{\gamma^*_{ij}p}|u_j-v_j|^p\Big)\\
&\quad+\sum_{i=1}^{m}\alpha_i\sum_{j=1,j\neq i}^{m}\Big(\frac{p-1}{p}|c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}L_j^{\gamma_{ij}\frac{p}{p-1}}|u_i-v_i|^p+\frac{1}{p}|c_{ij}|^{\sigma^*_{ij}p}L_j^{\gamma^*_{ij}p}|u_j-v_j|^p\Big)\\
&\leqslant\sum_{i=1}^{m}\alpha_i\,\frac{p-1}{p}\sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}}+|c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}\Big)L_j^{\gamma_{ij}\frac{p}{p-1}}|u_i-v_i|^p\\
&\quad+\sum_{i=1}^{m}\alpha_i\,\frac{1}{p}\sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}\Big(|b_{ji}|^{\delta^*_{ji}p}+|c_{ji}|^{\sigma^*_{ji}p}\Big)L_i^{\gamma^*_{ji}p}|u_i-v_i|^p
\end{aligned}
$$

which then gives

$$
\begin{aligned}
&\sum_{i=1}^{m}\alpha_i\Big\{[a_i-(b_{ii}+c_{ii})\kappa_i]-\frac{p-1}{p}\sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}}+|c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}\Big)L_j^{\gamma_{ij}\frac{p}{p-1}}\\
&-\frac{1}{p}\sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}\Big(|b_{ji}|^{\delta^*_{ji}p}+|c_{ji}|^{\sigma^*_{ji}p}\Big)L_i^{\gamma^*_{ji}p}\Big\}|u_i-v_i|^p\leqslant 0.
\end{aligned}
$$

It is not difficult to see that this system under the condition (3.2) yields $u_i=v_i$ which means $\boldsymbol{u}=\boldsymbol{v}$. Hence, the map $\boldsymbol{h}$ is injective on $\mathbb{R}^m$.

To show that $\|\boldsymbol{h}(\boldsymbol{u})\|_p\to\infty$ as $\|\boldsymbol{u}\|_p\to\infty$, we consider for convenience the map $\hat{\boldsymbol{h}}(\boldsymbol{u})=\boldsymbol{h}(\boldsymbol{u})-\boldsymbol{h}(\boldsymbol{0})$, where

$$
\hat{h}_i(\boldsymbol{u})=h_i(\boldsymbol{u})-h_i(\boldsymbol{0})=-a_iu_i+(b_{ii}+c_{ii})g_i(u_i)+\sum_{j=1,j\neq i}^{m}(b_{ij}+c_{ij})f_j(u_j)
$$

for $u_i \in \mathbb{R}$. Let us assume that $\left\|\hat{\boldsymbol{h}}(\boldsymbol{u})\right\|_p \to \infty$ is not true as $\|\boldsymbol{u}\|_p \to \infty$. In other words, there is a sequence $\{\boldsymbol{u}_n\}$ such that $\|\boldsymbol{u}_n\|_p \to \infty$ and $\left\|\hat{\boldsymbol{h}}(\boldsymbol{u}_n)\right\|_p$ is bounded as $n \to \infty$. One can pick a subsequence $\{\boldsymbol{u}'_n\}$ of $\{\boldsymbol{u}_n\}$ such that $\|\boldsymbol{u}'_n\|_p \to \infty$ and $\left\|\hat{\boldsymbol{h}}(\boldsymbol{u}'_n)\right\|_p \to \lambda$ as $n \to \infty$. Now,

$$\begin{aligned}
&\sum_{i=1}^{m} \alpha_i |u'_{in}|^{p-1} \operatorname{sgn}(u'_{in}) \hat{h}_i(\boldsymbol{u}'_n) \\
&= \sum_{i=1}^{m} \alpha_i |u'_{in}|^{p-1} \operatorname{sgn}(u'_{in}) \Big\{ -a_i u'_{in} + (b_{ii} + c_{ii}) g_i(u'_{in}) + \sum_{j=1, j\neq i}^{m} (b_{ij} + c_{ij}) f_j(u'_{jn}) \Big\} \\
&\leqslant \sum_{i=1}^{m} \alpha_i \Big\{ -[a_i - (b_{ii} + c_{ii})\kappa_i] |u'_{in}|^p + \sum_{j=1, j\neq i}^{m} (|b_{ij}| + |c_{ij}|) L_j |u'_{in}|^{p-1} |u'_{jn}| \Big\}.
\end{aligned}$$

One can apply similar steps like before to obtain

$$\begin{aligned}
&\sum_{i=1}^{m} \alpha_i |u'_{in}|^{p-1} \operatorname{sgn}(u'_{in}) \hat{h}_i(\boldsymbol{u}'_n) \\
&\leqslant -\sum_{i=1}^{m} \alpha_i \Big\{ [a_i - (b_{ii} + c_{ii})]\kappa_i - \frac{p-1}{p} \sum_{j=1, j\neq i}^{m} \left(|b_{ij}|^{\delta_{ij} \frac{p}{p-1}} + |c_{ij}|^{\sigma_{ij} \frac{p}{p-1}} \right) L_j^{\gamma_{ij} \frac{p}{p-1}} \\
&\quad - \frac{1}{p} \sum_{j=1, j\neq i}^{m} \frac{\alpha_j}{\alpha_i} \left(|b_{ji}|^{\delta^*_{ji} p} + |c_{ji}|^{\sigma^*_{ji} p} \right) L_i^{\gamma^*_{ji} p} \Big\} |u'_{in}|^p \\
&\leqslant -\varepsilon \sum_{i=1}^{m} \alpha_i |u'_{in}|^p,
\end{aligned}$$

where

$$\begin{aligned}
\varepsilon = \min_{i\in\mathcal{I}} \Big\{ &[a_i - (b_{ii} + c_{ii})]\kappa_i - \frac{p-1}{p} \sum_{j=1, j\neq i}^{m} \left(|b_{ij}|^{\delta_{ij} \frac{p}{p-1}} + |c_{ij}|^{\sigma_{ij} \frac{p}{p-1}} \right) L_j^{\gamma_{ij} \frac{p}{p-1}} \\
&- \frac{1}{p} \sum_{j=1, j\neq i}^{m} \frac{\alpha_j}{\alpha_i} \left(|b_{ji}|^{\delta^*_{ji} p} + |c_{ji}|^{\sigma^*_{ji} p} \right) L_i^{\gamma^*_{ji} p} \Big\} > 0.
\end{aligned}$$

We then have

$$\begin{aligned}
\varepsilon \min_{i\in\mathcal{I}} \{\alpha_i\} \sum_{i=1}^{m} |u'_{in}|^p &\leqslant -\sum_{i=1}^{m} \alpha_i |u'_{in}|^{p-1} \operatorname{sgn}(u'_{in}) \hat{h}_i(\boldsymbol{u}'_n) \\
&\leqslant \max_{i\in\mathcal{I}} \{\alpha_i\} \sum_{i=1}^{m} |u'_{in}|^{p-1} |\hat{h}_i(\boldsymbol{u}'_n)|.
\end{aligned}$$

By applying a Hölder inequality to the above, we obtain

$$\sum_{i=1}^{m} |u'_{in}|^p \leqslant \frac{1}{\varepsilon} \frac{\max_{i\in\mathcal{I}}\{\alpha_i\}}{\min_{i\in\mathcal{I}}\{\alpha_i\}} \Big(\sum_{i=1}^{m} |u'_{in}|^{(p-1)q} \Big)^{1/q} \Big(\sum_{i=1}^{m} |\hat{h}_i(\boldsymbol{u}'_n)|^p \Big)^{1/p},$$

where $\frac{1}{p} + \frac{1}{q} = 1$, which in turn yields

$$\left\|\boldsymbol{u}'_n\right\|_p \leqslant \frac{1}{\varepsilon}\, \frac{\max_{i\in\mathcal{I}}\{\alpha_i\}}{\min_{i\in\mathcal{I}}\{\alpha_i\}} \left\|\hat{\boldsymbol{h}}(\boldsymbol{u}'_n)\right\|_p = \frac{\lambda}{\varepsilon}\, \frac{\max_{i\in\mathcal{I}}\{\alpha_i\}}{\min_{i\in\mathcal{I}}\{\alpha_i\}} \quad \text{as } n \to \infty.$$

This contradicts our choice of $\{\boldsymbol{u}'_n\}$ which satisfies $\left\|\boldsymbol{u}'_n\right\|_p \to \infty$ as $n \to \infty$. It then follows that the map $\boldsymbol{h}$ satisfies $\left\|\boldsymbol{h}(\boldsymbol{u})\right\|_p \to \infty$ as $\left\|\boldsymbol{u}\right\|_p \to \infty$.

We have established that the map $\boldsymbol{h}$ is a homeomorphism on $\mathbb{R}^m$ and this guarantees the existence of a unique fixed point $\boldsymbol{x}^*$ of the map $\boldsymbol{h}$. This fixed point represents the unique equilibrium state of the DCNN (2.1). The proof is now complete. □

4. Exponential stability theorem

We proceed to analyse the global exponential stability of the unique equilibrium state $\boldsymbol{x}^*$ of the network (2.1) under the sufficient condition (3.2). Let us denote

$$\left\|\boldsymbol{\phi} - \boldsymbol{x}^*\right\|_p = \left[\sum_{i=1}^{m} \Big(\sup_{s\in[-\infty,0]} |\phi_i(s) - x_i^*|^p \Big)\right]^{1/p} < \infty,$$

where $p > 1$ is a real number, and $\phi_i(s)$ for $s \in (-\infty, 0]$ denote arbitrary initial value functions of the network (2.1).

Now, we are ready to prove the exponential stability of the equilibrium state $\boldsymbol{x}^*$ in the norm $\left\|\cdot\right\|_p$. For the convenience of the reader, we give below the definition of exponential stability. For definitions of exponential stability involving the usual norms (i.e., $\left\|\cdot\right\|_1$, $\left\|\cdot\right\|_2$ and $\left\|\cdot\right\|_\infty$) and the general norm $\left\|\cdot\right\|_p$, we refer the reader to the articles [5, 9, 37, 42]. We remark that Sasagawa [42] defined the exponential stability involving the norm $\left\|\cdot\right\|_p$ as the exponential p-stability. Though the definitions of exponential stability distinguished under different norms is not an important notion due to the equivalence in norms within the Euclidean space $\mathbb{R}^m$, the sufficient conditions obtained under the general norm $\left\|\cdot\right\|_p$ certainly can generate a family of conditions, which in turn provides us with a vast range for choosing the neural parametric values that will ensure the exponential convergence of the network (2.1).

Definition 4.1. The unique equilibrium state $\boldsymbol{x}^*$ of the network (2.1) is said to be globally exponentially stable if there exist real numbers $\beta \geqslant 1$ and $\nu > 0$ for which

$$\left[\sum_{i=1}^{m} |x_i(t) - x_i^*|^p\right]^{1/p} \leqslant \beta \mathrm{e}^{-\nu t} \left\|\boldsymbol{\phi} - \boldsymbol{x}^*\right\|_p \quad \text{for } t > 0,$$

where the constants β, ν are independent of the initial values of the network (2.1).

Theorem 4.2. *Let $p > 1$ be a real number, and let the assumptions* (2.2)–(2.4) *hold. If the condition* (3.2) *is satisfied, then the unique equilibrium state $\boldsymbol{x}^*$ of the network* (2.1) *is globally exponentially stable.*

Proof. The uniqueness of the equilibrium state $\boldsymbol{x}^*$ of the network (2.1) follows from Theorem 3.1. Let $\boldsymbol{x}(t)$ denote a solution of the network (2.1) corresponding to a given but arbitrary initial value function $\boldsymbol{\phi}$. We obtain that

$$\begin{aligned}\frac{\mathrm{d}(x_i(t)-x_i^*)}{\mathrm{d}t} &= -a_i(x_i(t)-x_i^*)+(b_{ii}+c_{ii})(g_i(x_i(t))-g_i(x_i^*))\\ &\quad+\sum_{j=1,j\neq i}^{m} b_{ij}[f_j(x_j(t))-f_j(x_j^*)]\\ &\quad+\sum_{j=1,j\neq i}^{m} c_{ij}\Big[f_j\Big(\int_0^\infty K_{ij}(s)x_j(t-s)\mathrm{d}s\Big)-f_j\Big(\int_0^\infty K_{ij}(s)x_j^*\mathrm{d}s\Big)\Big].\end{aligned}$$

By applying the upper right derivative $\frac{\mathrm{d}^+}{\mathrm{d}t}|x_i(t)-x_i^*| = \frac{\mathrm{d}}{\mathrm{d}t}(x_i(t)-x_i^*)\mathrm{sgn}(x_i(t)-x_i^*)$ and the assumptions (2.2) and (2.3) to the above, we obtain the system

$$\begin{aligned}\frac{\mathrm{d}^+}{\mathrm{d}t}|x_i(t)-x_i^*| &\leqslant -[a_i-(b_{ii}+c_{ii})\kappa_i]|x_i(t)-x_i^*| + \sum_{j=1,j\neq 1}^{m}|b_{ij}|L_j|x_j(t)-x_j^*|\\ &\quad+\sum_{j=1,j\neq i}^{m}|c_{ij}|L_j\int_0^\infty K_{ij}(s)|x_j(t-s)-x_j^*|\mathrm{d}s\end{aligned} \tag{4.1}$$

for $t > 0$.

Let us introduce a number $0 < \nu < \mu$, where μ is defined in (2.4), to the condition (3.2) so that

$$\begin{aligned}&a_i-(b_{ii}+c_{ii})\kappa_i-\nu\\ &-\frac{p-1}{p}\sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}}+|c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}\mathrm{d}s\Big)L_j^{\gamma_{ij}\frac{p}{p-1}}\\ &-\frac{1}{p}\sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}\Big(|b_{ji}|^{\delta_{ji}^*p}+|c_{ji}|^{\sigma_{ji}^*p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}\mathrm{d}s\Big)L_i^{\gamma_{ji}^*p}\geqslant 0\end{aligned} \tag{4.2}$$

for all $i \in \mathcal{I}$. Corresponding to this number ν, we define nonnegative functions $w_i(\cdot)$ by

$$w_i(t) = \mathrm{e}^{\nu t}|x_i(t)-x_i^*| \quad \text{for } t \in (-\infty,\infty) \tag{4.3}$$

and from which we derive the following:

$$\begin{aligned}\frac{\mathrm{d}^+ w_i(t)}{\mathrm{d}t} &= \nu \mathrm{e}^{\nu t}|x_i(t)-x_i^*| + \mathrm{e}^{\nu t}\frac{\mathrm{d}^+}{\mathrm{d}t}|x_i(t)-x_i^*| \\ &\leqslant \nu \mathrm{e}^{\nu t}|x_i(t)-x_i^*| - [a_i-(b_{ii}+c_{ii})\kappa_i]\mathrm{e}^{\nu t}|x_i(t)-x_i^*| \\ &\quad + \sum_{j=1,j\neq i}^{m}|b_{ij}|L_j\mathrm{e}^{\nu t}|x_j(t)-x_j^*| \\ &\quad + \sum_{j=1,j\neq i}^{m}|c_{ij}|L_j\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu t}|x_j(t-s)-x_j^*|\mathrm{d}s\end{aligned}$$

and hence

$$\begin{aligned}\frac{\mathrm{d}^+ w_i(t)}{\mathrm{d}t} &\leqslant -[a_i-(b_{ii}+c_{ii})\kappa_i-\nu]w_i(t) + \sum_{j=1,j\neq i}^{m}|b_{ij}|L_j w_j(t) \\ &\quad + \sum_{j=1,j\neq i}^{m}|c_{ij}|L_j\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}w_j(t-s)\mathrm{d}s\end{aligned} \tag{4.4}$$

for $t > 0$. Associated with the solution $w_i(\cdot)$ of (4.4) is a Lyapunov functional $V(t) = V(\boldsymbol{w}(t))$ defined by

$$V(t) = \sum_{i=1}^{m}\alpha_i\Big(w_i^p(t) + \sum_{j=1,j\neq i}^{m}|c_{ij}|^{\sigma_{ij}^* p}L_j^{\gamma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}\Big(\int_{t-s}^{t} w_j^p(r)\mathrm{d}r\Big)\mathrm{d}s\Big) \tag{4.5}$$

for $t > 0$. We shall see in the following that this nonnegative functional is nonincreasing and bounded above by $V(0)$, namely, $V(t) \leqslant V(0) < \infty$ for all $t > 0$. Firstly, we observe that

$$\begin{aligned}V(0) &= \sum_{i=1}^{m}\alpha_i\Big(w_i^p(0) + \sum_{j=1,j\neq i}^{m}|c_{ij}|^{\sigma_{ij}^* p}L_j^{\gamma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}\Big(\int_{-s}^{0} w_j^p(r)\mathrm{d}r\Big)\mathrm{d}s\Big) \\ &= \sum_{i=1}^{m}\alpha_i\Big(w_i^p(0) + \sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}|c_{ji}|^{\sigma_{ji}^* p}L_i^{\gamma_{ji}^* p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}\Big(\int_{-s}^{0} w_i^p(r)\mathrm{d}r\Big)\mathrm{d}s\Big) \\ &\leqslant \sum_{i=1}^{m}\alpha_i\Big(1 + \sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}|c_{ji}|^{\sigma_{ji}^* p}L_i^{\gamma_{ji}^* p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}s\,\mathrm{d}s\Big)\Big(\sup_{s\in(-\infty,0]} w_i^p(s)\Big).\end{aligned}$$

Since $\sup_{s\in(-\infty,0]}|x_i(s)-x_i^*|^p < \infty$ and

$$\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}s\mathrm{d}s \leqslant \int_0^\infty K_{ji}(s)\mathrm{e}^{\mu s}\mathrm{d}s < \infty$$

for $0 < \nu < \mu$, we assert therefore that $V(0) < \infty$. Next, we calculate the rate of change of $V(\cdot)$ along the solution of (4.4) to obtain

$$
\begin{aligned}
\frac{\mathrm{d}^+V(t)}{\mathrm{d}t} &= \sum_{i=1}^{m}\alpha_i\Big[p w_i^{p-1}(t)\frac{\mathrm{d}^+w_i(t)}{\mathrm{d}t} \\
&\quad + \sum_{j=1,j\neq i}^{m} |c_{ij}|^{\sigma_{ij}^* p} L_j^{\gamma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}[w_j^p(t)-w_j^p(t-s)]\mathrm{d}s\Big] \\
&\leqslant \sum_{i=1}^{m}\alpha_i\Big[-p[a_i-(b_{ii}+c_{ii})\kappa_i-\nu]w_i^p(t) + p\sum_{j=1,j\neq i}^{m}|b_{ij}|L_j w_i^{p-1}(t)w_j(t) \\
&\quad + p\sum_{j=1,j\neq i}^{m}|c_{ij}|L_j\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}w_i^{p-1}(t)w_j(t-s)\mathrm{d}s \\
&\quad + \sum_{j=1,j\neq i}^{m} |c_{ij}|^{\sigma_{ij}^* p} L_j^{\gamma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}[w_j^p(t)-w_j^p(t-s)]\mathrm{d}s\Big] \\
&\leqslant \sum_{i=1}^{m}\alpha_i\Big[-p[a_i-(b_{ii}+c_{ii})\kappa_i-\nu]w_i^p(t) \\
&\quad + \sum_{j=1,j\neq i}^{m}\Big((p-1)|b_{ij}|^{\delta_{ij}\frac{p}{p-1}}L_j^{\gamma_{ij}\frac{p}{p-1}}w_i^p(t) + |b_{ij}|^{\delta_{ij}^* p}L_j^{\gamma_{ij}^* p}w_j^p(t)\Big) \\
&\quad + \sum_{j=1,j\neq i}^{m}\Big((p-1)|c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}L_j^{\gamma_{ij}\frac{p}{p-1}}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}w_i^p(t)\mathrm{d}s \\
&\quad + |c_{ij}|^{\sigma_{ij}^* p}L_j^{\gamma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}w_j^p(t-s)\mathrm{d}s\Big) \\
&\quad + \sum_{j=1,j\neq i}^{m} |c_{ij}|^{\sigma_{ij}^* p} L_j^{\gamma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}[w_j^p(t)-w_j^p(t-s)]\mathrm{d}s\Big] \\
&= \sum_{i=1}^{m}\alpha_i\Big[-p[a_i-(b_{ii}+c_{ii})\kappa_i-\nu]w_i^p(t) \\
&\quad + (p-1)\sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}} + |c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}\mathrm{d}s\Big)L_j^{\gamma_{ij}\frac{p}{p-1}}w_i^p(t) \\
&\quad + \sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}^* p} + |c_{ij}|^{\sigma_{ij}^* p}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}\mathrm{d}s\Big)L_j^{\gamma_{ij}^* p}w_j^p(t)\Big]
\end{aligned}
$$

which can be summarised as

$$
\begin{aligned}
\frac{\mathrm{d}^+V(t)}{\mathrm{d}t} &\leqslant -\sum_{i=1}^{m}p\alpha_i\Big[[a_i-(b_{ii}+c_{ii})\kappa_i-\nu] \\
&\quad - \frac{p-1}{p}\sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}} + |c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}\int_0^\infty K_{ij}(s)\mathrm{e}^{\nu s}\mathrm{d}s\Big)L_j^{\gamma_{ij}\frac{p}{p-1}}
\end{aligned}
$$

$$-\frac{1}{p}\sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}\Big(|b_{ji}|^{\delta^*_{ji}p}+|c_{ji}|^{\sigma^*_{ji}p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}\mathrm{d}s\Big)L_i^{\gamma^*_{ji}p}\Big]w_i^p(t)$$

for $t>0$. By using (4.2) in the above, we have $\mathrm{d}^+V(t)/\mathrm{d}t\leqslant 0$ for $t>0$. This in turn implies that $V(t)\leqslant V(0)<\infty$ for all $t>0$.

We obtain from (4.5) that

$$\sum_{i=1}^{m}\alpha_i w_i^p(t)\leqslant\sum_{i=1}^{m}\Big(\alpha_i+\sum_{j=1,j\neq i}^{m}\alpha_j|c_{ji}|^{\sigma^*_{ji}p}L_i^{\gamma^*_{ji}p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}s\mathrm{d}s\Big)\Big(\sup_{s\in(-\infty,0]}w_i^p(s)\Big)$$

for $t>0$. By using (4.3) in the above, we obtain

$$\begin{aligned}&\sum_{i=1}^{m}\alpha_i\mathrm{e}^{p\nu t}|x_i(t)-x_i^*|^p\\&\quad\leqslant\sum_{i=1}^{m}\Big(\alpha_i+\sum_{j=1,j\neq i}^{m}\alpha_j|c_{ji}|^{\sigma^*_{ji}p}L_i^{\gamma^*_{ji}p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}s\mathrm{d}s\Big)\Big(\sup_{s\in(-\infty,0]}|x_i(s)-x_i^*|^p\Big),\end{aligned}$$

which in turn gives

$$\sum_{i=1}^{m}|x_i(t)-x_i^*|^p\leqslant\gamma\mathrm{e}^{-p\nu t}\sum_{i=1}^{m}\Big(\sup_{s\in(-\infty,0]}|x_i(s)-x_i^*|^p\Big)\tag{4.6}$$

for $t>0$, where

$$\gamma=\max_{i\in\mathcal{I}}\Big\{\alpha_i+\sum_{j=1,j\neq i}^{m}\alpha_j|c_{ji}|^{\sigma^*_{ji}p}L_i^{\gamma^*_{ji}p}\int_0^\infty K_{ji}(s)\mathrm{e}^{\nu s}s\mathrm{d}s\Big\}\Big/\min_{i\in\mathcal{I}}\{\alpha_i\}\geqslant 1.$$

Noting that $x_i(s)=\psi_i(s)$ for $s\in(-\infty,0]$, it then follows from (4.6) that

$$\Big(\sum_{i=1}^{m}|x_i(t)-x_i^*|^p\Big)^{1/p}\leqslant\beta\mathrm{e}^{-\nu t}\left\|\boldsymbol{\psi}-\boldsymbol{x}^*\right\|_p\quad\text{for }\ t>0,$$

where $\beta=\sqrt[p]{\gamma}$. The global exponential stability of the unique equilibrium $\boldsymbol{x}^*$ of the network (2.1) has been established, and this completes the proof. □

One of the consequences of Theorem 4.2 is the following corollary:

Corollary 4.3. *Let $p>1$ be a real number, and let the assumptions* (2.2)–(2.4) *hold. Suppose $-(b_{ii}+c_{ii})\kappa_i>0$ and*

$$\begin{aligned}-(b_{ii}+c_{ii})\kappa_i>\frac{p-1}{p}&\sum_{j=1,j\neq i}^{m}\Big(|b_{ij}|^{\delta_{ij}\frac{p}{p-1}}+|c_{ij}|^{\sigma_{ij}\frac{p}{p-1}}\Big)L_j^{\gamma_{ij}\frac{p}{p-1}}\\&+\frac{1}{p}\sum_{j=1,j\neq i}^{m}\frac{\alpha_j}{\alpha_i}\Big(|b_{ji}|^{\delta^*_{ji}p}+|c_{ji}|^{\sigma^*_{ji}p}\Big)L_i^{\gamma^*_{ji}p}\end{aligned}\tag{4.7}$$

are satisfied, where the other additional parameters are defined in Theorem 3.1. *Then the equilibrium state $\boldsymbol{x}^*$ of the DCNN* (2.1) *is unique and globally exponentially stable.*

We remark that the network (1.4) denotes a special case of the network (2.1), and it is always assumed in (1.4) that the discrete delays τ_{ij} satisfy $\tau_{ij} \geqslant 0$ for $i \neq j$. The following Corollary 4.4 is an adaptation of Theorem 4.2 in which we establish the exponential convergence of the network (1.4) towards a unique equilibrium state $\boldsymbol{x}^*$. We remark that this corollary improves the results of Theorem 1 in Cao [4] on grounds that our condition (3.2) is more relaxed than the sufficient condition obtained in [4] in terms of restricting the parameter a_i, and the activation functions of the network (1.4) have been assumed to be unbounded and not monotonically increasing in general, while all the activation functions of the network studied in [4] have been assumed to be bounded and monotonically increasing.

Corollary 4.4. *Let $p > 1$ be a real number, and let the assumptions* (2.2) *and* (2.3) *hold. If the condition* (3.2) *is satisfied, then the unique equilibrium state $\boldsymbol{x}^*$ of the network* (1.4) *is globally exponentially stable.*

Of course, one may opt to adopt the results of Corollary 4.3 to the network model (1.4) as a variation to Corollary 4.4.

In the following, we give a couple of corollaries of Theorem 4.2 which further improve the global exponential stability results obtained earlier in the literature. The comparisons are made based on the assumptions that a delayed cellular neural network has self-connections which are inhibitory and instantaneous and the corresponding activation functions are monotonically increasing. These properties have been included intrinsically in the earlier studies of the network. The following Corollary 4.5 establishes the exponential convergence of the DCNN (2.1) (or DCNN (1.4)) in the norm $\|\cdot\|_2$. One can see that the condition (4.8) below is weaker than the sufficient conditions obtained in [7, 8, 10, 11, 31].

Corollary 4.5 (Case $p = 2$). *Let the assumptions* (2.2)–(2.4) *hold. Suppose the condition*

$$\begin{aligned} a_i - (b_{ii} + c_{ii})\kappa_i > {} & \frac{1}{2} \sum_{j=1, j\neq i}^{m} \Big(|b_{ij}|^{2\delta_{ij}} + |c_{ij}|^{2\sigma_{ij}} \Big) L_j^{2\gamma_{1,ij}} \\ & + \frac{1}{2} \sum_{j=1, j\neq i}^{m} \frac{\alpha_j}{\alpha_i} \Big(|b_{ji}|^{2\delta_{ji}^*} + |c_{ji}|^{2\sigma_{ji}^*} \Big) L_i^{2\gamma_{ji}^*} \end{aligned} \tag{4.8}$$

is satisfied, where the other additional parameters are defined as in Theorem 3.1. *Then the unique equilibrium state $\boldsymbol{x}^*$ of the DCNN* (2.1) *(or DCNN* (1.4)*) is globally exponentially stable.*

The following Corollary 4.6 establishes the exponential convergence of the network (2.1) (or DCNN (1.4)) in the norm $\|\cdot\|_3$. The results of this corollary can improve some of the results obtained by Cao [8].

Corollary 4.6 (Case $p = 3$). *Let the assumptions* (2.2)–(2.4) *hold. Suppose the condition*

$$\begin{aligned} a_i - (b_{ii} + c_{ii})\kappa_i > \frac{2}{3} \sum_{j=1,j\neq i}^{m} \Big(|b_{ij}|^{3\delta_{ij}/2} + |c_{ij}|^{3\sigma_{ij}/2}\Big) L_j^{3\gamma_{ij}/2} \\ + \frac{1}{3} \sum_{j=1,j\neq i}^{m} \frac{\alpha_j}{\alpha_i} \Big(|b_{ji}|^{3\delta_{ji}^*} + |c_{ji}|^{3\sigma_{ji}^*}\Big) L_i^{3\gamma_{ji}^*} \end{aligned} \tag{4.9}$$

is satisfied, where the other additional parameters are defined in Theorem 3.1. *Then the unique equilibrium state* $\boldsymbol{x}^*$ *of the DCNN* (2.1) *(or DCNN* (1.4)*) is globally exponentially stable.*

We remark that the general Euclidean norm $\|\cdot\|_p$ of a vector $\boldsymbol{u} \in \mathbb{R}^m$ tends to the norm $\|\cdot\|_\infty$ if we let $p \to \infty$. Henceforth, we assert in the following Corollary 4.7 the exponential convergence of the network (2.1) (or DCNN (1.4)) in the norm $\|\cdot\|_\infty$. One finds that the sufficient condition (4.10) given below is less restrictive when compared to the condition obtained earlier by Mohamad and Gopalsamy [37]. To get the condition (4.10) from (3.2), we let $\delta_{ij} = \sigma_{ij} = \gamma_{ij} = \frac{p-1}{p}$ and $\delta_{ij}^* = \sigma_{ij}^* = \gamma_{ij}^* = \frac{1}{p}$ for $i \neq j$ so that (3.2) reduces to

$$a_i - (b_{ii} + c_{ii})\kappa_i > \frac{p-1}{p} \sum_{j=1,j\neq i}^{m} (|b_{ij}| + |c_{ij}|)L_j + \frac{1}{p} \sum_{j=1,j\neq i}^{m} \frac{\alpha_j}{\alpha_i}(|b_{ji}| + |c_{ji}|)L_i.$$

By letting $p \to \infty$ in the above, we obtain the condition (4.10).

Corollary 4.7 (Case $p = \infty$). *Let the assumptions* (2.2)–(2.4) *hold. Suppose the condition*

$$a_i - (b_{ii} + c_{ii})\kappa_i > \sum_{j=1,j\neq i}^{m} (|b_{ij}| + |c_{ij}|)L_j \tag{4.10}$$

is satisfied. Then the unique equilibrium state $\boldsymbol{x}^*$ *of the DCNN* (2.1) *(or DCNN* (1.4)*) is globally exponentially stable.*

5. Conclusion

We have established the exponential convergence of the neural states of the network (2.1) toward a unique equilibrium state under a set of sufficient conditions which are delay independent and easily verifiable. The verifiable nature of our results gives us a better advantage over those results obtained by Liao et al [32] especially when the network has a large number of processing units. This can be found useful during the actual implementation of the network in which the choice of the parameter values in desiring the exponential convergence of the network can be tested easily against the sufficient condition (3.2).

The other advantage of our results which can contribute to the circuit implementation of the network is the usage of the monotonic property of the activation functions incorporated only within the instantaneous and inhibitory self-connectivity. We remark from most of the earlier studies of neural network models (particularly, [23, 49, 50]) that the use of non-monotonic activation functions can increase the memory storage capacity of a network when applied to associative memory related problems. We believe that this aspect may not be affected by our network (2.1) as the activation functions $f_i(\cdot)$ can have the non-monotonic property. In fact, the monotonic factor $\kappa_i > 0$ coming from the property $\kappa_i \leqslant \frac{g_i(u)-g_i(v)}{u-v}$ of the activation functions $g_i(\cdot)$ coupled with the parameters $b_{ii}, c_{ii} \leqslant 0$ can play a significant role in relaxing the restriction on the parameter value of a_i. We are aware that the parameter a_i is always related reciprocally with the resistance R_i within the circuitry of the cell i. Maintaining a high value of a_i (due to maintaining a low value of the resistance R_i) in responding to high values of $b_{ij}, c_{ij} \in \mathbb{R}$ for all $i, j \in \mathcal{I}$, as depicted from most of the conditions obtained in the earlier studies, can be an impossible task to achieve, particularly when a computational run of the network is attenuated over a long period of time, as this can cause the value of R_i to increase as a result of heating. Even worse is when the resistance R_i (during the circuit implementation) reaches a critical value where the network starts to destabilize. This concern can no longer exist if a network follows the proposed set-up similar to the circuitry of (2.1) in which the neural parameter values satisfy the sufficient condition (4.7) of the Corollary 4.3. We can see from the condition (4.7) that the choice of the parameter value of a_i becomes insignificant. This aspect added with the unbounded property of the activation functions lifts the network to a greater height in terms of real-world applications, especially for solving optimization problems which consist of unbounded constraints.

References

[1] S. Arik, *An improved global stability result for delayed cellular neural networks*, IEEE Trans. Circ. Syst.–I **49** (2002), 1211–1214.

[2] G. Avitabile, M. Forti, S. Manetti, M. Marini, *On a class of nonsymmetrical neural networks with application to ADC*, IEEE Trans. Circuits Syst. **38** (1991), 202–209.

[3] E. F. Beckenbach, R. Bellman, *Inequalities*, Springer Verlag, New York, 1965.

[4] J. Cao, *New results concerning exponential stability and periodic solutions of delayed cellular neural networks*, Phys. Lett. A **307** (2003), 136–147.

[5] J. Cao, *Global exponential stability and periodic solutions of delayed cellular neural networks*, J. Comp. Syst. Sci. **60** (2000), 38–46.

[6] J. Cao, *On exponential stability and periodic solutions of CNN's with delays*, Phys. Lett. A **267** (2000), 312–318.

[7] J. Cao, Q. Li, *On the exponential stability and periodic solutions of delayed cellular neural networks*, J. Math. Anal. Appl. **252** (2000), 50–64.

[8] J. Cao, *Global stability analysis in delayed cellular neural networks*, Phys. Rev. E **59** (1999), 5940–5944.

[9] J. Cao, *Periodic solutions and exponential stability in delayed cellular neural networks*, Phys. Rev. E **60** (1999), 3244–3248.

[10] J. Cao, *On stability of delayed cellular neural networks*, Phys. Lett. A **261** (1999), 303–308.

[11] J. Cao, D. Zhou, *Stability analysis of delayed cellular neural networks*, Neural Networks **11** (1998), 1601–1605.

[12] A. Chen, J. Cao, L. Huang, *Global robust stability of interval cellular neural networks with time-varying delays*, Chaos Solitons Fractals **23** (2005), 787–799.

[13] A. Chen, J. Cao, L. Huang, *Periodic solution and global exponential stability for shunting inhibitory delayed cellular neural networks*, Electr. J. Diff. Equ. **2004** (2004), 1–16.

[14] L. O. Chua, T. Roska, *Stability of a class of nonreciprocal cellular neural networks*, IEEE Trans. Circ. Syst. **37** (1990), 1520–1527.

[15] L. O. Chua, L. Yang, *Cellular neural networks: Theory*, IEEE Trans. Circ. Syst. **35** (1988), 1257–1272.

[16] L. O. Chua, L. Yang, *Cellular neural networks: Applications*, IEEE Trans Circ. Syst. **35** (1988), 1273–1290.

[17] P. P. Civalleri, M. Gilli, L. Pandolfi, *On stability of cellular neural networks with delay*, IEEE Trans. Circ. Syst. I: Fund. Theor. Appl. **40** (1993), 157–165.

[18] M. Forti, A. Tesi, *New conditions for global stability of neural networks with application to linear and quadratic programming problems*, IEEE Trans. Circuits Syst. I: Fund. Theor. Appl. **42** (1995), 354–366.

[19] M. Forti, *On global asymptotic stability of a class of nonlinear systems arising in neural network theory*, J. Diff. Equ. **113** (1994), 246–264.

[20] M. Forti, S. Manetti, M. Marini, *Necessary and sufficient condition for absolute stability of neural networks*, IEEE Trans. Circ. Syst. I: Fund. Theor. Appl. **41** (1994), 491–494.

[21] M. Forti, S. Manetti, M. Marini, *A condition for global convergence of a class of symmetric neural circuits*, IEEE Trans. Circ. Syst. I: Fund. Theor. Appl. **39** (1992), 480–483.

[22] C. J. Fu, *A sufficient condition for exponential stability of cellular neural networks with time-varying delays*, J. Math. (Wuhan) **22** (2002), 266–270.

[23] M. Gilli, *Stability of cellular neural networks and delayed cellular neural networks with nonpositive templates and nonmonotonic output functions*, IEEE Trans. Circ. Syst. I: Fund. Theor. Appl. **41** (1994), 518–528.

[24] K. Gopalsamy, *Stability of artificial neural networks with impulses*, Appl. Math. Computat. **154** (2004), 783–813.

[25] K. Gopalsamy, X. Z. He, *Stability in asymmetric Hopfield nets with transmission delays*, Phys. D **76** (1994), 344–358.

[26] K. Gopalsamy, *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, Kluwer Academic Press, The Netherlands, 1992.

[27] C. Guzelis, L. O. Chua, *Stability analysis of generalised cellular neural networks*, Int. J. Circuit Theor. Appl. **21** (1993), 1–33.

[28] T. Lara, P. Ecimovciz, J. Wu, *Delayed cellular neural networks: model, applications, implementations, and dynamics*, Diff. Equ. Dynam. Systems **10** (2002), 71–97.

[29] X. Li, L. Huang, *Exponential stability and global stability of cellular neural networks*, Appl. Math. Computat. **147** (2004), 843–853.

[30] X. M. Li, L. H. Huang, H. Zhu, *Global stability of cellular neural networks with constant and variable delays*, Nonlin. Anal. **53** (2003), 319–333.

[31] X. X. Liao, J. Wang, *Algebraic criteria for global exponential stability of cellular neural networks with multiple time delays*, IEEE Trans. Circuits Systems I Fund. Theory Appl. **50** (2003), 268–275.

[32] X. Liao, Z. Wu, J. Yu, *Stability analyses of cellular neural networks with continuous time delay*, J. Computat. Appl. Math. **143** (2002), 29–47.

[33] D. Liu, A. N. Michel, *Cellular neural networks for associative memories*, IEEE Trans. Circ. Syst. II: Analog Dig. Sig. Proc. **40** (1993), 119–121.

[34] Z. Liu, L. Liao, *Existence and global exponential stability of periodic solution of cellular neural networks with time-varying delays*, J. Math. Anal. Appl. **290** (2004), 247–262.

[35] S. Mohamad, *Convergence dynamics of delayed Hopfield-type neural networks under almost periodic stimuli*, Acta Appl. Math. **76** (2003), 117–135.

[36] S. Mohamad, *Global exponential stability in discrete-time analogues of delayed cellular neural networks*, J. Differ. Equ. Appl. **9** (2003), 559–575.

[37] S. Mohamad, K. Gopalsamy, *Exponential stability of continuous-time and discrete-time cellular neural networks with discrete delays*, Appl. Math. Computat. **135** (2003), 17–38.

[38] S. Mohamad, K. Gopalsamy, *Dynamics of a class of discrete-time neural networks and their continuous-time counterparts*, Math. Comput. Simul. **53** (2000), 1–39.

[39] T. Roska, L. Chua, D. Wolf, T. Kozek, R. Tetzlaff, F. Puffer, *Simulating nonlinear waves and PDEs via CNN – Part I: Basic Techniques, Part II: Typical examples*, IEEE Trans. Circuits Systems–I **42** (1995), 809–820.

[40] T. Roska, W. Chai, L. Chua, *Stability of CNN with dominant nonlinear and delay-type templates*, IEEE Trans. Circuits Systems–I **40** (1993), 270–272.

[41] T. Roska, L. O. Chua, *Cellular neural networks with nonlinear and delay-type template elements*, Int. J. Circuit Theory Appl. **20** (1992), 469–481.

[42] T. Sasagawa, *Sufficient condition for the exponential p-stability and p-stabilizability of linear stochastic systems*, Int. J. Syst. Sci. **13** (1982), 399–408.

[43] A. Slavova, *Cellular neural network models of some equations from Biology, Physics and Ecology*, Funct. Diff. Equ. **10** (2003), 579–591.

[44] A. Slavova, *Applications of some mathematical methods in the analysis of cellular neural networks*, J. Computat. Appl. Math. **114** (2000), 387–404.

[45] N. Takahashi, L. O. Chua, *On the complete stability of nonsymmetric cellular neural networks*, IEEE Trans. Circuits Syst. I: Fund. Theor. Appl. **45** (1998), 754–758.

[46] N. Takahashi, L. O. Chua, *A new sufficient condition for nonsymmetric CNN's to have a stable equilibrium point*, IEEE Trans. Circuits Syst. I: Fund. Theor. Appl. **44** (1997), 1092–1095.

[47] D. W. Tank, J. J. Hopfield, *Neural computation by concentrating information in time*, Proc. Natl. Acad. Sci. USA **84** (1987), 1896–1990.

[48] H. Xie, Q. Wang, *The existence of almost periodic solution for cellular neural networks with variable coefficients and delays*, Ann. Diff. Equ. **21** (2005), 65–72.

[49] H. Yanai, S. Amari, *Auto-associative memory with two-stage dynamics of non-monotonic neurons*, IEEE Trans. Neural Networks **7** (1996), 803–815.

[50] S. Yoshizawa, M. Morita, S. I. Amari, *Capacity of associative memory using a non-monotonic neuron model*, Neural Networks **6** (1993), 167–176.

[51] J. Zhang, *Absolute stability analysis in cellular neural networks with variable delays and unbounded delay*, Comp. Math. Appl. **47** (2004) 183–194.

[52] Q. Zhang, X. Wei, J. Xu, *New stability conditions for neural networks with constant and variable delays*, Chaos Solitons Fractals **26** (2005), 1391–1398.

[53] Q. Zhang, X. Wei, J. Xu, *On global exponential stability of nonautonomous delayed neural networks*, Chaos Solitons Fractals **26** (2005), 965–970.

[54] Q. Zhang, X. Wei, J. Xu, *Delay-dependent exponential stability of cellular neural networks with time-varying delays*, Chaos Solitons Fractals **23** (2005) 1363–1369.

[55] H. Zhao, *Existence and global attractivity of almost periodic solution for cellular neural network with distributed delays*, Appl. Math. Computat. **154** (2004), 683–695.

[56] D. Zhou, L. Zhang, J. Cao, *On global exponential stability of cellular neural networks with Lipschitz-continuous activation function and variable delays*, Appl. Math. Comput. **151** (2004), 379–392.

[57] D. Zhou, J. Cao, *Globally exponential stability conditions for cellular neural networks with time-varying delays*, Appl. Math. Computat. **131** (2002), 487–496.

Sannay Mohamad
Department of Mathematics,
Faculty of Science,
Universiti Brunei Darussalam,
Gadong BE1410,
Brunei Darussalam
e-mail: sannay@fos.ubd.edu.bn

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 269–281

The Single-Vendor Multi-Buyer Integrated Inventory Problem: an Heuristic Solution Technique

Mohammad Abdul Hoque and Yong Shiaw Yin

Abstract. Researchers have paid a lot of attention to the single-vendor single-buyer integrated inventory system, but not to the integrated single-vendor multi-buyer problem — especially the delivery of a single product to many buyers. This may be a particular case of the Joint Replenishment Problem (JRP), but the JRP does not account for the set-up and inventory costs of the vendor in delivering a product to many buyers. These vendor costs, and also relevant costs of the buyers, are considered in this paper. An integrated inventory model for delivering a single product to many buyers is developed, using either equal or unequal (or mixed) sized batches. An heuristic solution algorithm is developed, and illustrated with a numerical solution. A comparative study of two single-vendor – single-buyer numerical problems shows the effectiveness of this new technique.

Mathematics Subject Classification (2000). Primary 90B05; Secondary 90B35.

Keywords. Synchronisation; Integrated inventory; Optimal solution.

1. Introduction

The integration of vendor-buyer inventory systems plays an important role in modern supply chain environments. Establishing a close vendor-buyer relationship may be of mutual benefit [2, 3, 8, 9, 10]. Although researchers have given considerable attention to the single-vendor – single-buyer integrated inventory system, there has been little research on its extension to the single-vendor – multi-buyer case. The Joint Replenishment Problem (JRP) — viz. the problem of coordinating the replenishment of a group of items from a single supplier to many buyers — has

Results in this paper were presented at the International Conference on Mathematical Modelling and Computation held at the University of Brunei Darussalam during 5–8 June 2006, in conjunction with the 20th anniversary celebration of the foundation of the University.

been studied extensively, but most work has ignored the set-up and inventory costs of the manufacturer (references [5] and [7] are exceptions).

The widely accepted joint economic lot size approach assumes that system benefits generated by this approach can be shared among the buyer(s) and the seller in a costless way, but it has been argued that negotiated benefit sharing is never costless [5]. It requires information sharing, communication, trust building, travel and executive time — and hence an alternative approach for minimising the total inventory and ordering costs for the vendor and buyer(s) was proposed, and claimed to be individually responsible and rational, so no further negotiation between the seller and the buyer(s) for benefit sharing would be required [5]. This approach charged the buyer(s) the cost of shipping and handling associated with their respective order by appropriate reduction in the unit selling price, and it was shown that the system costs reduced as much as in the joint economic lot size approach as available in the literature at that time. However, by ignoring the suggested costs of benefit sharing by the joint economic lot size approach, it was shown that there is an initial error in the recognised unit selling price charged by the vendor [2]. A lower total joint relevant cost for a new example with a modified way of defining the joint economic lot size was demonstrated, and it was argued that the ability of the vendor to entice the buyer by paying for order handling and processing costs depends on their interrelationship. Subsequently, the joint replenishment of items from the viewpoint of integrated inventory was addressed, considering major and minor set-up costs and also a fixed cost [7]. Although multiple items were considered, it was assumed that the selling of each item was to only one particular buyer, which obviously may be too restrictive in practice. A one-vendor multi-buyer supply chain for a single product had been proposed earlier, to analyze the benefits of coordinating the supply chain through the use of common replenishment time periods [11]. It was assumed that the vendor does not keep any inventory, but orders the required quantity from an outside supplier whenever an order from a buyer is received. Then without considering integrated inventory, the vendor may decide common replenishment periods and offer a price discount to entice buyers.

The single-vendor multi-buyer case has also been considered by others. An analytical model was proposed to integrate and synchronise the procurement of a raw material needed to produce multiple items, and then to deliver them to multiple retailers [6]. The objective was to find the production sequence of items, the common production cycle length, and the delivery frequencies and quantities that minimised the average total cost. It has also been assumed that the demand at each retailer for an item might be known and satisfied by the item stored at the warehouse [1], to try to determine single-cycle policies which minimise the average cost. However, these previous single-vendor multi-buyer models have not dealt with integrated inventory, in producing a single product by a vendor and its delivery to many buyers.

This paper develops a model for supplying an item to more than one buyer, after its production by a manufacturer. We assume that there can be transfer in equal or unequal sized batches, and that each batch is transferred as soon as its processing is finished, so there is only a transportation cost incurred. For the first batches, a later batch is a multiple of the previous one in the ratio of the production and total demand rates, and the remainder are equal to the largest batch. We present an heuristic solution technique, and illustrate it with a numerical example. To demonstrate the effectiveness of our technique, we have also carried out a comparative study on the results of two numerical problems solved earlier [4].

In the next section, we list our assumptions and notation, and then present the model and an optimal solution technique. In the following sections, we then discuss solutions of a numerical problem using our solution technique, and then a comparative study involving two single-vendor single-buyer numerical problems. We draw our conclusions in the final section.

2. Model formulation

2.1. Assumptions and notation

In developing the models we assume

1. Deterministic constant demand and production rates;
2. No backlogging or deliberate planning for shortages;
3. Insignificant set-up and transportation times;
4. Both the vendor and the buyers agree to share the benefit of integrated inventory system through negotiation.

We adopt the following notation.

For the vendor:

$D =$ Annual rate of demand;

$P =$ Annual rate of production ($P > D$ and $k = P/D$);

$h =$ Inventory carrying cost per item per year;

$S =$ Production set-up cost per lot;

$z =$ The smallest batch (part of a lot) size;

$Q =$ The lot transferred from the vendor to the buyers;

$n =$ Number of equal and/or unequal sized batches in a lot; and

$e =$ Number of unequal sized batches in a lot.

For the i^{th} purchaser $(i = 1, \cdots, m)$:

$D_i =$ Annual rate of demand ($D = \sum_{i=1}^{m} D_i$);

$h_i =$ Inventory carrying cost per item per year;

$s_i =$ Cost of placing an order;

$T_i =$ Cost of transporting a batch from the vendor to buyer i; and

$g =$ Transport capacity of the transport equipment.

2.2. Batch transfer on production model

2.2.1. The total cost function. We assume that $z_i = D_i z/D$, $kz_i = D_i kz/D$, ..., $k^{e-1}z_i = D_i k^{e-1}z/D$
– i.e. $k^{j-1}z_i = D_i k^{j-1}z/D$ so that

$$k^{j-1}z = \sum_{i=1}^{m} k^{j-1}z_i \text{ for } j = 1, 2, ..., e. \tag{1}$$

Let the vendor transfer the first batch of size z to the buyers. Note that

$$z_i = z_i \Rightarrow D_i z/D = z_i \Longrightarrow (P/D)(D_i z/P) = z_i \Longrightarrow kz/P = z_i/D_i = z/D \tag{2}$$

because $z_i = D_i z/D$ – i.e. the production time of the batch kz equals the time of meeting the demand by the previous batch z. In the same way, one can show that $k^2z/P = kz_i/D_i$ and so on. Generally,

$$k^j z/P = k^{j-1}z_i/D_i = k^{j-1}z/D \text{ for } j = 1, 2, ..., e. \tag{3}$$

– i.e. the production time of the batch $k^j z$ equals the time of meeting the demand by the previous batch $k^{j-1}z$. To keep to a minimum inventory, the synchronization of the production flow is achieved by transferring the lot Q from the vendor to buyers in e unequal sized batches of sizes $z, kz, k^2z, ..., k^{e-1}z$ and $n-e$ equal sized batches of size $k^{e-1}z$, so that

$$z+kz+k^2z+...+k^{e-1}z+(n-e)k^{e-1}z = Q \Longrightarrow z = \frac{Q}{\sum_{r=0}^{e-1} k^r + (n-e)k^{e-1}}\ . \tag{4}$$

The batch z_i for all i, is transferred first. It meets the demand for the time z_i/D_i, and in this time the vendor produces the 2^{nd} batch kz and transfers the batch kz_i to the buyer i. This 2^{nd} batch meets the demand for the time kz_i/D_i and in this time the vendor produces the 3^{rd} batch k^2z and shifts the batch k^2z_i to the ith buyer. This way of transferring batches to buyers continues until it transfers the batch $k^{e-1}z_i$. Then the vendor transfers this batch (when its processing at the vendor finishes) repeatedly $(n-e)$ times. The inventory pattern of the repeated batches at the ith buyer per production cycle is shown in the following figure:

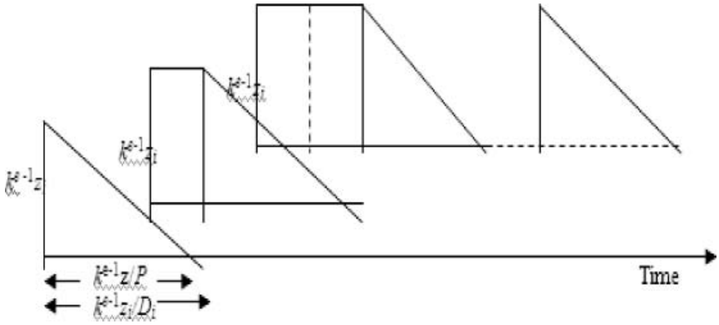


FIGURE 1. The inventory pattern for ith buyer per production cycle

Total cost = Cost of the vendor + Cost of the buyers. The vendor transfers each of batches as soon as it finishes its processing. Thus the WIP (Work-in-Progress) inventory for the manufacturer per cycle is

$$\frac{z^2}{2P}+\frac{k^2z^2}{2P}+\ldots+\frac{(k^{e-1}z)^2}{2P}+(n-e)\frac{(k^{e-1}z)^2}{2P}=\frac{z^2}{2P}\left[\sum_{r=0}^{e-1}k^{2r}+(n-e)k^{2(e-1)}\right]$$

The average inventory per cycle for the ith buyer can be evaluated as

$$\frac{z_i^2}{2D_i}+\frac{(kz_i)^2}{2D_i}\quad+\quad\ldots+\frac{(k^{e-1}z_i)^2}{2D_i}+(n-e)\frac{(k^{e-1}z_i)^2}{2D_i}$$
$$+\{1+2+\ldots+(n-e)\}\left(\frac{k^{e-1}z_i}{D_i}-\frac{k^{e-1}z}{P}\right)k^{e-1}z_i\ , \quad (5)$$

and substituting for $z_i = D_iz/D$ produces

$$\frac{D_iz^2}{2D}\left[\frac{\sum_{r=0}^{e-1}k^{2r}+(n-e)k^{2(e-1)}}{D}+(n-e)(n-e+1)(1/D-1/P)k^{2(e-1)}\right] \quad (6)$$

The set-up plus ordering plus transportation cost per cycle is given by

$$S+\sum_{i=1}^{m}(s_i+nT_i)\ , \quad (7)$$

so the total cost of the system per cycle is

$$\left[\left\{\sum_{r=0}^{e-1}k^{2r}+(n-e)k^{2(e-1)}\right\}\left\{\sum_{i=1}^{m}\frac{D_ih_i}{2D^2}+\frac{h}{2P}\right\}\right.$$
$$\left.+\ (n-e)(n-e+1)\left(\frac{1}{D}-\frac{1}{P}\right)k^{2(e-1)}\frac{\sum_{i=1}^{m}D_ih_i}{2D}\right]z^2$$
$$+S+\sum_{i=1}^{m}(s_i+nT_i)\ .$$

Since there are D/Q lots per year, we multiply this total cost by D/Q to obtain the average total cost per year – i.e. the total cost function

$$\left[\left\{\sum_{r=0}^{e-1}k^{2r}+(n-e)k^{2(e-1)}\right\}\left\{\sum_{i=1}^{m}\frac{D_ih_i}{2D}+\frac{Dh}{2P}\right\}\right.$$

$$\left.+(n-e)(n-e+1)\left(\frac{1}{D}-\frac{1}{P}\right)k^{2(e-1)}\frac{\sum_{i=1}^{m}D_ih_i}{2}\right]\frac{z^2}{Q}+\frac{D}{Q}\left[S+\sum_{i=1}^{m}(s_i+nT_i)\right]$$

or

$$C(n,e,Q)=\left[\begin{array}{c}\left\{\frac{\sum_{r=0}^{e-1}k^{2r}+(n-e)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1}k^{r}+(n-e)k^{e-1}\right\}^{2}}\right\}\left\{\frac{1}{2D}\sum_{i=1}^{m}D_ih_i+\frac{Dh}{2P}\right\}+\\ \frac{(n-e)(n-e+1)(1/D-1/P)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1}k^{r}+(n-e)k^{e-1}\right\}^{2}}\frac{\sum_{i=1}^{m}D_ih_i}{2}\end{array}\right]Q$$

$$+\frac{D}{Q}\left[S+\sum_{i=1}^{m}(s_i+nT_i)\right] \tag{8}$$

on substituting for z from (4).

2.2.2. The constraint. The largest batch $k^{e-1}z_i$ cannot exceed the capacity of the transport equipment, so $k^{e-1}z_i \le g$. Substituting for $z_i = D_iz/D$ we obtain

$$z \le \frac{Dg}{k^{e-1}D_i} \text{ or } k^{e-1}z \le \frac{Dg}{D_i} \text{ or } k^{e-1}z \le g', g' = Min(Dg/D_i)\ ; \tag{9}$$

and substituting for z from (4) and simplifying leads to

$$e-\sum_{r=0}^{e-1}k^{-r} \le n-\frac{Q}{g'}\ . \tag{10}$$

Thus we seek to minimise the total cost function (8) subject to constraint (10).

3. Solution technique

For given n and e, the cost function (8) is convex in Q, so the minimum value of Q and the associated minimum cost are respectively

$$Q_{\min}=\sqrt{E/F} \text{ and } C_{\min}=2\sqrt{EF} \tag{11}$$

where

$$E \;=\; \left\{\frac{\sum_{r=0}^{e-1}k^{2r}+(n-e)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1}k^{r}+(n-e)k^{e-1}\right\}^{2}}\right\}\left\{\frac{1}{2D}\sum_{i=1}^{m}D_ih_i+\frac{Dh}{2P}\right\} \tag{12}$$

$$+\;\frac{(n-e)(n-e+1)k^{2(e-1)}(1/D-1/P)}{\left\{\sum_{r=0}^{e-1}k^{r}+(n-e)k^{e-1}\right\}^{2}}\frac{\sum_{i=1}^{m}D_ih_i}{2} \tag{13}$$

and

$$F=D\left\{S+\sum_{i=1}^{m}(s_i+T_i)\right\}\ . \tag{14}$$

Constraint (10) is always satisfied for $n = e = 1$, hence an initial feasible solution can be found as $Q = g'$ for these values of n and e, together with the initial cost from (8) as the absolute minimal total cost. Then for these known values of n and e, we can calculate the minimal Q using equation (11). If $n0_i$ is the rounded up value of Q/g', then for $n = n0$ the inequality (10) is always satisfied for $e = 1$, so $(n0,1)$ is always an initial feasible solution. The integral values of e, from 1 up to the highest integer ($\le n0$) that satisfy the inequality

(10) along with $n = n0$, are the feasible solutions. Thus a set of feasible values of (n, e) can be determined satisfying (10). If (n, e) satisfies the constraint (10), we observe that $(n + x, e + x)$ for any integral x also satisfies (10). For a given Q and a feasible (n, e), it is proven in Appendix A that the coefficient of Q in (8) is a monotonically decreasing function of x, for a given set of feasible solutions $(n + x, e + x)$. Thus for any set $\{Q, n, e\}$, a part of the coefficient of Q in (8) is a monotonically decreasing function of x, and the remaining part is a linear function of x. The total cost function is therefore a convex function in x, with a minimum at the left-hand end of the range — i.e., where

$$C(n + x - 1, e + x - 1) > C(n + x, e + x) < C(n + x + 1, e + x + 1) \ , \tag{15}$$

implying

$$g(x-1) - \frac{D}{Q}\sum_{i=1}^{m} T_i > g(x) \ \text{ and } \ g(x) < g(x+1) + \frac{D}{Q}\sum_{i=1}^{m} T_i \tag{16}$$

with

$$g(x) = \left[\begin{array}{c} \left\{ \frac{\sum_{r=0}^{e+x-1} k^{2r} + (n-e)k^{2(e+x-1)}}{\left\{\sum_{r=0}^{e+x-1} k^{r} + (n-e)k^{e+x-1}\right\}^2} \right\} \left\{ \frac{1}{2D} \sum_{i=1}^{m} D_i h_i + \frac{Dh}{2P} \right\} + \\ \frac{(n-e)(n-e+1)(1/D-1/P)k^{2(e+x-1)}}{\left\{\sum_{r=o}^{e+x-1} k^{r} + (n-e)k^{e+x-1}\right\}^2} \frac{\sum_{i=1}^{m} D_i h_i}{2} \end{array} \right] \times Q \ , \tag{17}$$

leading to

$$g(x) - g(x+1) < \frac{D}{Q}\sum_{i=1}^{m} T_i < g(x-1) - g(x) \tag{18}$$

For each feasible solution, we obtain its minimal total cost along with values of n and e, and then the absolute minimum of the feasible solutions. Using the values of n, e for this absolute minimum, we calculate the value of Q and the associated cost from (11); and continue to compute the value of Q and associated absolute minimum total cost, until the absolute minimal total cost is equal or greater than its previous value. The absolute minimal total cost so obtained – along with the associated values of Q, n and e – is the desired minimal cost solution.

4. Numerical illustration

We illustrate our solution algorithm with a numerical example, where an item is supplied to 5 buyers. The artificially generated data for this are given in Table 1.

TABLE 1. Data for a single-vendor 5-buyer case

Purchaser	s_i	D_i	h_i	T_i
1	25	200	0.22	25
2	15	150	0.24	20
3	25	225	0.25	18
4	30	230	0.23	25
5	30	165	0.21	15

Relevant data are $S = 300$, $h = 0.20$, $P = 1500$, $D = \sum_{i=1}^{5} D_i = 970$, and $g = 300$.

Solution:

Step 1 Here $Q = g' = Min_i(Dg/D_i) = 843.48$ and from (8) 759.05 is the absolute minimum total cost. For $n0 = e0 = 1$, $Q_{\min} = 1686.68$ and $C_{\min} = 607.30$ from (11) – i.e. less than the previous absolute minimal total cost, and hence we adopt $C_{\min} = 607.30$ as the absolute minimal total cost.

Step 2 For $Q = 1686.68$, the feasible solution is $n0 = 2$ and $e0 = 1$ and $C(1686.68, 2, 1) = 549.08$. We increase the values of n and e by 1 at each step and calculate the respective total cost. Thus $C(1686.68, 3, 2) = 546.57$ and $C(1686.68, 4, 3) = 579.88$ – which is higher than previously, so it remains the absolute minimum at $n = 3$, $e = 2$. For these values of n and e, we then calculate $Q_{\min} = 3106.32$ and $C_{\min} = 458.41$ from (11). This value of $C_{\min}$ is smaller than the previous one, so we set the absolute minimal cost equal to the present $C_{\min} = 458.41$ and go to the next step.

Step 3 For $Q = 3106.32$, the feasible solution is $n0 = 4$ and $e0 = 1$ and $C(3106.32, 4, 1) = 495.97$. We again increase the values of n and e by 1 and calculate the respective total cost, obtaining $C(3106.32, 5, 2) = 478.14$ and $C(3106.32, 6, 3) = 485.08$ – higher than the previous value, which is then retained as the absolute minimum at $n = 5$, $e = 2$. For these values of n and e, we calculate $Q_{\min} = 3916.94$ and $C_{\min} = 465.57$ from (11), and this value of $C_{\min}$ is higher than the previous absolute minimal cost so we stop.

We conclude that the minimal solution is $Q = 3106.32$, $n = 3$, $e = 2$ and the total cost is 458.41. The lot is transferred from the vendor with batches of sizes $757.64, 1174.34, 1174.34$. Batch sizes for the buyers are shown in Table 2.

TABLE 2. Optimal batch sizes for the buyers

1^{st} buyer	$156.21, 242.13, 242.13$
2^{nd} buyer	$117.16, 181.6, 181.6$
3^{rd} buyer	$175.74, 272.40, 272.40$
4^{th} buyer	$179.63, 278.52, 278.52$
5^{th} buyer	$128.87, 199.76, 199.76$

Our solution technique can also be applied to solve single-vendor single-buyer problems, as we now illustrate by comparing the results obtained recently in such a case [4] with the outcome from our technique. Relevant data are $S = 400$, $h_1 = 4$, $P = 3200$, $D = D_1 = 1000$, $s_1 = 0$, $T_1 = 25$, $g = 300$ and this numerical problem is solved for $h = 5, 7$. The results for comparison appear in Table 3.

Thus our solution technique for the single-vendor multi-buyer model closely reproduces the results obtained recently for a single-vendor single-buyer problem by another procedure [4].

TABLE 3. Comparison of results obtained in a single-vendor single-buyer case

Reference [4]

h	n	e	Q	Cost	Batch sizes
5	3	3	553.84	1715.30	38.35, 122.73, 392.75
7	4	3	553.84	1786.44	22.60, 72.33, 231.44, 233.40

Our technique

n	e	Q	Cost	Batch sizes
3	3	553.84	1715.30	38.35, 122.73, 392.75
4	3	559.77	1786.44	22.68, 72.58, 232.27, 232.27

5. Conclusion

There has been considerable research on the single-vendor single-buyer problem, but the single-vendor multi-buyer problem has received relatively little attention in the literature. This paper has developed a model for the single-vendor multi-buyer problem, where integrated inventory is considered. The production flow is synchronised by transferring the lot from the vendor to the buyers in equal or unequal sized batches. Properties that lead to a step by step solution have been established, and our technique is illustrated for a particular numerical example. Our solution technique was also applied to a single-vendor single-buyer case recently solved in the literature. Our more general technique leads to the same result as obtained by a previous method in the literature. Thus our new single-vendor multi-buyer model and heuristic solution technique has enriched the vendor-buyer integrated inventory literature.

Appendix A

Consider

$$\frac{(n-e)(n-e+1)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1} k^r + (n-e)k^{e-1}\right\}^2} = \frac{(n-e)(n-e+1)}{\left\{\sum_{r=0}^{e-1} k^{-r} + (n-e)\right\}^2}. \tag{19}$$

Let

$$\begin{aligned} f(n+x, e+x) &= \frac{(n-e)(n-e+1)}{\left\{\sum_{r=0}^{e+x-1} k^{-r} + (n-e)\right\}^2} \leq \frac{(n-e)(n-e+1)}{\left\{\sum_{r=0}^{e+x} k^{-r} + (n-e)\right\}^2} \\ &= f(n+x+1, e+x+1)\ . \end{aligned}$$

Increasing each of the values of $n+x$ and $e+x$ by 1 yields

$$\frac{1}{\left\{\sum_{r=0}^{e+x-1} k^{-r} + (n-e)\right\}^2} \leq \frac{1}{\left\{\sum_{r=0}^{e+x} k^{-r} + (n-e)\right\}^2} \tag{20}$$

implying $1/(k^{e+x}) \le 0$, a contradiction since $k > 1$. Thus

$$\frac{(n-e)(n-e+1)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1} k^r + (n-e)k^{e-1}\right\}^2}$$

is a monotonically decreasing function of x. Now let us consider the remaining term

$$\frac{\sum_{r=0}^{e-1} k^{2r} + (n-e)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1} k^r + (n-e)k^{e-1}\right\}^2}. \tag{21}$$

The numerator can be transformed to

$$\begin{aligned}\frac{k^{2e}-1}{k^2-1} + (n-e)k^{2(e-1)} = \frac{k^{2e}-1}{k^2-1} - \frac{k^e-1}{k-1}\\ +(n-e)k^{2(e-1)} - (n-e)k^{e-1} + \left[\frac{k^e-1}{k-1} + (n-e)k^{e-1}\right]\\ = [(k^e+1)-(k+1)]\,\frac{k^e-1}{k^2-1} + (n-e)k^{e-1}(k^{e-1}-1) + f(n,e), \qquad (22)\end{aligned}$$

where

$$\begin{aligned} f(n,e) &= \frac{k^e-1}{k-1} + (n-e)k^{e-1}\\ &= \frac{k^e-k}{k}\,\frac{k(k^e-1)}{k^2-1} + (n-e)k^{e-1}(k^{e-1}-1) + f(n,e)\\ &= (k^{e-1}-1)\left(1-\frac{1}{k+1}\right)\frac{k^e-1}{k-1} + (n-e)k^{e-1}(k^{e-1}-1) + f(n,e)\\ &= (k^{e-1}-1)\left(\frac{k^e-1}{k-1} + (n-e)k^{e-1} - \frac{k^e-1}{k^2-1}\right) + f(n,e)\\ &= (k^{e-1}-1)\left(f(n,e) - \frac{k^e-1}{k^2-1}\right) + f(n,e)\\ &= k^{e-1}f(n,e) - \frac{(k^e-1)(k^{e-1}-1)}{k^2-1}\,, \end{aligned}$$

whence $$\frac{\sum_{r=0}^{e-1} k^{2r} + (n-e)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1} k^r + (n-e)k^{e-1}\right\}^2} = \frac{k^{e-1}}{f(n,e)} - \frac{(k^e-1)(k^{e-1}-1)}{(k^2-1)\{f(n,e)\}^2}. \tag{23}$$

Now let

$$h(n+x, e+x)$$

$$
\begin{aligned}
&= \frac{k^{e+x-1}}{(n-e)k^{e+x-1}+\sum_{r=0}^{e+x-1}k^r} \\
&- \frac{(k^{e+x}-1)(k^{e+x-1}-1)}{(k^2-1)\left\{(n-e)k^{e+x-1}+\sum_{r=0}^{e+x-1}k^r\right\}^2} \\
&\le \frac{k^{e+x}}{(n-e)k^{e+x}+\sum_{r=0}^{e+x}k^r} - \frac{(k^{e+x+1}-1)(k^{e+x}-1)}{(k^2-1)\left\{(n-e)k^{e+x}+\sum_{r=0}^{e+x}k^r\right\}^2} \\
&= h(n+x+1, e+x+1)
\end{aligned}
$$

so that

$$
\begin{aligned}
&\frac{k^{e+x-1}}{(n-e)k^{e+x-1}+\sum_{r=0}^{e+x-1}k^r} - \frac{k^{e+x}}{(n-e)k^{e+x}+\sum_{r=0}^{e+x}k^r} \qquad (24)\\
&\le \frac{1}{(k^2-1)}\left[\frac{(k^{e+x}-1)(k^{e+x-1}-1)}{\left\{(n-e)k^{e+x-1}+\sum_{r=0}^{e+x-1}k^r\right\}^2} - \frac{(k^{e+x+1}-1)(k^{e+x}-1)}{\left\{(n-e)k^{e+x}+\sum_{r=0}^{e+x}k^r\right\}^2}\right].
\end{aligned}
$$

Assuming the left-hand side of this inequality is non-positive, it can easily be shown that

$$
\sum_{r=0}^{e+x}k^r \le k\sum_{r=0}^{e+x-1}k^r = -1+\sum_{r=0}^{e+x}k^r \Longrightarrow 0 \le -1 -- \text{ a contradiction }, \qquad (25)
$$

so the left-hand side of this inequality is positive. Now if the right-hand side of the inequality (24) is positive, then

$$
\frac{k^{e+x-1}-1}{\left\{(n-e)k^{e+x-1}+\sum_{r=0}^{e+x-1}k^r\right\}^2} > \frac{k^{e+x+1}-1}{\left\{(n-e)k^{e+x}+\sum_{r=0}^{e+x}k^r\right\}^2}, \qquad (26)
$$

which after simplification becomes

$$
\begin{aligned}
&-(n-e)^2k^{2(e+x-1)}(k^2-1)+2(n-e)k^{2(e+x)-1}\left\{(1-k)\sum_{r=0}^{e+x-1}k^r+k^{e+x}-k\right\} \\
&+ \quad (k^{e+x-1}-1)(\sum_{r=0}^{e+x}k^r)^2-(k^{e+x+1}-1)(\sum_{r=0}^{e+x-1}k^r)^2 > 0\,. \qquad (27)
\end{aligned}
$$

Assuming $(1-k)\sum_{r=0}^{e+x-1}k^r+k^{e+x}-k$ is non-negative, it can be shown that $k \le 1$ (a contradiction), so this form is negative. Again let

$$
\left(k^{e+x-1}-1\right)\left(\sum_{r=0}^{e+x}k^r\right)^2-\left(k^{e+x+1}-1\right)\left(\sum_{r=0}^{e+x-1}k^r\right)^2 \ge 0\,, \qquad (28)
$$

which simplifies to

$$
\left(k^{e+x-1}-1\right)\left\{(k-1)\,k^{e+x+1}+2k\left(k^{e+x}-1\right)\right\} \ge (k+1)\left(k^{e+x}-1\right)^2. \qquad (29)
$$

Simplifying further this reduces to $k^{e+x+1} \leq 1$ (a contradiction), whence

$$\left(k^{e+x-1}-1\right)\left(\sum_{r=0}^{e+x} k^r\right)^2 - \left(k^{e+x+1}-1\right)\left(\sum_{r=0}^{e+x-1} k^r\right)^2 < 0\,, \tag{30}$$

which from (27) is a negative number greater than 0 (a contradiction), so that

$$\frac{k^{e+x-1}-1}{\left\{(n-e)k^{e+x-1}+\sum_{r=0}^{e+x-1} k^r\right\}^2} - \frac{k^{e+x}-1}{\left\{(n-e)k^{e+x}+\sum_{r=0}^{e+x} k^r\right\}^2} \text{ is negative.} \tag{31}$$

Now the right-hand side of the inequality (24) is negative, whereas its left-hand side is positive (another contradiction), whence

$$h(n+x,e+x) > h(n+x+1,e+x+1) \text{ such that } \frac{\sum_{r=0}^{e-1} k^{2r}+(n-e)k^{2(e-1)}}{\left\{\sum_{r=0}^{e-1} k^r+(n-e)k^{e-1}\right\}^2}$$

is also a monotonically decreasing function of x – and hence the coefficient of Q in (8) is a monotonically decreasing function of x.

Acknowledgment

The authors are grateful to the referees for valuable comments and suggestions, and to Dr Abby Tan for assisting with the preparation of this paper.

References

[1] Abdul-Jalbar, B., Gutierrez, J. M. and Sicilia, J., (2006). Single cycle policies for the one-warehouse N-retailer inventory/distribution system, Omega 34(2) 196-208.

[2] Goyal, S. K., Srinivasan, G., (1992). The individually responsible and rational decision approach to economic lot sizes for one vendor and many purchasers: a comment, Decision Sciences 23 777-784.

[3] Hill, R. M., (1999). The optimal production and shipment policy for the single-vendor single-buyer integrated production-inventory problem, International Journal of Production Research 37 2463-2475.

[4] Hill, R. M., (2006). Another look at the single-vendor single-buyer integrated production-inventory problem, International Journal of production Research 44(4), 791-800.

[5] Joglekar, P. N. and Tharthare, S., (1990). The individually responsible and rational decision approach to economic lot sizes for one vendor and many purchasers, Decision Sciences 21 492-500.

[6] Kim, T., Hong, Y. and Chang, S. Y., (2006). Joint economic procurement - production - delivery policy for multiple items in a single-manufacturer, multiple-retailer system, International Journal of Production Economics 103(1) 199-208.

[7] Lu, L., (1995). A one-vendor multi-buyer integrated inventory model, European Journal of Operational Research, 81 312-323.

[8] Martinich, J. C., (1997). Production and Operations Management. Wiley, New York.

[9] Pan, J. C. and Yang, J., (2002). A study of an integrated inventory with controllable lead time, International Journal of Production Research 40(5) 1263-1273.

[10] Thomas, D. J. and Griffin, P. M., (1996). Coordinated supply chain management, European Journal of Operational Research 94 1-15.

[11] Viswanathan, S. and Piplani, R., (2001). Coordinating supply chain inventories through common replenishment epochs, European Journal of Operational Research 129 277-286.

Mohammad Abdul Hoque
Department of Mathematics
Universiti Brunei Darussalam
Gadong BE1410
Brunei Darussalam
e-mail: hoque@fos.ubd.edu.bn

Yong Shiaw Yin
Department of Mathematics
Universiti Brunei Darussalam
Gadong BE1410
Brunei Darussalam
e-mail: yin@fos.ubd.edu.bn

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 283–291

A Term Structured Volatility Model of Poll Data and its Application to Election Timing

Dharma Lesmono

Abstract. A term structured volatility model is proposed to describe the dynamic of poll data measuring the difference in popularity between the Government and the Opposition in Australia. This model is then used to determine the best time for the Government to call for an election in the Majoritarian Parliamentary System. The results are in terms of the expected remaining life in government and the exercise boundary, given certain values of popularity.

Mathematics Subject Classification (2000). Primary 60H10; Secondary 90C39.

Keywords. Stochastic differential equation, Dynamic programming, Election timing.

1. Introduction

Stochastic Differential Equations (SDEs) have often been used to model various phenomena in daily life (e.g., see [4, 8, 10–12]). Applications can be found in finance, insurance, biology, physics, medicine and many other areas. In finance and insurance, the applications include pricing of options, portfolio optimisation, the optimal time to sell an asset, and the calculation of insurance risks. In biology and medicine, areas such as population dynamics, stochastic epidemic models and neuroscience require a strong background in stochastic processes and SDEs. In physics, applications include the development of the Langevin equation, kinetic models and quantum mechanics.

The application of SDEs in this paper is to politics, and in particular the mathematical modelling of poll data. The term "two-party-preferred" is used to refer to the distribution of preferences (votes) between the two major political groups in the Australian Parliament in Canberra — viz. the Coalition (Liberal

Results in this paper were presented at the International Conference on Mathematical Modelling and Computation held at the University of Brunei Darussalam during 5–8 June 2006, in conjunction with the 20th anniversary celebration of the foundation of the university.

Party and National Party) and the Labor Party. Figure 1 (above) represents the difference in the inferred two-party-preferred vote over a period with the Coalition in Government and the Labor party in Opposition, where the poll data were normally taken fortnightly but approximately weekly after an election was announced, and even more frequently in days leading up to the election date. The figure has some similarity with the dynamic of stock prices in finance, and the poll process may be modelled using the following SDE:

$$dS(t) = -\mu \frac{S(t)}{1 - S(t)^2} dt + \sigma(t) dW(t) \tag{1.1}$$

where $W(t)$ is a Wiener process, $S(t)$ is the difference in the two-party-preferred $(-1 < S < 1)$, and μ and σ are positive constants. The drift of the above SDE has a mean-reverting coefficient, which always reverts to zero as the less popular party reacts to make its popularity higher in the next poll. An assumption of constant volatility is analogous to that made in the original work of Black and Scholes, in their now famous model of option pricing in finance [2]. However, the poll data actually possess a weak time dependence, with clustering similar to that leading to stochastic volatility models of stock price data, so a term structured volatility model is preferable.

The organisation of the remainder of this paper is as follows. In Section 2, volatility estimates for the poll data are performed and a term structured volatility model is introduced. In Section 3, this term structured volatility model is applied to the election timing problem, along with a discrete time model (cf. [9]). Results in terms of the expected remaining life and exercise boundary are given in Section 4, and concluding remarks are made in Section 5.

2. Term structured volatility model

Volatility estimates for the two-party-preferred data from April 1993 — December 2002 are performed using the EWMA (Exponentially Weighted Moving Averages) method, which is basically an exponential smoothing procedure for analysing time series data. This method gives more weight to recent and less weight to earlier observations, in order to detect small changes in the volatility. The EWMA can also react to a jump in the data faster than the simple moving average method. It has been used in the *RiskMetrics* program introduced by the American Bank JP Morgan of October 1994, to obtain estimates of volatility and correlation in the framework of Value-at-Risk (VaR) (cf. [6] or Chapter 57 of [14]). In our case, a dynamic volatility estimate follows from

$$\widehat{\sigma}_{i+1}^2 = \lambda \widehat{\sigma}_i^2 + (1-\lambda) \frac{dX_i^2}{dt_i} \qquad i = 1, 2, ..., N$$

with $\lambda = 0.94$ as the weighting factor (as in *RiskMetrics*) and $dX_i = dS_i + \mu \left(S_i / (1 - S_i^2) \right) dt_i$. The result of this EWMA method is given in Figure 1 (below). Note that the EWMA volatility estimates capture the jump in the data

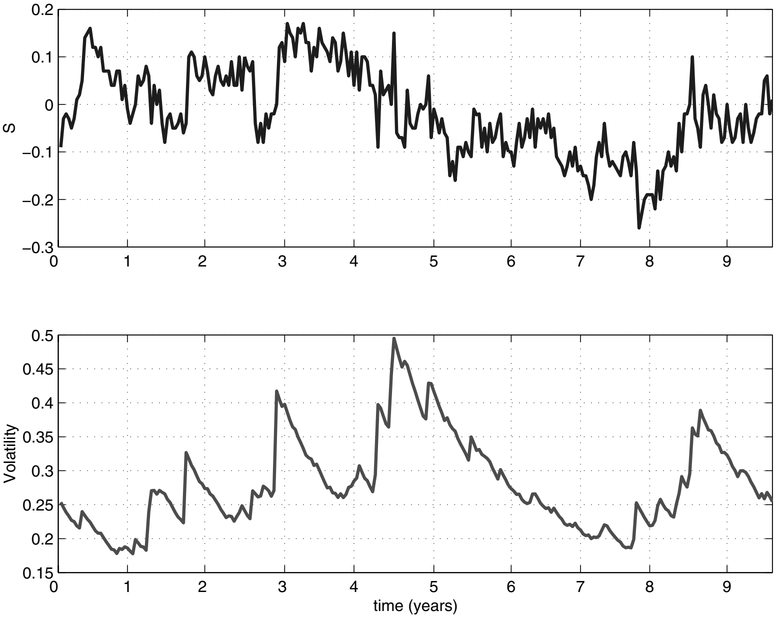


FIGURE 1. Volatility Estimates for Two-Party-Preferred Data.

between the fourth and fifth year; and it appears that the volatility estimates for the two-party-preferred data change over time, which may be due to volatility near the election date.

The term structured volatility model accommodates the dynamic volatility in the data similar to that for commodity markets [5] — viz.

$$dS(t) = -\mu \frac{S(t)}{1 - S(t)^2} dt + \sigma(t) dW(t), \quad \sigma(t) = \sigma_0 + \sigma_1 e^{q(T-t)}. \tag{2.1}$$

In Figure 2, a dynamic volatility estimate is performed for each period between four elections. This term structured volatility model is introduced in each of these four periods, to capture the dynamic of the EWMA estimates. Parameters in the term structured volatility model are estimated using the least-squares method, while μ is estimated using Maximum Likelihood Estimation (MLE). The parameter estimates of the term structured volatility model are summarised in Table 1, and results for the EWMA and a term structured volatility model for each period between elections are shown in Figure 2. The results seem promising for the period 1998–2001, where the proposed model matches the EWMA volatility estimates quite well, with a coefficient of determination R^2 of 96.96%. The rising volatility close to the election day is also as anticipated in this period. However, other factors such as volatility clustering in certain time intervals and jumps in the two-party-preferred data contribute in the other periods. In Figure 3, the best fit of the

TABLE 1. Parameter Estimates of Term Structured Volatility Model.

Parameters	1993–1996	1996–1998	1998–2001	2001–2004	Best fit
σ_0	0.2568	1.2851	0.2096	0.2420	0.2508
σ_1	–0.0191	–0.8720	0.2105	–0.0024	0.1713
q	0.3968	0.0734	–5.5610	0.9524	–17.4800

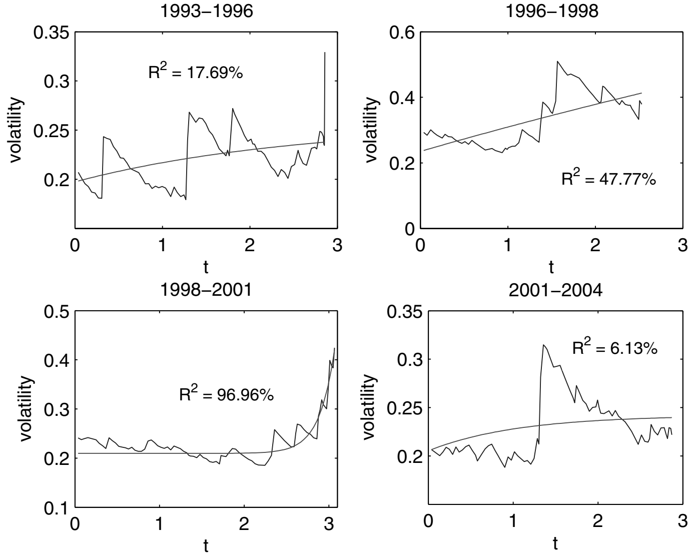


FIGURE 2. EWMA and Term Structured Volatility Model.

term structured volatility estimates from the model is given for each period, with coefficient of determination R^2 of 51%. These parameters are used to derive the model in the next section.

3. Election timing

In Australia and other countries with the Majoritarian Parliamentary System (e.g., Canada, New Zealand and Britain), governments have the constitutional right to call an early election within their term in office, which is typically between 3 to 5 years. The Australian Constitution and the Commonwealth Electoral Act of 1918 give the Australian Government the right to call an early election, subject to the approval of the representative of the Head of State (the Governor General). This is in contrast to presidential elections in the USA for example, where there is a fixed period of four years between elections.

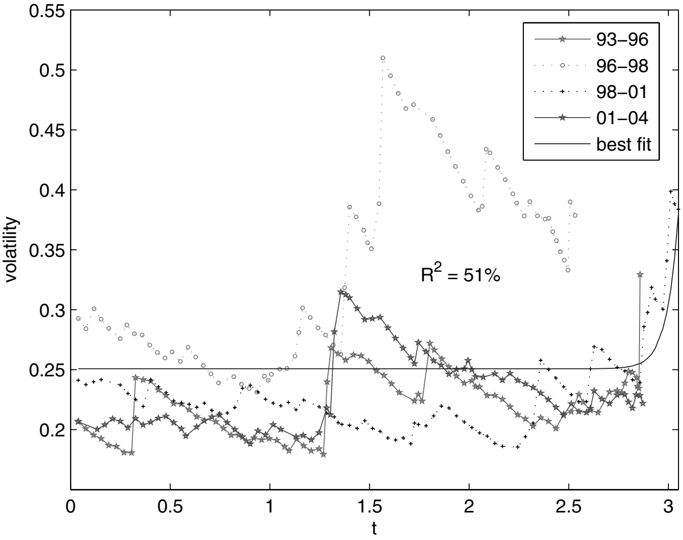


FIGURE 3. EWMA and Term Structured Volatility Model (Best Fit).

By announcing an election at the best time, an Australian Government can maximise its expected remaining life in power. There are many factors considered by the Government before it decides to call an early election — e.g., current economic growth, inflation rate, unemployment level or political issues. In this paper, Morgan Poll two-party-preferred data (www.roymorgan.com) are used to measure the popularity of the Government and the Opposition. An optimal control is devised for the Government, by locating an exercise boundary which indicates whether or not a snap election should be called. This problem can be compared with determination of the early exercise of an American option in finance.

In deriving the model, equation (2.1) and a discrete time model in [9] are used. The notation used in the formulation of the problem is as follows (details can also be found in [9]):

- $P_{iz}(t)$ is the transition probability from poll state S_i to state S_z over period t, measuring diffusion in the polls over period t. Stationarity of the process is assumed, so the transition probabilities remain unchanged even during an election campaign.
- Q_{zj} is the conditional probability that the true state of voting intentions is S_z, given that the poll state is S_j. This is a correction term to account for sampling errors.
- $P(W|S_j)$ is the conditional probability of winning the election from the true popularity state S_j, which contains the exaggerated majority effect (i.e., a party can win more than 50% of the votes and yet still lose the election). This quantity is derived from the resultant proportion of seats won by the

Coalition and the true state S, from the 22 Australian Federal Elections held since 1949.

- $\mathbf{E}(L|S_i,t)$ is the conditional expected remaining life of the Government, given poll state S_i and time $t \in [0,Y]$ into the current term.

One may adopt $P_{T_L} = P^k$ as $T_L = k\delta t$ and assume that the evolution process for S remains the same throughout the election process. (T_L here is a lead time, the period between announcing and holding the election, which can also be viewed as an election campaign period.)

The objective of the party in power is to maximise the expected remaining life in government by deciding whether or not to call an early election, with the exception that at terminal time t=Y (Y = 3 years in Australia) an election is compulsory. In each state at every time step, the Government must decide whether or not to call an early election by considering the expected remaining life between each alternative. The single control afforded to the Government in this problem is the action of stopping (calling an election), and the problem centres around this optimal stopping problem. Numerical dynamic programming is implemented to solve the recursive formulation for the expected remaining life in government. If an early election is called, time T_L will elapse with certainty after the announcement. If the Government wins the election, the expected remaining life is extended and the same problem is again considered, but with t reset to zero. If the Government chooses not to call an early election, it remains in power up to the next time step δt with certainty. At the new time t+δt the poll state will diffuse to a new value, and again the decision whether or not to call an early election can be re-evaluated. At the final time t=Y an election must be held, so the latest time to call an election is at t=Y-T_L.

The expected remaining life when calling an early election and calling no election are respectively

$$\mathbf{E}(L|S_i,t) = T_L + \sum_{j=1}^{m} \mathbf{E}(L|S_j,0) P_{iz}(T_L) Q_{zj} P(W|S_j) \qquad (3.1)$$

and

$$\mathbf{E}(L|S_i,t) = \delta t + \sum_{j=1}^{m} \mathbf{E}(L|S_j,t+\delta t) P_{ij}(\delta t), \qquad (3.2)$$

where the Einstein convention is used for summation over the repeated index z. Thus the expected remaining life is the maximum between calling an early election and calling no election — i.e.,

$$\mathbf{E}(L|S_i,t) = \max\left\{T_L + \sum_{j=1}^{m}\mathbf{E}(L|S_j,0)P_{iz}(T_L)Q_{zj}P(W|S_j),\right.$$
$$\left.\delta t + \sum_{j=1}^{m}\mathbf{E}(L|S_j,t+\delta t)P_{ij}(\delta t)\right\}. \qquad (3.3)$$

The algorithm for solving this problem starts with an initial estimate for the expected remaining life at $t = 0$ and calculates the expected remaining life at the final time. Then the algorithm moves backward until a new value at time $t = 0$ is obtained, and updates the initial estimate. This procedure is repeated until there is convergence — i.e., when the norm between the expected remaining life at $t = 0$ in two consecutive iterations is less than a chosen tolerance value ϵ.

4. Numerical results

For the computation, the popularity S (where usually $-0.5 < S < 0.5$) was divided into $m = 50$ states, with n equal time intervals during the Y years. A lead time of $T_L = 0.12$ year (around six weeks) and a time step $\delta t = 0.04$ year (around two weeks) were used. The number of time steps $n = Y/\delta t$, with Y either three or four (the maximum term in years).

Numerical results in terms of the expected remaining life and the exercise boundary are given in Figure 4(a) and (b). There are two surfaces in Figure 4(a), respectively representing the expected remaining life for a maximum term of three and four years. In those surfaces, the expected remaining life in general is almost constant at the beginning of the term regardless of the level of popularity, and then decreases as time elapses. The four-year maximum term gives a longer expected remaining life than the three-year maximum term (around 10 years and 7 years respectively), as the Government has more time to wait for its popularity to increase before calling an election. It is interesting to note that, in a referendum in 1988 to alter the maximum term of three-year to four-year, only 32.92% of the Australian people were in favour of the four-year maximum term (cf. [1] for details).

From Figure 4(b), it is apparent that the exercise boundary is monotonically decreasing in t. This agrees with proposition 1 of Balke [3]. Thus earlier in its term, the Government needs higher popularity before calling an election and less popularity as time elapses. This result agrees with one of the testable implications of Smith [13] — viz. that an election is called when the government is popular. It also agrees with other findings on the so-called *reservation growth rate property of election timing* and *the declining reservation growth rate of election timing* [7].

The influence of μ on the exercise boundary is also seen in Figure 4(b). As μ increases the exercise boundary moves up and the exercise region becomes narrower — although it has no significant impact as time elapses — so the Government is less likely to call an early election. Larger μ corresponds to strong mean reversion and the speed the process reverts to zero.

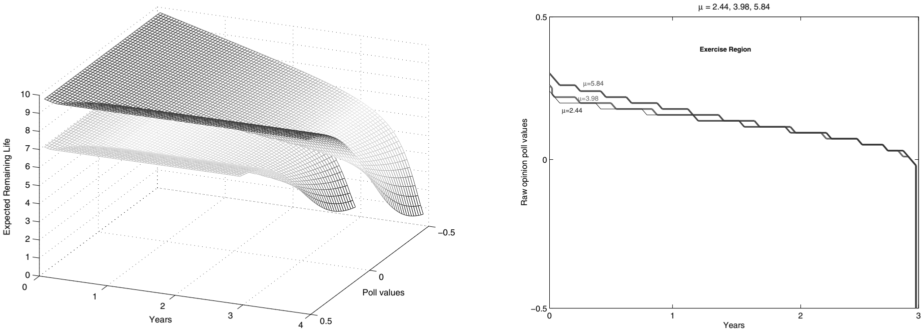


FIGURE 4. Non Constant Volatility (a) Expected Remaining Life (b) Exercise Boundary.

5. Conclusions

The dynamic of poll data has been described using a mean-reverting SDE with a non-constant volatility coefficient in a term structured volatility model. (It was evident from the data that the coefficient was not constant.)

This term structured volatility model was used together with a discrete time model for election timing, to devise an optimal control for the Government in the form of the exercise boundary. The expected remaining life in Government was found, and a comparison between a maximum term of three and four years was also provided. A four-year maximum term gave longer expected remaining life than a three-year maximum term, because the Government has more freedom to choose the best time to call an election before its term expires. The expected remaining life in Government is almost constant at the beginning of a term regardless of the level of popularity and then decreases as time elapses, especially for a low level of popularity.

From the exercise boundary discussed, the Government should call an early election when its popularity is higher than that of the Opposition. An earlier election needs a higher Government popularity, and as μ increases the exercise region becomes narrower, making the Government less likely to call an early election.

Acknowledgment

The author is grateful to Dr. Elliot Tonkes for interesting discussions on term structured volatility, and to anonymous referees for their constructive comments.

References

[1] Australian Electoral Commission, 1988 *Referendums: Statistics*, 1990.

[2] F. Black and M. Scholes, *The Pricing of Options and Corporate Liabilities.* J. Pol. Economy **81**(3) (1973), 637–654.

[3] N. S. Balke, *The Rational Timing of Parliamentary Elections.* Public Choice **65** (1990), 201–216.

[4] V. Capasso and D. Bakstein, *An Introduction to Continuous-Time Stochastic Processes: Theory, Models and Applications to Finance, Biology and Medicine.* Birkhäuser, 2005.

[5] L. Clewlow and C. Strickland, *Energy derivatives: pricing and risk management.* Lacima, 2000.

[6] R. Gençay, F. Selçuk, and B. Whitcher, *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics.* Academic Press, 2002.

[7] T. Ito, *The Timing of Elections and Political Business Cycles in Japan.* J. Asian Econ. **1** (1990), 135–156.

[8] D. S. Lemons, *An Introduction to Stochastic Processes in Physics.* The Johns Hopkins University Press, 2002.

[9] D. Lesmono, E. Tonkes and K. Burrage, *An Early Political Election Problem.* ANZIAM J. **45**(E) (2003), C16–C33.

[10] B. Øksendal, *Stochastic Differential Equations An Introduction with Applications*, 6th Edition, Springer-Verlag, 2003.

[11] W. Paul and J. Baschnagel, *Stochastic Processes From Physics to Finance.* Springer-Verlag, 1999.

[12] B. Øksendal and A. Sulem, *Applied Stochastic Control of Jump Diffussions.* Springer-Verlag, 2005.

[13] A. Smith, *Endogenous Election Timing in Majoritarian Parliamentary Systems.* Econ. Politics **8**(1996), 85–110.

[14] P. Wilmott, *Paul Wilmott on Quantitative Finance Volume Two.* John Wiley & Sons, 2000.

Dharma Lesmono
Department of Mathematics
Parahyangan Catholic University
Jalan Ciumbuleuit 94 Bandung 40141
West Java – Indonesia
e-mail: `jdharma@home.unpar.ac.id`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 293–304

Estimation for the Semiparametric Transformation Model under General Censorship

Bungon Kumphon and Vasudevan Mangalam

Abstract. In a semiparametric transformation model, an increasing transformation of the survival time is linearly related to a covariate Z with an error distribution ϵ — i.e., the survival time T has the property that $\alpha(T) = -\boldsymbol{\theta}\boldsymbol{z}+\epsilon$ given $\boldsymbol{Z} = \boldsymbol{z}$, where α is an unknown extended real-valued function on $\mathbb{R}$ and $\boldsymbol{\theta}$ is an unknown constant in $\mathbb{R}^d$. In this paper, we consider the estimation of the transformation function α and the regression coefficient $\boldsymbol{\theta}$ when the survival time data are subjected to general censorship. An observation is said to be censored by a general censorship scheme if there are random intervals which would hide the observation when it falls inside them. In such cases, we see the censoring interval instead of the actual observation. The maximum likelihood method is used to estimate the unknown parameters, and the asymptotic properties of the estimators are studied.

Mathematics Subject Classification (2000). Primary 62F10, 62F12, 62G05, 62G20; Secondary 65U05.

Keywords. Semiparametric model, Transformation model, Censored data, Interval censoring.

1. Introduction

Let ϵ be a mean zero random variable with a known continuous strictly increasing distribution function ψ. Let T be the variable of interest and $\boldsymbol{Z}$ a covariate, an element of $\mathbb{R}^d$. We are interested in the effect of $\boldsymbol{Z}$ on the response variable T. There are many such models considered in the literature, and parameter estimation under normal circumstances has been dealt with by various authors. The proportional

Results in this paper were presented at the International Conference on Mathematical Modelling and Computation held at Universiti Brunei Darussalam during 5–8 June 2006, in conjunction with the 20th anniversary of the foundation of the university.

hazards (PH) model [7] and the proportional odds (PO) model are two special cases that are well used in applications — cf. [20] for a discussion of some of the parametric and nonparametric regression models.

We consider a semiparametric transformation model where a transformation of the survival time is linearly related to a covariate Z with an error distribution ϵ. The transformation function α and $\boldsymbol{\theta}$ are the parameters to be estimated. Our motivation is that it is a strong generalisation of the ordinary regression problem (linear or nonlinear), where the relationship between the variable of interest and the covariate is precisely known. Several methods have been proposed to estimate the regression parameters under semiparametric transformation models. These include maximum semiparametric likelihood [1, 21], sieve maximum likelihood [23], maximum partial likelihood [8], and rank approximations [22]. [15] developed semiparametric estimators of α and $\boldsymbol{\theta}$ when the distribution function of the error is unknown. [12] extended this technique to censored data, and [11] developed better estimators of the transformation function and the error distribution.

In many situations, it is common to have incomplete data, and incomplete observation of the data often results from a random censoring mechanism. If data become unobservable whenever larger than the values of another variable (called the censoring variable), the observations are said to be right-censored. For right-censored data, the product-limit estimator by [18] is the nonparametric maximum likelihood estimator (NPMLE) of the unknown distribution function, and a similar estimator exists for the left-censored case. [2–4, 9, 14, 24] and others considered doubly censored data (i.e., both left and right censoring occur simultaneously), where estimators for the distribution function and its asymptotic properties have been studied. [10, 13] and others studied the case of interval-censored data, where one can only observe a censoring event and whether the time of the event of interest occurred before or after the occurrence of the censoring event. [16] studied the maximum likelihood estimator MLE for the PH model with Case 1 interval censored data, where all observations are censored by infinite intervals, and proved the asymptotic normality of the MLE for the regression parameter. [19] considered the MLE for the PH model with partly interval-censored data, where the data consist of exact data and interval-censored data. [6] studied an efficient semiparametric estimation of censored and truncated regression, based on a new approach for estimating the density function of the residual in a partially observed regression.

The type of censoring we consider, and refer to here as general censorship, is a generalisation of the different types of censoring in the literature. Under this censorship, some of the data become unobservable when they fall inside a finite or infinite random interval. Various combinations of finite and infinite intervals yield all the different types of censorship in the literature (left censoring, right censoring, double censoring and different cases of interval censoring) as special cases — cf. [17] for a detailed discussion.

In this paper, we consider the estimation of α and $\boldsymbol{\theta}$ when failure times are subjected to general censorship. We use the maximum likelihood method to estimate the unknown parameters, and investigate the large sample properties of the estimators. Numerical simulation is used to generate the data, and estimates of parameters are computed for these data. The accuracy of the estimates is demonstrated by computing the maximum and the average distance between the estimate and the true value. Histograms and quantile plots for the regression coefficient are given to demonstrate the asymptotic normality of the regression estimator.

The semiparametric transformation model (STM) is formally defined in Section 2, and the STM under general censorship and computation of the MLE are discussed in Section 3. The results of the numerical simulations are presented in Section 4.

2. Semiparametric transformation model

Let us consider a class of transformation models where a transformation of the survival time is linearly related to a covariate with error distribution ϵ — viz. a d-dimensional covariate $\boldsymbol{Z}$ such that

$$\alpha\left(T\right)=-\boldsymbol{\theta z}+\epsilon$$

given $\boldsymbol{Z}=\boldsymbol{z}$, where α is an unknown extended real-valued function on $\mathbb{R}$ and $\boldsymbol{\theta}$ is an unknown constant in $\mathbb{R}^d$. The distribution of ϵ is denoted by ψ, assumed to be known. The function α is called the *transformation function*, and $\boldsymbol{\theta}$ is referred to as the *regression coefficient*. The transformation function α is assumed to satisfy the following conditions:

1. α is monotonic increasing.
2. $\lim\limits_{t\to\pm\infty}\alpha\left(t\right)=\pm\infty$.

This model is known as the semiparametric transformation model. Further, let $F\left(\cdot|\boldsymbol{z}\right)\equiv F_z\left(\cdot\right)$ be the distribution function of T given $\boldsymbol{Z}=\boldsymbol{z}$. It is straightforward to verify that

$$F\left(t|\boldsymbol{z}\right)=\psi\left(\alpha\left(t\right)+\boldsymbol{\theta z}\right).$$

Well-known examples are the proportional hazards and proportional odds models. The PH model is obtained when the error term follows an extreme value distribution. Specifically, let λ be an unknown nonnegative continuous function from $[0,\infty)$ to $[0,\infty)$ so that $\Lambda=\int_0^t\lambda\left(t\right)dt$ is a monotonically increasing function from $[0,\infty)$ to $[0,\infty)$. Then the distribution function for the PH model is given by

$$F\left(t|\boldsymbol{z}\right)=1-\exp\left(-\Lambda\left(t\right)e^{\boldsymbol{\theta z}}\right).$$

The PH model can be seen to be an example of semiparametric transformation model, by setting

- $\psi\left(x\right)=1-\exp\left(-e^x\right)$, and
- $\alpha\left(t\right)=\log\Lambda\left(t\right)$ if $t>0$ and $-\infty$ otherwise.

The PO model arises when the error term has a logistic distribution. Setting $\psi\left(x\right)=\left(1+e^{-x}\right)^{-1}$, one has

$$F(t|\boldsymbol{z}) = \frac{\exp(\alpha(t) + \boldsymbol{\theta z})}{1 + \exp(\alpha(t) + \boldsymbol{\theta z})}.$$

3. STM under general censorship

Let T_i, $i = 1, ..., n$, be a sequence of i.i.d. random variables with distribution F. Let (L_i, R_i), $i = 1, ..., n$ represent the censoring mechanism, consisting of n pairs of i.i.d. extended real-valued random variables (L_i, R_i) such that $P(L < R) = 1$. The i^{th} observation is said to be censored if $T_i \in (L_i, R_i)$, when we do not get to observe T_i. Let $\delta_i = I[T_i \notin (L_i, R_i)]$, so that $\delta_i = 0$ if the i^{th} observation is censored and 1 if it is uncensored. We assume that T_i and (L_i, R_i) are independent given the concomitant variable $\boldsymbol{Z}$.

The density function of T is given by $f(t) = \psi'(\alpha(t) + \boldsymbol{\theta z})\alpha'(t)$, so the likelihood and the log-likelihood functions in the presence of censoring are given by

$$L_n(\boldsymbol{\theta}, \alpha) = \prod_{i=1}^{n} [f_z(t_i)]^{\delta_i} [F_z(r_i-) - F_z(l_i)]^{1-\delta_i}$$

and

$$\begin{aligned} \log L_n &= \sum_{i=1}^{n} \{\delta_i \left(\log[\psi'(\alpha(t_i) + \boldsymbol{\theta z}_i)] + \log \alpha'(t_i)\right) \\ & \quad +(1 - \delta_i) \log[\psi(\alpha(r_i-) + \boldsymbol{\theta z}_i) - \psi(\alpha(l_i) + \boldsymbol{\theta z}_i)]\}. \end{aligned}$$

The value of this expression depends on the function α and its derivative only at the jump points, and it can be made arbitrarily large by making $\alpha'(t_i)$ as large as we want without affecting the values of $\alpha(t_i)$. Consequently, we work with a discretised version of the likelihood function, where $\alpha'(t_i)$ is replaced by a discrete jump size a_i and $\alpha(t_i)$ by $A_i = \sum_{k=0}^{i} a_k$. Thus we find the function that maximises this modified log-likelihood among all α's such that α is an increasing step function that is a constant for $t < t_1$ and has jumps at t_i's.

We replace any empty censoring interval (a censoring interval that contains no uncensored observations) by its midpoint as an uncensored observation, group the censored and uncensored observations separately, and reorder the uncensored observations in ascending order. Let n_1 be the size of the exact data and $n_2 = n - n_1$ be the size of the censored data. Let $a_0 = \alpha(t_1-)$ and a_i be the jump size at t_i for $i = 1, ..., n_1$. Then the modified version of the log-likelihood is given by

$$\begin{aligned} l_n(\boldsymbol{\theta}, \boldsymbol{a}) &= \sum_{i=1}^{n_1} \{\log [\psi' (A_i + \boldsymbol{\theta z}_i)] + \log a_i\} \\ & \quad + \sum_{j=1}^{n_2} \log \left[\psi \left(\sum_{k:t_{(k)} < r_j} a_k + \boldsymbol{\theta z}_j \right) - \psi \left(\sum_{k:t_{(k)} \leq l_j} a_k + \boldsymbol{\theta z}_j \right) \right] , \end{aligned}$$

which is to be maximised under the constraint that all of the a_i's except a_0 are non-negative. If $\hat{a_i}$, $i = 0$ to n_1 maximises $l_n(\boldsymbol{\theta}, \boldsymbol{a})$, then the function $\hat{\alpha}$ defined as an increasing step function with jumps $\hat{a_i}$ at t_i and value a_0 for $t < t_1$ is the MLE of α.

Let $\Theta \subset \mathbb{R}^d$ be the d-dimensional parameter space of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0$ and α_0 denote the true value of the parameters and let F_0 be the true conditional distribution function of T given $\boldsymbol{Z}$. The MLE's of α, $\hat{\alpha}$ and $\hat{\boldsymbol{\theta}}$, and $\boldsymbol{\theta}$, are obtained by maximising $l_n(\boldsymbol{\theta}, \boldsymbol{a})$ over $\Theta \times \boldsymbol{A}$ where $\boldsymbol{A} = \mathbb{R} \times \mathbb{R}^{+n_1}$, the set of $n_1 + 1$ dimensional vectors whose coordinates are all nonnegative except the first. Under some mild assumptions, it can be shown that $l_n(\boldsymbol{\theta}, \boldsymbol{a})$ is strictly concave for each n, and that it is bounded above. It therefore has a unique maximiser, and the maximiser $(\hat{\theta}, \hat{a})$ can be obtained by equating the first derivative to zero and using the multivariate Newton–Raphson algorithm.

4. Numerical simulation

We now present simulations of two semiparametric transformation models — viz. the PH model where ϵ has a standard extreme value distribution yielding $P(\epsilon \leq t) = 1 - \exp(-e^t)$, and the PO model where ϵ has a logistic distribution given by $P(\epsilon \leq t) = \frac{1}{1+e^{-t}}$. The transformation functions chosen were $\alpha_0(t) = \log t$ and $\alpha_0(t) = t$, the dimension d of the regression coefficient $\boldsymbol{\theta}$ was taken to be 1, and two true values of $\boldsymbol{\theta}$ (viz. $\boldsymbol{\theta}_0 = 0$ and $\boldsymbol{\theta}_0 = 1$) were considered. The values of the covariate $\boldsymbol{Z}$ were randomly generated from a standard normal distribution. The random variables L_i and R_i were produced according to four different levels of censoring, set by letting L_i and R_i be the minimum and the maximum of c independent exponential random variables for $c = 1, 2, 3$ and 5. The sample size n was taken to be 100, and each combination of all these factors was replicated 300 times. The implementations of data generation and computation were written in the statistical software package S-Plus. The non-linear optimisation routine *Nonlinear Minimization subject to Box Constraints* (NLMINB) based on the multivariate Newton–Raphson algorithm was used for numerical maximisation. The percentage of censoring, denoted by r, was taken to be the average of the censoring percentages for the 300 replications.

The transformation function α is estimated by $\hat{\alpha}_n$, and the distribution function F by $\hat{F}_n$. The performance of $\hat{F}_n$ is more important than that of $\hat{\alpha}_n$ [5], and can be measured by $\|\hat{F}_n - F_0\| = \max_t |\hat{F}_n(t) - F_0(t)|$. Another measure of this performance that we calculated is the mean of $|\hat{F}_n - F_0|$, defined as $n_1^{-1} \sum_{i=1}^{n_1} |\hat{F}_n(t_i) - F_0(t_i)|$. The averaging was done over n_1 uncensored data points, because those are the points where $\hat{F}_n$ has jumps. The results are reported in Tables 1 to 8. The accuracy of estimation when there is no censoring, produced by the case $c = 1$, is presented as $r = 0$ for comparison.

Tables 1 to 4 show the estimated value of $\boldsymbol{\theta}$ and the performance of $\hat{F}_n$ for PH model. Results show that the estimates of $\boldsymbol{\theta}$ are good overall and are not

TABLE 1. Simulation results for estimated values of $\alpha_0(t) = \log(t)$ and $\boldsymbol{\theta}_0 = 0$ in PH model.

$\boldsymbol{\theta}_0 = 0$	Censoring percentage							
	$r = 0$		$r = 30.89$		$r = 45.99$		$r = 62.06$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	0.0101	0.0111	0.0032	0.0120	0.0039	0.0103	0.0107	0.0146
$\lVert \hat{F}_n - F_0 \rVert$	0.0788	0.0007	0.0920	0.0010	0.1211	0.0018	0.2224	0.0069
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0314	0.0002	0.0344	0.0002	0.0402	0.0003	0.0465	0.0002

TABLE 2. Simulation results for estimated values of $\alpha_0(t) = \log(t)$ and $\boldsymbol{\theta}_0 = 1$ in PH model.

$\boldsymbol{\theta}_0 = 1$	Censoring percentage							
	$r = 0$		$r = 26.98$		$r = 39.98$		$r = 54.25$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	1.0357	0.0187	1.0254	0.0187	1.0063	0.0103	0.9866	0.0301
$\lVert \hat{F}_n - F_0 \rVert$	0.0839	0.0008	0.1042	0.0011	0.1284	0.0023	0.2500	0.0073
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0295	0.0001	0.0302	0.0001	0.0314	0.0001	0.0386	0.0001

TABLE 3. Simulation results for estimated values of $\alpha_0(t) = t$ and $\boldsymbol{\theta}_0 = 0$ in PH model.

$\boldsymbol{\theta}_0 = 0$	Censoring percentage							
	$r = 0$		$r = 9.93$		$r = 14.14$		$r = 19.92$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	-0.0033	0.0112	-0.0005	0.0124	0.0070	0.0140	0.0204	0.0094
$\lVert \hat{F}_n - F_0 \rVert$	0.0808	0.0008	0.0817	0.0005	0.0872	0.0006	0.1096	0.0013
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0318	0.0002	0.0319	0.0001	0.0337	0.0002	0.0349	0.0003

TABLE 4. Simulation results for estimated values of $\alpha_0(t) = t$ and $\boldsymbol{\theta}_0 = 1$ in PH model.

$\boldsymbol{\theta}_0 = 1$	Censoring percentage							
	$r = 0$		$r = 12.93$		$r = 24.33$		$r = 25.09$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	1.0337	0.0160	1.0061	0.0179	0.9954	0.0305	0.9389	0.0251
$\lVert \hat{F}_n - F_0 \rVert$	0.0899	0.0010	0.0890	0.0009	0.0924	0.0861	0.1145	0.0012
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0300	0.0002	0.0316	0.0002	0.0384	0.0002	0.0342	0.0002

TABLE 5. Simulation results for estimated values of $\alpha_0(t) = \log(t)$ and $\boldsymbol{\theta}_0 = 0$ in PO model.

$\boldsymbol{\theta}_0 = 0$	Censoring percentage							
	$r = 0$		$r = 26.33$		$r = 40.08$		$r = 53.77$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	0.0004	0.0350	0.0024	0.0238	0.0119	0.0310	0.0054	0.0450
$\lVert \hat{F}_n - F_0 \rVert$	0.0782	0.0007	0.0939	0.0014	0.1125	0.0019	0.1988	0.0054
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0304	0.0002	0.0327	0.0003	0.0319	0.0001	0.0379	0.0002

TABLE 6. Simulation results for estimated values of $\alpha_0(t) = \log(t)$ and $\boldsymbol{\theta}_0 = 1$ in PO model.

$\boldsymbol{\theta}_0 = 1$	Censoring percentage							
	$r = 0$		$r = 23.56$		$r = 35.43$		$r = 48.44$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	1.0115	0.0308	1.0298	0.0408	1.0148	0.0405	0.9993	0.0469
$\lVert \hat{F}_n - F_0 \rVert$	0.0829	0.0008	0.0981	0.0010	0.1205	0.0020	0.2131	0.0066
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0296	0.0002	0.0300	0.0001	0.0311	0.0001	0.0342	0.0001

TABLE 7. Simulation results for estimated values of $\alpha_0(t) = t$ and $\boldsymbol{\theta}_0 = 0$ in PO model.

$\boldsymbol{\theta}_0 = 0$	Censoring percentage							
	$r = 0$		$r = 17.96$		$r = 26.84$		$r = 35.44$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	-0.0125	0.0324	0.0125	0.0277	-0.0005	0.0296	-0.0136	0.0346
$\lVert \hat{F}_n - F_0 \rVert$	0.0800	0.0007	0.0871	0.0007	0.1045	0.0011	0.1869	0.0038
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0311	0.0002	0.0340	0.0002	0.0344	0.0002	0.0387	0.0002

TABLE 8. Simulation results for estimated values of $\alpha_0(t) = t$ and $\boldsymbol{\theta}_0 = 1$ in PO model.

$\boldsymbol{\theta}_0 = 1$	Censoring percentage							
	$r = 0$		$r = 18.60$		$r = 28.31$		$r = 36.73$	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
$\hat{\boldsymbol{\theta}}$	1.0236	0.0326	1.0042	0.0413	1.0440	0.0332	0.9867	0.0425
$\lVert \hat{F}_n - F_0 \rVert$	0.0836	0.0008	0.0889	0.0008	0.1049	0.0016	0.1827	0.0034
Mean of $\lvert \hat{F}_n - F_0 \rvert$	0.0297	0.0002	0.0307	0.0002	0.0329	0.0002	0.0380	0.0003

TABLE 9. The value of W for the PH and PO models.

Model	$\boldsymbol{\theta}_0$	$\alpha_0(t)$	Censoring percentage			
PH	$\boldsymbol{\theta}_0 = 0$	$\log t$	$r = 0$	$r = 30.89$	$r = 45.99$	$r = 62.06$
			0.0466	**0.0436**	**0.0451**	**0.0321**
		t	$r = 0$	$r = 9.93$	$r = 14.14$	$r = 19.92$
			0.0342	**0.0350**	**0.0319**	**0.0359**
	$\boldsymbol{\theta}_0 = 1$	$\log t$	$r = 0$	$r = 26.98$	$r = 39.98$	$r = 54.25$
			0.0317	**0.0445**	**0.0400**	**0.0420**
		t	$r = 0$	$r = 12.93$	$r = 24.33$	$r = 25.09$
			0.0885	**0.0311**	**0.0452**	**0.0272**
PO	$\boldsymbol{\theta}_0 = 0$	$\log t$	$r = 0$	$r = 26.33$	$r = 40.88$	$r = 53.77$
			0.0369	**0.0260**	**0.0231**	**0.0381**
		t	$r = 0$	$r = 17.96$	$r = 26.84$	$r = 35.44$
			0.0404	**0.0321**	**0.0383**	**0.0370**
	$\boldsymbol{\theta}_0 = 1$	$\log t$	$r = 0$	$r = 23.56$	$r = 35.43$	$r = 48.44$
			0.0450	**0.0240**	**0.0339**	**0.0324**
		t	$r = 0$	$r = 18.60$	$r = 28.31$	$r = 36.73$
			0.0361	**0.0415**	**0.0405**	**0.0475**

drastically influenced by the severity of censoring. As for the performance of $\hat{F}_n$, the values of both $\|\hat{F}_n - F_0\|$ and mean of $|\hat{F}_n - F_0|$ perform fairly well but they are influenced by the rate of censoring. As the censorship rate increases, the error increases, but continues to be within acceptable limits.

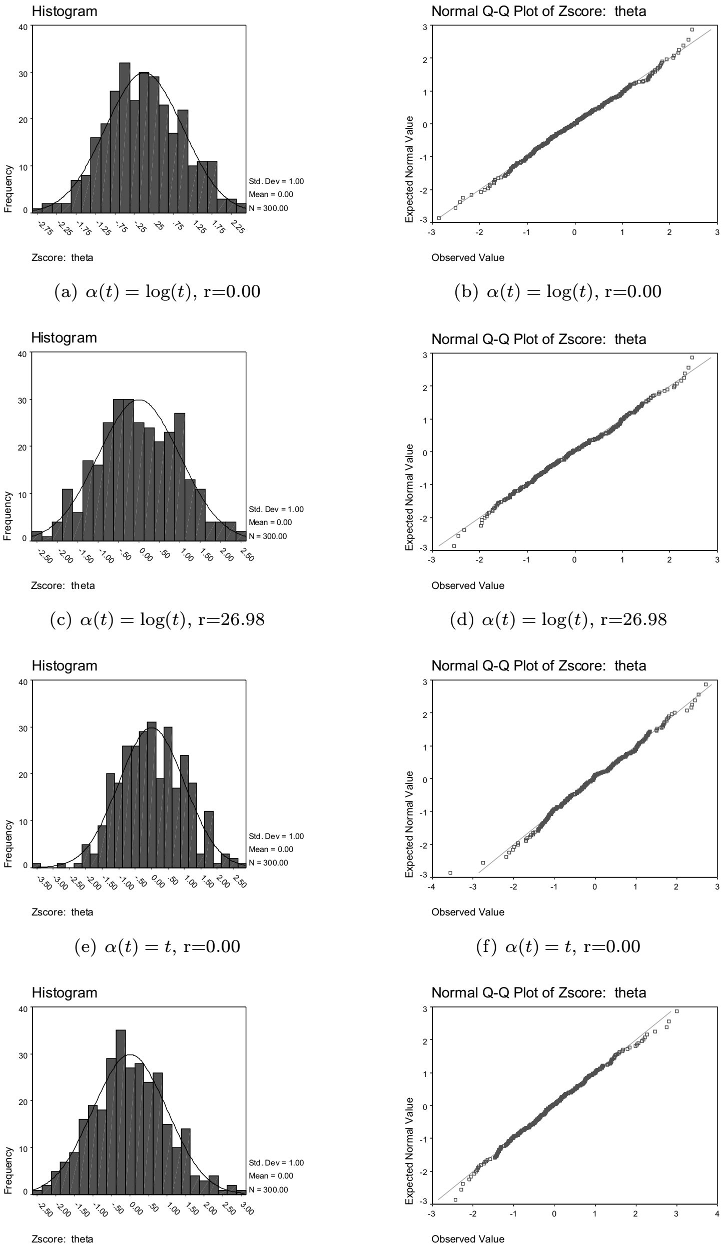
Tables 5 to 8 show the results for the PO model, and the conclusions are similar. Thus under both of the PH and PO models, our method of parameter estimation of $\boldsymbol{\theta}$ and F performs fairly well.

We also calculated the distance between the observed distribution of standardised $\boldsymbol{\theta}$ and the standard normal distribution, measured by

$$W = \max_{1 \le i \le m} \left(\left| \frac{i}{m} - \Phi(T_i) \right| \vee \left| \frac{i-1}{m} - \Phi(T_i) \right| \right)$$

where T_i is the standardised value of $\hat{\boldsymbol{\theta}}$, Φ is the cumulative distribution function of the standard normal distribution and m is the number of replications (viz. 300, as mentioned above). The results given in Table 9 show that the calculated values of W are small for all cases, and hence the observed distribution of standardised $\boldsymbol{\theta}$ is close to the standard normal distribution.

The asymptotic normality of $\boldsymbol{\theta}$ was also examined by the Kolmogorov–Smirnov test. The SPSS program gave 0.200 as a lower bound for the p-value, for all cases studied. This high p-value is strong evidence that it is reasonable to infer the distribution of $\hat{\boldsymbol{\theta}}$ is normal.



(a) $\alpha(t) = \log(t)$, r=0.00

(b) $\alpha(t) = \log(t)$, r=0.00

(c) $\alpha(t) = \log(t)$, r=26.98

(d) $\alpha(t) = \log(t)$, r=26.98

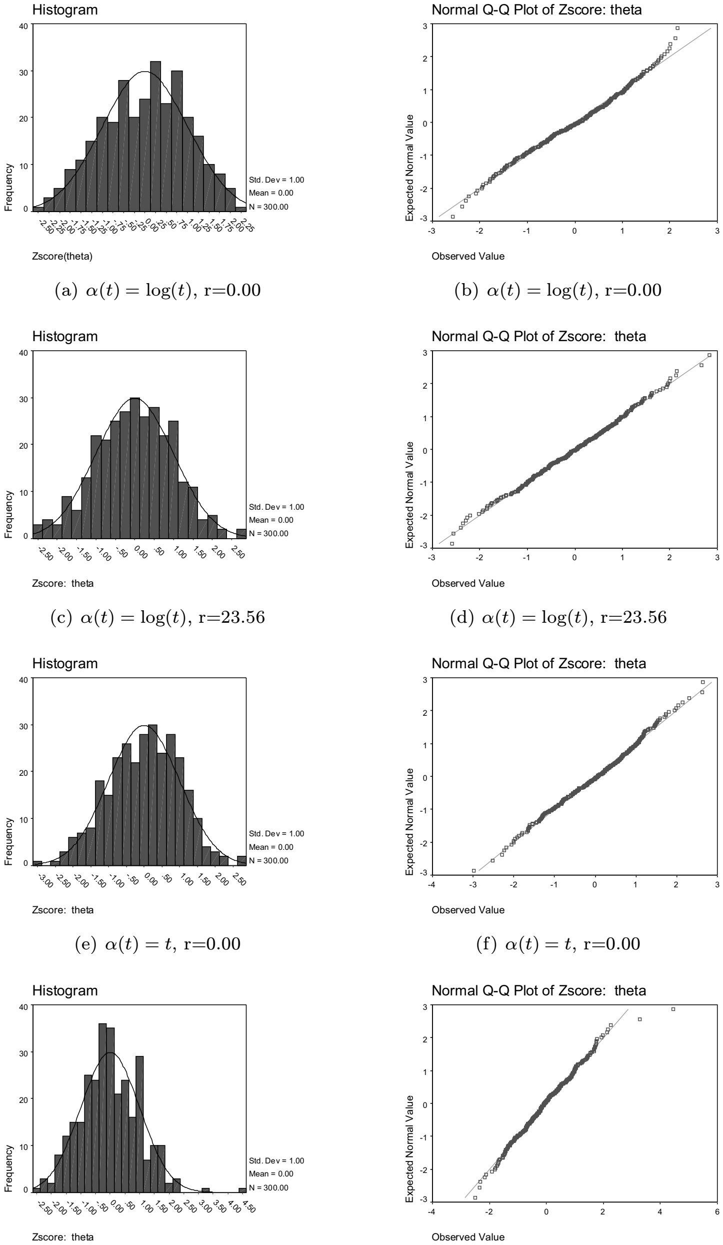
(e) $\alpha(t) = t$, r=0.00

(f) $\alpha(t) = t$, r=0.00

(g) $\alpha(t) = t$, r=12.93

(h) $\alpha(t) = t$, r=12.93

FIGURE 1. The histogram and normal Q-Q plot of standardised $\hat{\boldsymbol{\theta}}$ for PH model when $\boldsymbol{\theta} = 1$ with various rate of censoring, and $\alpha(t) = \log(t)$ and $\alpha(t) = t$.



(a) $\alpha(t) = \log(t)$, r=0.00

(b) $\alpha(t) = \log(t)$, r=0.00

(c) $\alpha(t) = \log(t)$, r=23.56

(d) $\alpha(t) = \log(t)$, r=23.56

(e) $\alpha(t) = t$, r=0.00

(f) $\alpha(t) = t$, r=0.00

(g) $\alpha(t) = t$, r=18.60

(h) $\alpha(t) = t$, r=18.60

FIGURE 2. The histogram and normal Q-Q plot of standardised $\hat{\boldsymbol{\theta}}$ for PO model when $\boldsymbol{\theta} = 1$ with various rate of censoring, and $\alpha(t) = \log(t)$ and $\alpha(t) = t$.

5. Conclusions

Maximum likelihood estimation for the regression coefficients and the transformation function were carried out for a semiparametric transformation model, where survival time data are subjected to general censorship. The optimisation steps were carried out through a multivariate Newton–Raphson method using the NLMINB procedure of S-Plus. In all our simulations the global maximum was attained and the maximum likelihood estimates performed fairly well in the sense that the estimated values were close to the true values of the corresponding parameter. Asymptotic normality of the maximum likelihood estimators for the regression coefficients was verified by a Kolmogorov–Smirnov test using the SPSS package.

References

[1] Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273-277.

[2] Cai, T. and Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika* **91**, 277-290.

[3] Chang, M. N. and Yang, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15**, 1536–1547.

[4] Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* **18**, 391–404.

[5] Cheng, Y.-C. (2002). *Estimation in semiparametric transformation models with doubly censored data.* Ph.D. thesis, Department of Statistics, Rutgers University.

[6] Cosslett, S. R. (2004). Efficient semiparametric estimation of censored and truncated regressions via a smoothed self-consistency equation. *Econometrica.* **72**, 1277-1293.

[7] Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc.* Ser. B **34**, 187-220.

[8] Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scand. J. Statist.* **5**, 1-23.

[9] Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* **52** 650-653.

[10] Geskus, R. B. and Groeneboom, P. (1996). Asymptotically optimal estimation of smooth functionals for interval censoring. I. *Statist. Neerlandica* **50**, 69-88.

[11] Gørgens, T. (2003). Semiparametric estimation of censored transformation models. *J. Nonparametr. Statist* **15**, 377-393.

[12] Gørgens, T. and Horowitz, J. L. (1999). Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *J. Econometrics* **90**, 155-191.

[13] Groeneboom P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* DMV Seminar, **19**. Birkhäuser Verlag, Basel.

[14] Gu, M. G. and Zhang, C.-H. (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.* **21**, 611–624.

[15] Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* **64**, 103-137.

[16] Huang, J. (1996) Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540-568.

[17] Jammalamadaka, S. Rao and Mangalam, V. (2003). Nonparametric estimation for middle censored data. *J. Nonparametr. Statist.* **15**, 253-265.

[18] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J.Amer. Statist. Assoc.* **53**, 457-481.

[19] Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *J. Roy. Statist. Soc.* Ser. B **65**, 489-502.

[20] Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. John Wiley & Sons, New York.

[21] Murphy, S. A., Rossini, A. J and Van Der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Am. Statist. Assoc.* **92**, 968-976.

[22] Pettitt, A. N. (1984). Proportional odds model for survival data and estimates using ranks. *Applied Statistics* **33**, 169-175.

[23] Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **85**, 165-177.

[24] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69**, 169-173.

Bungon Kumphon
Department of Mathematics
Universiti Brunei Darussalam
Jalan Tungku Link
Gadong BE1410
Negara Brunei Darussalam
e-mail: `m03h8451@stu.ubd.edu.bn`

Vasudevan Mangalam
Department of Mathematics
Universiti Brunei Darussalam
Jalan Tungku Link
Gadong BE1410
Negara Brunei Darussalam
e-mail: `mangalam@fos.ubd.edu.bn`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 305–320

Integer Programming Models of Bookmobile Routing

Les R. Foulds, Stein W. Wallace, John Wilson and Martin West

Abstract. A bookmobile is a specially adapted bus or van used as part of the outreach operations of public library systems. Bookmobiles play a significant part in the service of the public library system in Buskerud County, Norway. They are used to deliver and collect library materials (printed books, audio books, periodicals, and music) to and from borrower groups throughout the County, many in remote areas. The question of how best to utilise the County's bookmobile resources can be modelled as an interesting variation of one of the classical models of operational research — the travelling salesman problem. The combination of the features that make this scenario non-standard include multiple depots, simultaneous cost minimisation and prize collection objectives, differing customer service levels, time windows, route start time flexibility for some routes, multiple route duration restrictions, route lunch breaks, and overnight stays on certain routes. We report on models for the bookmobile problem, and the outcome of its application to the Buskerud County bookmobile system.

Mathematics Subject Classification (2000). 90B06, 90C10.

Keywords. Bookmobile problem, Vehicle scheduling, Integer programming, Case study.

1. Introduction

Buskerud County, Norway, is located to the west of Oslo. The public libraries of the County, based in the towns of Drammen and Gol, both operate bookmobiles to serve identified groups of its borrowers that cannot visit a library. These bookmobiles, which are specially designed vans, carry various library materials, such as: printed books, audio books, periodicals, and music, to road lay-bys and private farmhouses, in remote parts of the County, and to schools and other libraries. The Buskerud Library Department is under considerable financial pressure to make efficient and effective use of its bookmobile operations. This led to us being invited by those responsible for the bookmobile system to examine their operations, with view to making suggestions for improvement. It became clear to us that the question of how to best utilise the County's bookmobile resources can be modelled as an interesting variation of one of the classical models of operational research; the travelling salesman (TSP). The TSP can be stated as: given a number of locations and the costs of travelling from any location to any other location, what is the least cost round-trip tour that visits each city exactly once and then returns to the starting location? The TSP model is quite important in industry as a number of common endeavours, such as vehicle routing, scheduling, integrated physical mapping, and circuit design can be formulated as TSP's. In the present chapter we focus on a scheduling solution, hence we review some of the relevant TSP literature to highlight some key developments in this area.

The origins of the TSP are obscure. A discussion of the early work on the problem by the British mathematicians Hamilton and Kirkman, can be found in the book by Biggs et al [4]. In the 1930s, the mathematician and economist Karl Menger [18] popularized the TSP as the "messenger problem". Eventually, the TSP gained notoriety as the prototype of a hard problem in combinatorial optimization. A breakthrough came in 1954 when Dantzig et al. [10] published a TSP method and illustrated its power by solving an instance with 49 locations, an impressive size at that time. In 1963, Little et al. [17] were among the first to apply integer programming to the TSP and coined the term "branch-and-bound" In their approach, subsets of tours are conveniently represented as the nodes of a decision tree and the process of partitioning these subsets follows the branching of the tree. Later, the application of spanning trees to devise effective solution methods was also an important step forward (Held and Karp[14]). In the 1970s and 80s, when it was established that constructing optimal solutions to large-scale, general numerical instances of the TSP was beyond current computational capacity, research into the problem was widened from exact methods to heuristic (approximate) solution methods. For surveys of progress during that period, the reader is referred to the reports of Burkard [6], Lawler et al. [16], and Rosencrantz et al. [23]. Among the numerous TSP heuristics that have been devised since then, some have been based on the "genetic algorithm" (GA) learning meta-heuristic approach, which takes ideas from genetics and natural selection. A GA is sometimes employed to solve difficult optimization problems, where traditional techniques are less efficient. However, the main limitation with GA's is that they are not very effective at solving large TSP's (Tsai et al. [25]). When the number of locations is above a few thousand, then a more advanced approach, termed a "hybrid genetic algorithm", can sometimes be successful (Nguyen [20]). See Gutin and Punnen [13] for a comprehensive summary of other TSP work to date.

One of the variations of the TSP is termed the multiple travelling salesman problems (MTSP). It involves a given set of locations, one of which is designated as the "home location" where a given set of salesmen are based. The problem is to find a minimum-cost assignment of some, or all of the salesmen to individual tours that each begin and end at the home location, such that each other location is visited by exactly one salesman. Although MTSP models are employed in many industrial applications (such as: crew scheduling, school bus routing, print press scheduling, and mission planning), Tolga [24] stated that the MTSP has not received the same amount of research attention as the TSP. Previous studies investigated solving the MTSP with GA's using standard TSP chromosomes and operators, as discussed earlier. However, Catera and Ragsdale [8] proposed novel GA chromosome and related operators for the MTSP and compared the theoretical properties and computational performance of the proposed technique to previous methods. Computational testing shows that the new approach results in a smaller search space and, in many cases, produces better solutions than the previous techniques. A recent paper by Bektas [3] studied the MTSP thoroughly, surveying the literature on it, discussing its practical applications, highlighting key formulations, and describing exact and heuristic solution procedures.

A further variation of the TSP is termed the travelling salesman problem with time windows (TSPTW). In this case, the travel costs are given as travel times and each location has associated with it a "time window" – representing the earliest and latest times by which the location can be visited. An early paper by Baker [2] reported an exact algorithm for a version of this problem. More recent work involving exact methods for the problem include a dynamic programming approach by Dumas et al. [11], a branch-and-cut method due to Ascheuer et al. [1] and the method of Focacci et al. [12] who developed a hybrid exact algorithm. Heuristic methods for the TSPTW include those by Carlton and Barnes [7] who used tabu search, Pessant et al. [22] who used constraint logic programming, and Wolfler-Calvo [26] who based his approach on the assignment problem with a parametric objective function. Also, Ohlmann and Thomas [21] describe a variant of the "simulated annealing" learning meta heuristic, incorporating a variable penalty method to solve the TSPTW. Augmenting temperature from traditional simulated annealing with the concept of pressure (analogous to the value of a penalty multiplier), compressed annealing relaxes the time-window constraints by integrating a penalty method within a stochastic search procedure. Computational results validate the value of a variable-penalty method versus a static-penalty approach. Compressed annealing compares favourably with benchmark results in the literature, obtaining best-known results for numerous instances.

Extensions to the TSPTW include dynamic time windows (Larsen et al. [15]).These authors examine the TSPTW for various degrees of dynamism, in the sense that part or all of the necessary information becomes available during the day of operation. They seek to minimize lateness, and examine the impact of this criterion choice on the total travel cost. Their focus on lateness is motivated by the problem faced by overnight mail service providers. As the TSPTW is usually harder to solve than the TSP for problems of the same size, instances solved are necessarily of smaller-scale than those reported in the TSP literature, where very large-scale instances can now be solved to optimality, albeit requiring large amounts of computer time.

When the MTSP and the TSPTW above are combined we have a scenario with both multiple travelling salesmen and time windows (MTSPTW). This model is often more appropriate for practical applications than the above-mentioned models. It is also possible to extend the model to a wide variety of routing and scheduling problems with immediate applications in road vehicle, ship, and airline scheduling problems. This can be achieved by incorporating some additional side constraints, often related to vehicle capacity and fixed service costs. Mitrovic-Minic and Krishnaurti [19] use precedence graphs to establish bounds on the minimum number of salesmen needed to visit all locations in the MTSPTW. Also, Chandran et al. [9] employ a clustering approach for the MTSPTW to address the issue of workload balance among the salesmen.

As will be seen in the discussion that follows, the bookmobile routing problem is an MTSPTW with additional side constraints. Unfortunately, the side constraints are many, varied, and significant, including: the existence of multiple home locations, simultaneous cost minimization and prize collection objectives, differing customer service levels, route start time flexibility, multiple route duration restrictions, route lunch breaks, and overnight stays on certain routes. As far as the authors are aware, this particular combination of side constraints has not been considered previously in the literature. Indeed, the only related side constraints that appear to have been discussed in the open literature involve meal-break and start-time flexibility for manpower planning (Brusco and Jacobs [5], but these have little relevance for the MTSPTW. For these reasons, it seems difficult to apply the solution techniques previously discussed to the bookmobile problem. Thus, the authors elected to develop integer programming models of the problem and to apply powerful, existing integer programming solution techniques.

In the next section we describe a special case of the County's bookmobile system, a model of its operations, and the outcome when we applied our approach to the actual problem at hand. In Section 3 we generalise the description of the operation of Section 2 to more general scenarios. We end the paper with conclusions drawn from this case study and suggested directions for further research.

2. The Drammen bookmobile system

The first, and most pressing, question that we were asked to examine was how to improve the bookmobile service to a selected number of borrower locations surrounding the town of Drammen.

2.1. A Description of the Drammen operation

A single bookmobile is operated, once per week, over a four-weekly cycle. It must service certain borrower locations (termed "compulsory borrowers") once every fourteen days (i.e., twice), but at the same time of day. This is because some of the boroughs that comprise the County pay for certain borrowers located within them to receive this service. There is also another type of borrower (termed "optional borrowers") each of whom may be visited at most once during the cycle. The optional borrowers are private individuals it is desirable, but not strictly necessary, to visit. The regime of identified routes (termed "runs") is carried out once, in the same way, during each time cycle. The primary issue is one of devising a set of feasible runs for the bookmobile so that the maximum number of (optional) unweighted borrower locations are visited. The secondary issue is one of achieving this at minimal cost. We now go on to describe the operation in more detail, and highlight factors that constrain the operation.

Each run begins at Drammen, visits a given sequence of borrowers that it services in turn, and finally returns to Drammen. The locations of all borrowers are known, as is the (season-dependent) travel time for the bookmobile to traverse all feasible road segments that link them. (Because of the terrain of the County, the road network is somewhat sparse.) We can assume, without loss of generality, a given set of road segment driving times for a given season of the year. Because it is assumed by the planners that costs are directly proportional to time, the secondary issue, mentioned above, reduces to one of minimising the total elapsed time (travel, service, break, and idle time, combined over all runs).

Because the capacity of the bookmobile is ample to service any combination of borrowers on any run, the vehicle is not "capacitated". If any borrower is visited at all, its servicing takes a known time (termed its "duration"). Moreover, each borrower can be serviced only during a known time window, which remains constant for the borrower throughout the cycle. Servicing starts immediately upon arrival. Suppose that the time windows of two borrowers who are visited one immediately after the other, on a run, are such that idle time is necessary. The idle time must occur just before the departure from the earlier borrower. Furthermore, none of the runs can exceed a given number of time periods; there is a restriction on the average time of all runs; and each run must contain a continuous break that must overlap with the time interval from three time periods before the midpoint (in terms of time) to three time periods after the midpoint. Each break takes place at a borrower, immediately after servicing that borrower. Finally, the start time of each run is arbitrary.

2.2. A Model of the Drammen operation

We now create a model of the operation that has just been described. To this end we first introduce the necessary notation.

Basic dimensions.

n. the number of locations. (The depot, being the town of Drammen, is denoted by location 1. The locations of the compulsory borrowers are denoted by locations: $2, 3, \dots, q$. The locations of the optional borrowers are denoted by locations: $q+1, q+2, \dots, n$; where $1 < q < n$.)

m. the given number of runs to be carried out in each time cycle.

k. the index of the run carried out in the kth week of the time cycle.

Input data.

M, N. relatively large, given numbers,

d_i. duration of borrower i,

a_i. the earliest time period during which the servicing of borrower i can begin,

b_i. the latest time period during which the servicing of borrower i can begin. (The interval $[a_i, b_i]$ represents the time window during which the servicing of borrower i must begin, if it takes place at all.)

t_{ij}. travel time in proceeding directly from location i to location j. (For technical reasons it is assumed that all $t_{ij} > 0$. The triangle inequalities $t_{ij} \leq t_{ih} + t_{hj}$, for all i, h, j do not have to be satisfied unless column generation is used.)

u. the length of the break on each run, expressed as a number of time periods,

P. maximum allowable number of time periods for any run, and

T. maximum average number of time periods over all runs of a complete cycle.

Decision variables.

s_k the start time of the kth run,

$$v_{ik} \begin{cases} 1 & \text{if there is a break on the } k\text{th run,} \\ & \text{immediately after servicing borrower } i, \\ 0 & \text{otherwise,} \end{cases}$$

$$x_{ijk} \begin{cases} 1 & \text{if the } k\text{th run proceeds directly from location } i \text{ to location } j, \\ 0 & \text{otherwise,} \end{cases}$$

y_{ik} the arrival time of the kth run at location i. (If the kth run does not visit location i, y_{ik} is defined arbitrarily.)

$$z_j^o \begin{cases} 1 & \text{if compulsory borrower } j \text{ is visited in weeks 1 and 3,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_j^e \begin{cases} 1 & \text{if compulsory borrower } j \text{ is visited in weeks 2 and 4,} \\ 0 & \text{otherwise.} \end{cases}$$

We now develop a model of the Drammen operation. The primary objective is to maximise the number of optional borrowers serviced. The secondary objective is to carry out the primary objective in the minimum feasible total elapsed time

$$\max \left(M \sum_{i=q+1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{m} x_{ijk} \right) - \sum_{k=1}^{m} (y_{1k} - s_k) \,, \tag{1}$$

subject to the following.

Each borrower can be visited at most once on any given run:

$$\sum_{i=1}^{n} x_{ijk} \leq 1 \qquad j = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{2}$$

Each optional borrower is visited at most once:

$$\sum_{i=1}^{n} \sum_{k=1}^{m} x_{ijk} \leq 1 \qquad j = q+1, q+2, \ldots, n. \tag{3}$$

Each compulsory borrower must be visited every fourteen days:

$$\left.\begin{aligned} \sum_{i=1}^{n} \sum_{k=1,3} x_{ijk} &= 2z_j^o \\ \sum_{i=1}^{n} \sum_{k=2,4} x_{ijk} &= 2z_j^e \\ z_j^o + z_j^e &= 1 \end{aligned}\right\} \qquad j = 2, 3, \ldots, q. \tag{4}$$

Each run must depart from Drammen:

$$\sum_{j=2}^{n} x_{1jk} = 1 \qquad k = 1, 2, \ldots, m. \tag{5}$$

Each run must return to Drammen:

$$\sum_{i=2}^{n} x_{i1k} = 1 \qquad k = 1, 2, \ldots, m. \tag{6}$$

If a run arrives at a borrower, it must leave that borrower:

$$\sum_{i=1}^{n} x_{ihk} = \sum_{j=1}^{n} x_{hjk} \qquad h = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{7}$$

Arrival times must account for duration, travel and break times:

$$\begin{aligned} & y_{ik} + d_i + uv_{ik} + t_{ij} + N(x_{ijk} - 1) \leq y_{jk} \\ & \qquad\qquad i = 2, 3, \ldots, n; \quad j = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, m, \\ & s_k + t_{1j} + N(x_{1jk} - 1) \leq y_{jk} \qquad j = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \end{aligned} \tag{8}$$

Each run has exactly one break:

$$\sum_{i=2}^{n} v_{ik} = 1 \qquad k = 1, 2, \ldots, m. \tag{9}$$

If the break on the kth run occurs at borrower i, then the kth run must service borrower i:

$$\sum_{j=1}^{n} x_{ijk} \geq v_{ik} \qquad i = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{10}$$

The servicing of a borrower must occur during the borrower's time window:

$$a_i \leq y_{ik} \leq b_i \qquad i = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{11}$$

There is a time limit on the duration of each run:

$$y_{1k} - s_k \leq P \qquad k = 1, 2, \ldots, m. \tag{12}$$

There is a time limit on the average duration of all runs:

$$\sum_{k=1}^{m} (y_{1k} - s_k) \leq mT. \tag{13}$$

Part of the break of each run must occur approximately half-way through the run:

$$\begin{gathered} y_{ik} + d_i \leq \frac{1}{2}(y_{1k} + s_k) + 3 + N(1 - v_{ik}), \\ \frac{1}{2}(y_{1k} + s_k) - 3 \leq y_{ik} + d_i + u + N(1 - v_{ik}), \\ i = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \end{gathered} \tag{14}$$

Each compulsory borrower must be visited at the same time of day on both of its runs:

$$\left.\begin{aligned} y_{i1} &= y_{i3} \\ y_{i2} &= y_{i4} \end{aligned}\right\} \quad i = 2, 3, \ldots, q. \tag{15}$$

And, finally, simple constraints:

$$\left.\begin{aligned} x_{ijk} &\in \{0,1\} \\ v_{ik} &\in \{0,1\} \\ z_j^o, z_j^e &\in \{0,1\} \\ y_{ik} &\in \{1,2,3,\ldots\} \\ s_k &\in \{1,2,3,\ldots\} \end{aligned}\right\} \quad \begin{aligned} i &= 1, 2, \ldots, n, \\ j &= 1, 2, \ldots, n, \\ k &= 1, 2, \ldots, m. \end{aligned} \tag{16}$$

2.3. Application of the model

We now report on the outcome when the numerical instance for the Drammen system was solved. The input data are given in Table 1, with 156 time periods of 5 minutes each defining the maximum length of any working day. Period 0 begins at 8.00 am, period 1 begins at 8.05 am, and so on. Also, $n = 22$, $q = 18$, $m = 4$, $M = 1000$, $N = 100$, $u = 8$, $P = 108$, and $T = 96$. Borrowers 9 and 15 are actually the same, except for their time windows. This means that there is one special borrower that is visited four times, alternately in the afternoons and in the mornings, week by week. This factor is reflected in the following specialised version of constraint (4):

$$\begin{aligned} \sum_{j=1}^{n} x_{9jk} &= 1 \qquad k = 1, 3, \\ \sum_{j=1}^{n} x_{15,jk} &= 1 \qquad k = 2, 4. \end{aligned} \tag{17}$$

The problem was solved to optimality using a standard IP software system — viz. XPRESS-MP [27]. The best solution achieved by XPRESS-MP was established as optimal by conducting further implicit enumeration based on the time windows and driving times associated with certain key locations. This optimal solution corresponds to the following runs, with the sequence of borrowers serviced (starting and ending at location 1) as indicated by the order of the columns, as shown in Tables 2–5. Table entries represent time period numbers when each event begins, including the break.

TABLE 1. The Input Data for the Drammen System.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
d_i	12	3	4	4	4	4	3	6	6	4	4	4	4	6	4	4	4	4	4	4	4	
a_i	6	0	48	72	84	96	108	96	108	12	12	24	12	12	12	12	12	96	0	0	48	
b_i	42	151	96	150	150	149	149	147	147	148	149	150	150	66	150	150	150	148	149	151	148	

TABLE 2. The run for the first week (Duration = 84).

Location	1	18	2	17	15	13	11	12	5	6	4	1
Arrive		21	27	41	48	57	64	78	84	90	96	103
Break							68					
Depart	19	25	39	45	54	61	76	82	88	94	100	

TABLE 3. The run for the second week (Duration = 52).

Location	1	14	3	16	10	9	8	7	1
Arrive		94	99	103	108	124	132	137	144
Break					114				
Depart	92	98	102	107	122	130	135	141	

TABLE 4. The run for the third week (Duration = 103).

Location	1	18	2	17	15	13	11	12	5	6	4	19	20	22	1
Arrive		21	27	41	48	57	64	78	84	90	96	101	109	114	122
Break							68								
Depart	19	25	39	45	54	61	76	82	88	94	100	105	113	118	

TABLE 5. The run for the fourth week (Duration = 56).

Location	1	21	14	3	16	10	9	8	7	1
Arrive		89	94	99	103	108	124	132	137	144
Break						114				
Depart	88	93	98	102	107	122	130	135	141	

The combination of the above four runs provides an objective function value of $(1000)4 - 84 - 52 - 103 - 56 = 3706$, with all four optional borrowers serviced, and a total elapsed time for all four runs of 295 time periods. These runs represent the following improvements to the Drammen system. For the first time, the optional borrowers are serviced. Indeed all four of them are serviced. Also, the total elapsed time is 17% less than the schedule that was originally in use. That original schedule was not only longer, but did not include any of the four optional borrowers. The runs shown in Tables 2–5 were accepted by the Drammen librarians, and have become the modus operandi for Drammen. We now go on to describe larger problems.

3. Operations with larger numerical instances

We now briefly describe extensions of the Drammen operation, discussed in the last section, to the routing and scheduling of the bookmobiles that serve firstly, the southern part of Buskerud County, and secondly, the whole County. As will be seen, these operation are significantly more complex than the Drammen operation, both in terms of constraints and size.

3.1. A Description of the Southern Buskerud County operation

The Southern Buskerud County operation has most of the same features and constraints as the previously mentioned operation, except for the following factors. Most of the borrowers are compulsory, and most of the cumpulsory borrowers must be serviced exactly once during each time cycle. The rest of the compulsory borrowers must be serviced exactly twice, as in the Drammen operation. But as we now have more than one run per week, this requirement must be reflected in a specific constraint in any model of the operation. However each of the borrowers that must be visited twice must not only be visited at the same time of day, as before, but also on the same day of the week (e.g., every second Thursday at 11.00 am). This last restriction is now important, as there are to be fourteen runs per four-weekly time cycle, as opposed to one per week, as in the Drammen case. There are no borrowers that must be visited more than twice. There are also a few optional borrowers, that as before, can be visited at most once.

However, the main difference from the Drammen operation comes about due to the fact that some of the borrowers are located in remote areas far from the depot, which is still the town of Drammen. This makes it necessary to incorporate an overnight stay (of a single night) at a certain hotel at a given location, into one of the runs. This means that most of the runs are completed within one working day as before, but one run is completed in exactly two working days. Certain identified borrowers must be serviced by this unique, two-day run. However, there is time available on that run to service additional borrowers, whether compulsory or optional. The question of which, if any, additional borrowers are serviced on the hotel run is part of the decision problem. Naturally, for the two-day run, the hotel departure time and all subsequent times on the run must be reset with respect to the hotel arrival time. Once again, each working day must have a break during the middle of its elapsed time, hence the two-day run has two breaks. The above imply that, in terms of TSP terminology, the Southern Buskerud County operation is a MTSPTW with additional compulsory borrowers requiring differing service levels, and also with one run being up to twice the normal allowable duration. The numerical instance corresponding to this latter operation is also far bigger than for the Drammen operation. Our formulation is as follows:

$$\max\left(M\sum_{i=q+1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}x_{ijk}\right)-\sum_{i=1,n+1}\sum_{k=1}^{m}(y_{1k}-s_{ik}). \tag{18}$$

Subject to:

Each borrower can be visited at most once on any given run:

$$\sum_{i=1}^{n+1}x_{ijk}\le 1 \qquad j=2,3,\dots,n;\quad k=1,2,\dots,m. \tag{19}$$

Each optional borrower is visited at most once:

$$\sum_{i=1}^{n+1}\sum_{k=1}^{m}x_{ijk}\le 1 \qquad j=q+1,q+2,\dots,n. \tag{20}$$

Each borrower in L_1 must be visited exactly once:

$$\sum_{i=1}^{n+1}\sum_{k=1}^{m}x_{ijk}=1 \qquad j=2,3,\dots,p. \tag{21}$$

Each compulsory borrower in L_2 must be visited twice, at the same time of the day, every fourteen days:

$$\left.\begin{aligned}&\sum_{i=1}^{n+1}\sum_{k=1}^{8}x_{ijk}=2z_j^o\\&\sum_{i=1}^{n+1}\sum_{k=9}^{14}x_{ijk}=2z_j^e\\&z_j^o+z_j^e=1\\&y_{jk}=y_{j,k+4}\quad k=1,2,3,4\\&y_{jk}=y_{j,k+3}\quad k=9,10,11\end{aligned}\right\}\qquad j=p+1,p+2,\dots,q. \tag{22}$$

Each run, except for the second day of the overnight run, must depart from the depot:

$$\begin{aligned}&\sum_{j=2}^{n}x_{1jk}=1 \qquad k=1,3,4,\dots,m,\\&\sum_{j=2}^{n}x_{n+1,j,2}=1.\end{aligned} \tag{23}$$

Each run, apart from the first day of the overnight run, must return to the depot:

$$\begin{aligned} \sum_{i=2}^{n} x_{i1k} &= 1 \qquad k = 2, 3, \ldots, m, \\ \sum_{i=2}^{n} x_{i,n+1,1} &= 1. \end{aligned} \tag{24}$$

If a run arrives at a borrower, it must leave that borrower:

$$\sum_{i=1}^{n+1} x_{ihk} = \sum_{j=1}^{n+1} x_{hjk} \qquad h = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{25}$$

Arrival times must account for duration, travel, and break times:

$$\begin{aligned} &y_{ik} + d_i + uv_{ik} + t_{ij} + N(x_{ijk} - 1) \leq y_{jk} \\ &\qquad i = 2, 3, \ldots, n; \quad j = 1, 2, \ldots, n+1; \quad k = 1, 2, \ldots, m, \\ &s_{1k} + t_{1j} + N(x_{1jk} - 1) \leq y_{jk} \qquad j = 2, 3, \ldots, n; \quad k = 1, 3, 4, \ldots, m, \\ &s_{n+1,2} + t_{n+1,j} + N(x_{n+1,j2} - 1) \leq y_{j2} \qquad j = 2, 3, \ldots, n. \end{aligned} \tag{26}$$

Each run has exactly one break:

$$\sum_{i=2}^{n} v_{ik} = 1 \qquad k = 1, 2, \ldots, m. \tag{27}$$

If the break on a run occurs at borrower i, then the run must service borrower i:

$$\sum_{j=1}^{n+1} x_{ijk} \geq v_{ik} \qquad i = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{28}$$

The servicing of a borrower must occur during the borrower's time window:

$$a_i \leq y_{ik} \leq b_i \qquad i = 2, 3, \ldots, n; \quad k = 1, 2, \ldots, m. \tag{29}$$

There is a time limit on the duration of each run:

$$\begin{aligned} y_{1k} - s_{1k} &\leq P \qquad k = 3, 4, \ldots, m, \\ y_{n+1,1} - s_{1,1} &\leq P, \\ y_{1,2} - s_{n+1,2} &\leq P. \end{aligned} \tag{30}$$

There is a time limit on the average duration of all runs:

$$(y_{n+1,1} - s_{1,1}) + (y_{1,2} - s_{n+1,2}) + \sum_{k=3}^{m} (y_{1k} - s_{1k}) \leq mT. \tag{31}$$

Part of the break of each run must occur approximately half-way through the run:

$$\left.\begin{aligned} y_{ik} + d_i &\le \frac{1}{2}(y_{1k} + s_{1k}) + 3 + N(1 - v_{ik}) \\ \frac{1}{2}(y_{1k} + s_{1k}) - 3 &\le y_{ik} + d_i + u + N(1 - v_{ik}) \end{aligned}\right\} \begin{aligned} i &= 2, 3, \ldots, n, \\ k &= 3, 4, \ldots, m, \end{aligned}$$

$$\begin{aligned} y_{i1} + d_i &\le \frac{1}{2}(y_{n+1,1} + s_{1,1}) + 3 + N(1 - v_{i1}), \\ \frac{1}{2}(y_{n+1,1} + s_{1,1}) - 3 &\le y_{i1} + d_i + u + N(1 - v_{i1}), \\ y_{i2} + d_i &\le \frac{1}{2}(y_{1,2} + s_{n+1,2}) + 3 + N(1 - v_{i1}), \\ \frac{1}{2}(y_{1,2} + s_{n+1,2}) - 3 &\le y_{i2} + d_i + u + N(1 - v_{i1}). \end{aligned} \tag{32}$$

Each borrower in DH must be on the overnight run:

$$\sum_{i=1}^{n} x_{ij1} + \sum_{i=2}^{n+1} x_{ij2} = 1 \qquad j \in DH. \tag{33}$$

The hotel departure time is fixed:

$$s_{n+1,2} = A. \tag{34}$$

And, finally, the simple conditions:

$$\left.\begin{aligned} x_{ijk} &\in \{0, 1\} \\ v_{ik} &\in \{0, 1\} \\ z_j^o, z_j^e &\in \{0, 1\} \\ y_{ik} &\in \{1, 2, 3, \ldots\} \\ s_{ik} &\in \{1, 2, 3, \ldots\} \end{aligned}\right\} \qquad \begin{aligned} i &= 1, 2, \ldots, n + 1, \\ j &= 1, 2, \ldots, n + 1, \\ k &= 1, 2, \ldots, m. \end{aligned} \tag{35}$$

3.2. A Description of the County-wide operation

The County-wide operation differs from the previous case in the following ways. The operation is based at two towns, Drammen and Gol, that serve as depots — i.e., each bookmobile is based at exactly one of the towns, in the sense that it begins and ends all of its runs at that town. The number of runs based in each town is given. However, the question of which depot will service each borrower is part of the decision problem. Further, in addition to the known hotel with its dedicated borrowers serviced out of Drammen, there are to be four overnight runs based in Gol. There are a number of available hotels, from which four must be chosen. Also, finally, the two bookmobiles from the two depots must meet once during a month. We are free to choose the place and time. During this encounter they exchange library material.

4. Conclusions and suggested directions for further research

There do not seem to be any models or solution techniques for bookmobile routing and scheduling reported in the open literature. This version of the TSP is complicated by a unique combination of factors. We have reported the outcome of a successful implementation of a standard integer programming model for a small, practical scenario. However, it is likely that more effective techniques will have to be employed to produce useful schedules for instances of the dimensions of the Southern Buskerud and County-wide scenarios. To this end, the authors are currently investigating the application of the learning meta-heuristic tabu search as a solution technique for the extended problems that are discussed in Section 3.

References

[1] N. Ascheuer, M. Fischetti and M. Grotschel, *Solving the asymmetric travelling salesman problem with time windows by branch-and-cut.* Mathematical Programming **90** (2001) 475–506.

[2] E. Baker, *An exact algorithm for the time constrained travelling salesman problem.* Operations Research **31** (1983) 938–945.

[3] T. Bektas, *The multiple traveling salesman problem: An overview of formulations and solution procedures.* OMEGA **34** (2006), 209–219.

[4] N.L. Biggs, E.K. LLoyd, and R.J. Wilson, *Graph Theory 1736-1936.* Clarendon Press, Oxford, 1976.

[5] M.J. Brusco, and L.J. Jacobs, *Optimal models for meal-break and start-time flexibility in continuous tour scheduling.* Management Science **46** (2000), 1630–1641.

[6] R.E. Burkard, *Travelling salesman and assignment problems: A survey.* In: Discrete Optimization 1 (P. L. Hammer, E. L. Johnson, and B. H. Korte, eds.), Annals of Discrete Mathematics **4**, North-Holland, Amsterdam (1979), 193–215.

[7] W.B. Carlton and J.W. Barnes, *Solving the travelling- salesman problem with time windows using tabu search.* IEE Transactions **28** (1996), 617–629.

[8] A.E. Cartera and C.T. Ragsdale, *A new approach to solving the multiple traveling salesperson problem using genetic algorithms.* European Journal of Operational Research **175** (2006), 246–257.

[9] N. Chandran, T.T. Narendran, and K. Ganesh, *A clustering approach to solve the multiple travelling salesmen problem.* International Journal of Industrial and Systems Engineering **1** (2006), 1–20.

[10] G.B. Dantzig, R. Fulkerson, and S.M. Johnson, *Solution of a large-scale traveling salesman problem.* Operations Research **2** (1954), 393–410.

[11] Y. Dumas, J. Desrosiers, and E. Gelinas, *An optimal algorithm for the travelling salesman problem with time windows.* Operations Research **43** (1995), 367–371.

[12] F. Focacci, A. Lodi, and M. Milano, *A hybrid exact algorithm for the TSPTW.* INFORMS Journal on Computing **14** (2002), 403–17.

[13] G. Gutin and A.P. Punnen, *The Traveling Salesman Problem and Its Variations.* Springer-Verlag, Berlin, 2006.

[14] M. Held and R. M. Karp, *The traveling-salesman problem and minimum spanning trees.* Operations Research **18** (1970), 1138–1162.

[15] A. Larsen, O. B. G. Madsen, and M. M. Solomon, *The A-priori Dynamic Traveling Salesman Problem with Time Windows.* Transportation Science **38** (2004), 459–472.

[16] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy-Khan, and D. B. Shmoys, *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization.* John Wiley & Sons, New York, 1985.

[17] J. D. C. Little, K. G. Murty, D. W. Sweeney, and C. Karel, *An algorithm for the traveling salesman problem*, Operations Research **11** (1963), 972–989.

[18] K. Menger, *Das Botenproblem.* In Ergebnisse eines Mathematischen Kolloquiums 2 (K. Menger, editor), Teubner, Leipzig (1932), 11–12.

[19] S. Mitrovic-Minic and R. Krishnamurti, *The multiple TSP with time windows: vehicle bounds based on precedence graphs.* Operations Research Letters **34** (2006), 111–120.

[20] H. D. Nguyen, I. Y. K. Yamamori, and M. Yasunaga, *Implementation of an Effective Hybrid GA for Large-Scale Traveling Salesman Problems.* IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics **37** (2007), 92–99.

[21] J. W. Ohlmann and B. W. Thomas, *A Compressed-Annealing Heuristic for the Traveling Salesman Problem with Time Windows.* INFORMS Journal on Computing **19** (2007), 80-90.

[22] G. Pessant, M. Gendreau, J.-Y. Potvin, and J.-M. Rousseau, *An exact constraint logic programming algorithm for the travelling salesman problem with time windows.* INFORMS Journal on Computing **19** (1998), 12–29.

[23] D. J. Rosenkrantz, R. E. Stearns, and P. M. Lewis II *An Analysis of Several Heuristics for the Traveling Salesman Problem.* SIAM Journal of Computing **6** (1977), 563-581.

[24] B. Tolga, *The multiple traveling salesman problem: an overview of formulations and solution procedures.* OMEGA **34** (2006), 209–219.

[25] H-K. Tsai, J-M. Yang, Y-F. Tsai, and C-Y. Kao, *An evolutionary algorithm for large traveling salesman problems.* IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics 34 (2004) 1718–1729.

[26] R. Wolfler-Calvo, *A new heuristic for the travelling salesman problem with time windows.* Transportation Science **34** (2000) 113–124.

[27] *XPRESS-MP.* Dash Optimization, Blisworth NN73BX, UK.

Les R. Foulds
Universidade Federal de Goiás
Campus II Samambaia
Goiania/GO 74000-001
Brazil
e-mail: lfoulds@waikato.ac.nz

Stein W. Wallace
Molde University College,
P.O. Box 2110
NO-6402 Molde
Norway
e-mail: Stein.W.Wallace@himolde.no

John Wilson
University of Loughborough,
Great Britain

Martin West
School Information Systems
Curtin Business School
Curtin University of Technology
GPO Box U1987
Perth, WA 6845
Australia
e-mail: martin.west@cbs.curtin.edu.au

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 321–342

Instability and Sustained Oscillations in Neo-Classical Growth Models with Unemployment

Luciano Fanti and Piero Manfredi

Abstract. The paper deals with a long-standing problem in the debate on economic growth — viz. the issue of stability of the balanced growth path in macroeconomic growth models. We first consider an early claim that the replacement of a fixed coefficients technology by a neoclassical production function provides the solution to the Harrod–Domar knife-edge instability problem, and proceed to investigate the stability and onset of oscillations in an "augmented" neoclassical model of macroeconomic growth. The model embeds a Constant Elasticity of Substitution (CES) production function, sluggishly adjusting and non-market-clearing real wages, and endogenous fertility. The analysis shows that (i) Solow models may suffer instability; (ii) a spark-triggering instability may be due to the presence of a too strong "neoclassicity" in production; and (iii) strong "neoclassicity" may lead to sustained oscillations of the economy and also to knife-edge instability.

Mathematics Subject Classification (2000). JEL classification codes E3, J0.

Keywords. Neoclassical growth, Neoclassical production, Unemployment, Stability and oscillations, Balanced growth path.

1. Introduction

The first epoch of modern economic growth theory was dominated by the issue of the intrinsic instability of the equilibrium growth path of the then dominant Harrod–Domar model. The corresponding "knife-edge" problem was elegantly solved by the advent of the neoclassical growth model by Solow [17]. By simply postulating a neoclassical production function instead of the fixed coefficients technology used by Harrod and Domar, he proved the (global) asymptotic stability of the equilibrium growth path of the economy, the so-called balanced growth path. Solow [20] observed that the knife-edge problem stems from the assumption that production takes place under fixed proportions (i.e., without substitution between labour and capital in production); but that if this assumption is replaced by neoclassical production, "...the knife-edge notion of unstable balance seems to go with it" [17, p. 65]. This result is a fundamental achievement in economic growth theory.

A more general question is whether neoclassical production theory is definitely the "panacea" ensuring the stability of the equilibrium of a growing economy.[1] The usual answer given in subsequent literature seems to have been positive. Moreover, the stabilising role played by neoclassical production seems valid for other growth models not based on neoclassical production schemes — viz. Harrod–Domar–type models, which do not feature cyclical growth, and growth cycle models (e.g., Goodwin [8]). Indeed, Van der Ploeg [22] has shown that, by enriching Goodwin's model with the hypothesis of profit maximisation according to a constant return to scale Constant Elasticity of Substitution (CES) production function, the Lotka–Volterra–Goodwin perpetual cycles are replaced either by damped oscillations or by monotonic convergence to the balanced growth state. Consequently, contrary to fixed coefficients production schemes when the economy enters a phase unfavourable to firms where workers can claim higher wages (thereby reducing profits and investments), neoclassical firms can replace labour with capital and so maintain their profitability in a smoothly approached balanced growth state. Thus the introduction of the neoclassical production function removes the conservative Lotka–Volterra oscillations of the Goodwin system, and renders asymptotically stable trajectories (Van der Ploeg [22, p. 228]).

A second fundamental result obtained by Van der Ploeg [22] is that the presence of damped oscillations, instead of a monotonic convergence towards the balanced growth path, is the consequence of a small factor substitution – thus "A small elasticity of substitution...is more likely to lead to cycles than to monotonic convergence to the balanced growth trajectories" (p. 229). A higher degree of factor substitution (i.e., a stronger degree of "neo-classicity" in production) implies stronger stability of the balanced growth equilibrium. In brief, the presence of a neoclassical production function seems sufficient to ensure the stability of both aggregate descriptive growth models and growth cycle models.

In this paper, we reconsider the central issue of stability of the balanced growth path from a broader departure point. We take a highly flexible neoclassical production function (a CES function) as the core of a general neoclassical Solow-type model. In addition to traditional Solow features, our model involves three additional assumptions — viz. (i) the wage earners do not save; (ii) structural unemployment may persist and wages adjust sluggishly in the labour market; and (iii) the fertility of individuals is endogenous and heterogeneous between population subgroups, depending on their employment status.

[1] The term "neoclassical production theory" used here means a theory having production as one building block, based on optimising firms with a production function showing a certain degree of factor substitution. This definition is consistent with the notion of Solow et al. [22] and others that the term "neoclassical" means an approach where, besides the assumption of Say's Law, capital and labour are directly and smoothly substitutable for one another.

The stability of the unique balanced growth equilibrium of our model is investigated by using as the control ("bifurcation") parameter the parameter tuning in the production function — i.e., the degree of factor substitution that is a measure of the degree of "neo-classicity" in production. Our results, regarding the relationship between the presence of a neoclassical production function and the stability of the steady state of balanced growth, differ substantially from those in the literature. There are economically plausible situations where a higher factor substitution can be destabilising, while a technology tending to a strong complementarity between factors (the old Harrod–Domar–Goodwin world) can favour stability. Whenever instability occurs, steady oscillations arise through Hopf bifurcations of the balanced growth path, and extensive simulation suggests that the ensuing oscillations are at least locally stable. Thus the model provides plausible avenues to (stable) steady oscillations around the balanced growth trend. This was considered a central issue by classical and neo-Keynesian economists — Kaldor [23] lists "cycles in the growth" as one of his stylised facts of economic growth — but it is surprisingly overlooked in the modern growth agenda [2]. Our model provides a unified endogenous explanation of this stylised fact, in shedding new light on the relation between the stability of growing economies and the assumption of neoclassical production theory, important for stabilisation policies.

The paper is organised as follows. Section 2 introduces the model. Section 3 analyses the existence and stability of the balanced growth equilibrium, its bifurcation into steady cycles, and how these features are affected by the elasticity of substitution of the neoclassical production function. Section 4 is devoted to numerical illustrations, and concluding remarks follow.

2. The model of the economy

Solow's original model [2, 17] assumes perfect competition in the labour market, and consequently full employment at any time (or at most an exogenously given constant unemployment). Moreover, the rate of growth of the labour force, the main engine of growth in the model, was assumed to be fully exogenous. As previously indicated, in our model, we allow for persistent unemployment and endogenous dynamics of the labour force, by endogeneising the population fertility.[2] The amendments we make were proposed by Solow himself [17, 18]. In particular, on the issue of unemployment Solow [17, 28, p. xviii] acknowledges the need to take into account the stylised fact of wage stickiness: "this is the sort of amendment that I mentioned in 1956, but did not pursue very far." In such circumstances, "the new equilibrium path will depend on the amount of capital accumulation that has taken place during the period of disequilibrium, and probably also on the amount of

[2] Models with unemployment and endogenous fertility are investigated in Fanti and Manfredi [4], focusing on the role of unemployment benefits; and in Manfredi and Fanti [14, 15], focusing on age structure. Endogenous fertility may enrich the standard neoclassical growth model with full employment, as shown in Fanti and Manfredi [4, 5].

unemployment, especially long-term unemployment, that has been experienced." (ibidem, p. xviii). A parsimonious way to model this extension in a growth framework is to consider a sluggishly-adjusting labour market, for instance according to a Phillips real wage equation, without necessarily specifying the underlying model of real wage determination. Growth scholars know that the Phillips equation is the fundamental ingredient of another cornerstone economic growth model — viz. the non-hortodox model of Goodwin, which is the famous Lotka–Volterra representation of Marxian class conflict between capitalists and workers [8]. It is notable that the two most famous descriptive growth models (due to Solow and Goodwin) have distinctively different assumptions — viz. whether or not the production is neoclassical, and whether or not the wage is flexible to instantaneously clear the labour market. In the Goodwin model, the impossibility of a factor substitution (however small) is a special restrictive assumption. In the Solow model, the impossibility of stickiness in wage adjustments is special and restrictive. Our model blends some features of these two models, and the main differences are discussed further below in more detail.

In the Solow model, the accumulation rate is independent of the distribution, and therefore there is the drawback that the investment would be the same for any level of the profit.[3] On the other hand, the Goodwin model assumes that profits alone finance investments, in that wage earners do not save. This assumption is less heroic than it might seem at first glance, and certainly more realistic than the corresponding Solow hypothesis — cf. empirical and theoretical research on the determinants of investment by firms [7]. Solow himself [18] defends Goodwin's choice against his own, noting one of Kaldor's [23] "stylized facts" of growth — viz. that "economies with a high share of profits in income tend to have a high ratio of investment to output" (p. 3), and that Goodwin's assumption of equality between profit share and investment/output ratio is just a representation of this stylized fact.

The other major feature distinguishing the Goodwin from the Solow model is the assumption of sluggish wage adjustment, which leads to the establishment of some "natural" rate of unemployment. This feature is certainly more realistic of the continuous instantaneous adjustment in the labour market assumed by Solow. Finally, in the Goodwin model — as in the Harrod–Domar model — the production is not neoclassical. The main dynamical consequence is that the Goodwin model displays a steady-state cyclical growth rather than the monotonic growth of the neoclassical Solow model, and unfortunately suffers the well-known problem of "structural instability" of Lotka–Volterra conservative cycles. In passing,

[3] Solow [18] explicitly explores situations in which the saving rate is variable. He postulates a "fixed saving ratio from wage and profit income, a larger one from profits than from wages" (p.29), and furthermore argues that in any case "any theory of saving that makes the saving rate depend only on the variables of the model — the capital/output ratio, the labour/capital ratio, the return of capital — can be handled in the same way" (p. 29). Thus the assumption of saving-profits equality does not modify the properties of his original model.

we note however that structural stability may be obtained in suitable variants of the original Goodwin model (e.g., Fanti-Manfredie [3]).

Our model is a Solow–Goodwin synthesis for a closed “real”economy with rational individuals, maximising firms, and a labour market governed by a real wage bargaining system represented by a linear Phillips curve. The dynamics of this economy arise from the rate of accumulation, from wage bargaining, and from population dynamics. Let us now discuss the major features of our model.

2.1. Firms

Firms seek to maximise profit in a competitive market. Technology is represented by the Constant Elasticity of Substitution (CES) production function with constant returns to scale:

$$Y = c\left[zK^{-\theta} + (1-z)L_E^{-\theta}\right]^{-\frac{1}{\theta}}, \quad 0 < z < 1, \qquad -1 < \theta < \infty\,. \tag{1}$$

The function $\eta = 1/\left(1+\theta\right)$ is known as the elasticity of substitution, and z is a distribution parameter that becomes a distributive share for $\theta \to 0$. Equation (1) can represent any possible elasticity of substitution, and includes both the Cobb–Douglas ($\theta \to 0$) and the Leontief fixed-coefficients ($\theta \to +\infty$) as special cases. The labour input is measured in efficiency units: $L_E\left(t\right) = L\left(t\right)\beta(t)$,where $L\left(t\right)$ denotes physical labour, and $\beta(t)$ the stock of knowledge or the labour augmenting technical progress. Firms hire labour until the productivity of the marginal worker equals the real wage. Thus the optimal factor demand ratio, expressed in terms of the distributive workers' share of the national income (V), follows from (1) after some manipulations — i.e.,

$$\frac{K}{L} = \phi(V)\beta\,, \tag{2}$$

where

$$\phi(V) = \left[\frac{(1-z)(1-V)}{zV}\right]. \tag{3}$$

Under the neo-Keynesian growth theory assumption that the saving-investment equality holds and assuming that the wage earners do not save, the profits-investment equality follows — i.e., profits (P) are reinvested by firms according to a fraction s$_p$ ($I = s_pP$). The rate of accumulation $\dot{K}/K$, where the dot denotes time differentiation, can then be expressed as a function of the distributive share of profit $1-V$ — i.e.,

$$\frac{\dot{K}}{K} = s_p c z^{\frac{-1}{\theta}}(1-V)^{\frac{1+\theta}{\theta},} \tag{4}$$

such that firms finance their investments by their profit income. As mentioned previously, this is consistent with empirical evidence, as well as neo-Keynesian investment theory [1].

2.2. Employment status and individual fertility behaviour

We consider two sub-population types, the employed and unemployed. We follow a modern approach where family choice determines the birth rate per unit time [9, 16], rather than the lifetime stock of children, as in basic overlapping generation models. Thus every time each individual family determines the crude birth rate b, trading off between consumption c and children. The income of the employed individual is the wage earned per unit time (w), while the income of the unemployed individual is the unemployment benefit, assumed to be a constant fraction h (the so-called replacement ratio) of the wage. We assume that rearing children is more expensive for employed individuals because of the opportunity-cost of the wage, during the time of child care (e.g., due to a fraction of the wage being paid to a nurse). Cost may include both direct costs (e.g., clothes) and indirect costs (e.g., nursing), and it is assumed to be an exogenous constant fraction of the income actually received. We denote such fractions as q ($q \leq 1$) for the worker and q_u ($q_u \leq 1$) for the unemployed individual, with $q > q_u$ consistent with our assumption on the opportunity cost of the children. The quantities qw and $q_u hw$ therefore define the real cost per child, for the workers and for the unemployed individuals, respectively.

At any time t, the representative employed individual maximises utility — i.e.,

$$\max_{c_t, b_t} U(c_t, b_t) \tag{5}$$

where the utility function U is well-behaved, subject to the income constraint

$$c_t + q w_t b_t \leq w_t \ . \tag{6}$$

Assuming that preferences are represented by a log-linear utility function (for simplicity we suppress suffix t),

$$U(c, b) = c^a b^{1-a}, \qquad 0 < a < 1 \ , \tag{7}$$

whence the demand for children by employed individuals is

$$b_e = \frac{(1-a)}{q} \ . \tag{8}$$

Similarly, the demand for children by unemployed individuals is

$$b_u = \frac{(1-a)h}{q_u} \ . \tag{9}$$

The quantities b_e, b_u denote the fertility rate of the two sub-populations of employed and unemployed individuals.[4] Since a fraction E of individuals is employed

[4] Such static optimisations can be derived from a fully dynamic optimisation problem by assuming that: (i) the utility depends on the flow of births rather than on the stock of children; and (ii) each individual faces a probability of death which depends on aggregate per capita consumption, which individuals take as given [9].

(and $1 - E$ unemployed) at any time t, where E is the rate of employment, the overall fertility rate is

$$b = b_e E + b_u (1 - E)) = (1 - a)\left(\frac{h(1 - E)}{q_u} + \frac{E}{q}\right) = (1 - a)\frac{qh(1 - E) + q_u E}{qq_u} . \tag{10}$$

Equation (10) represents employment status as a source of heterogeneity in fertility behaviour, consistent with an empirically documented significant correlation between the unemployment rate (the presence of structural unemployment) and fertility, which appears to have been a fundamental stylised fact of European history [1, 20]. Despite this evidence, with a few exceptions [4, 14] the dynamic interaction between fertility and employment status has not been included in economic macro-models, and the present paper addresses this issue as well.

2.3. Labour market

The wage dynamics factor is represented by a simple linear "real wage" Phillips equation, as in Goodwin [8] — viz.

$$\dot{w} = w(-\gamma + \rho E) , \qquad 0 < \gamma \leq \rho \tag{11}$$

where $E = L/N$ is the employment rate (the ratio of the total labour actually employed L to the total labour supply N) and γ, ρ are characteristic labour market parameters. Equation (11) states that when unemployment decreases, workers become more "powerful" and claim higher real wages, (and vice-versa), and is consistent with different economic viewpoints. Thus it may arise not only from a unionised labour market with a relative bargaining power depending on the state of unemployment or from a "Marxian" labour market with industrial reserve army, but also from a purely neoclassical labour market obeying a Marshallian [10, 11] or a Walrasian adjustment process with persistent excess demand [6], in the presence of a "natural rate of unemployment" and voluntary unemployment. The concept of real-wage Phillips curve is supported by recent empirical work [21].

Since the wage dynamics factor affects the distributive share of labour, from equation (2) it also affects the optimal factor demand ratio. Thus when workers are able to obtain a larger share of national income, firms find it less profitable to hire workers, and therefore switch away from labour to machinery.

2.4. Our model of the economy

The previous economic relationships lead to a two-dimensional dynamical system involving the employment rate (E) and the labour share (V). Taking growth rates in equation (2) $K/L = \phi(V)$, we have

$$\frac{\dot{L}}{L} = \frac{\dot{K}}{K} - \frac{\dot{\phi}}{\phi} - \frac{\dot{\beta}}{\beta} . \tag{12}$$

From (3) we have

$$\frac{\dot{\phi}}{\phi} = -\frac{1}{\theta}\left[\frac{\dot{(1 - V)}}{(1 - V)} - \frac{\dot{V}}{V}\right] = \frac{1}{\theta(1 - V)}\frac{\dot{V}}{V} , \tag{13}$$

whence (12) becomes

$$\frac{\dot{L}}{L} = s_p(1-V)^{\frac{1+\theta}{\theta}} c z^{\frac{-1}{\theta}} - \frac{1}{\theta(1-V)}\frac{\dot{V}}{V} - \frac{\dot{\beta}}{\beta} \,. \tag{14}$$

The dynamic equation for the employment ratio E then follows from the identity

$$\dot{E}/E = \dot{L}/L - n \tag{15}$$

where $n = \dot{N}/N$ is the growth rate of the labour supply.

In this paper, we disregard participation and assume that n is defined by its sole demographic component, the difference between the fertility rate b defined in (10) and the mortality rate μ, taken to be constant — thus

$$n = b - \mu = \frac{(1-a)}{qq_u}\left[qh(1-E) + q_u E\right] - \mu = n\left(E\right) \ , \tag{16}$$

and hence from (14)–(15),

$$\frac{\dot{E}}{E} = s_p(1-V)^{\frac{1+\theta}{\theta}} c z^{\frac{-1}{\theta}} - \frac{1}{\theta(1-V)}\frac{\dot{V}}{V} - n - \frac{\dot{\beta}}{\beta} \,. \tag{17}$$

Finally, since $V = w/A$ where $A = Y/L$, after a further time differentiation we obtain the equation for the wage share dynamics — viz.,

$$\frac{\dot{V}}{V} = \frac{\dot{w}}{w} - \frac{\dot{A}}{A} = \frac{\dot{w}}{w} - \frac{1}{\theta}\frac{\dot{V}}{V} - \frac{\dot{\beta}}{\beta} \Rightarrow \frac{\dot{V}}{V} = \frac{\theta}{1+\theta}\left[\frac{\dot{w}}{w} - \frac{\dot{\beta}}{\beta}\right] \,. \tag{18}$$

Assuming for simplicity $\dot{\beta}/\beta = a_0$ (i.e., an exogenous constant productivity growth rate), the economy is therefore described by the following two-dimensional model in terms of the employment rate E and the share of labour V:

$$\begin{aligned} \frac{\dot{V}}{V} &= \frac{\theta}{1+\theta}\left[-\gamma + \rho E - a_0\right], \\ \frac{\dot{E}}{E} &= s_p(1-V)^{\frac{1+\theta}{\theta}} c z^{\frac{-1}{\theta}} - a_0 - \frac{1}{\theta(1-V)}\frac{\dot{V}}{V} - \frac{(1-a)}{qq_u}\left[qh(1-E) + q_u E\right] + \mu \,. \end{aligned} \tag{19}$$

This model encompasses most descriptive growth models as its special cases, and allows for an endogenous determination of income, population growth, employment and distribution. The last feature is not shared by models based on the Cobb–Douglas production function, since the distribution is then determined by the assumed technology.

3. Properties of our model: equilibria, stability and oscillations

Let us now proceed to investigate system (19). Since our focus is on the role played by neoclassical substitution in the stability of the balanced growth equilibrium, and perhaps the appearance of oscillations, we are especially concerned with the role played by ϑ in determining the existence and local stability (or instabilty) of an admissibile positive equilibrium. Preliminary investigations show that, provided $0 < V_0 < 1$, the system (19) always admits a meaningful unique positive solution — i.e., assuming it is positive initially, the solution always stays positive.

3.1. Equilibria

System (19) not only admits the "zero " equilibrium $P_0 = (0, 0)$,[5] but also an axis equilibrium $P_2 = (0, E_2)$ that have interesting economic properties, particularly from a welfare perspective [4, 14]. However, since we are mainly interested in states of balanced growth, we restrict our search to strictly positive steady states. It may be shown that at most one strictly positive equilibrium $P_1 = (V_1, E_1)$ exists, when the equilibrium level of employment is $E_1 = (a_0 + \gamma)/\rho$ which is meaningful for $a_0 + \gamma \leq \rho$. Let

$$n_1 = \frac{(1-a)}{qq_u}\left[qh(1-E_1) + q_u E_1\right] - \mu \tag{20}$$

be the corresponding rate of growth of the population. Setting $f_1(V) = (1-V)^{\frac{1+\theta}{\theta}}$, from (19) meaningful equilibrium values of the wage share are solutions V^* $(0 < V^* < 1)$ of the equation

$$f_1(V) - Gz^{\frac{1}{\theta}} = 0 \tag{21}$$

where

$$G = \frac{n_1 + a_0}{cs_p}\ . \tag{22}$$

For $\vartheta > 0$, $f_1(V)$ is monotonically decreasing and convex in the admissible set $0 < V < 1$, whereas for $-1 \leq \vartheta \leq 0$ it is monotonically increasing and convex on $0 < V < 1$. Let $\Theta = (-1, +\infty)$ be the set of admissible ϑ values. The role played by ϑ on existence and admissibility of the positive equilibrium is summarised by the following result.

PROPOSITION 1.
(A) For $G > 1$ (and therefore $G > z$), the system admits a unique positive and admissible equilibrium P_1 in the set $\Omega_1 = \{-1 < \vartheta < \vartheta_{**}\}$, where $\vartheta_{**} > 0$.
(B) For $z < G \leq 1$, a unique equilibrium that is always admissible exists for all ϑ.
(C) For $0 < G < z$, the system admits a unique positive and admissible equilibrium in the set $\Omega_2 = \{\vartheta_* < \vartheta < +\infty\}$, where $\vartheta_* < 0$. In particular, the equilibrium value of the wage share is

$$V_1 = 1 - \left[\frac{n_1 + a_0}{cs_p}\right]^{\frac{\theta}{1+\theta}} z^{\frac{1}{1+\theta}}. \tag{23}$$

The proof is given in the Appendix. Note that (23) follows straightforwardly from (21) and (22).

Proposition 1 summarises how different degrees of neoclassical substitution, as measured by ϑ, affect the existence and admissibility of a positive equilibrium of the economy — a central issue of debate in growth theory. Thus when $G/z > 1$

[5] Correctly speaking, since $V > 0$, the point $E_0 = (0, 0)$ may be considered an "extended" equilibrium point.

the balanced growth state exists in intervals of the form $-1 < \theta < \theta_{**}$ — i.e., when $G/z > 1$ the balanced growth state always exists under "neoclassical" technologies, and is only lost when the technology moves towards strong complementarity between factors. Consequently, the balanced growth state exists in a neoclassical world but does not exist in a Harrod–Domar world. When $G/z < 1$, the balanced growth state exists in intervals of the form $\theta_{*} < \theta < +\infty$, with $\theta_{*} < 0$ — i.e., it exists in the presence of complementarity, but may be lost when the neoclassical substitution is too strong $(-1 < \theta < \theta_{*})$. Thus $G < z$ implies $scz > a_0 + n_1$, when the economy saves and invests enough to overcome the expansion of the labour force and the productivity, leading to endogenous growth. Factor substitution is therefore the engine that allows endogenous growth, a result well known in the growth literature [2].

To sum up, the "beneficial" role that Solow paid to neoclassical substitution appears to largely be confirmed by the existence of the growth steady state, according to our analysis. Indeed, in one case $(G/z > 1)$ an increasing degree of neoclassical substitution allows the onset of the growth steady state in circumstances, where it did not exist in a Harrod–Domar's world; while in the other case $(G/z < 1)$, the steady growth state can actually be lost due to too strong a neoclassical substitution, but the economy nevertheless enters an even more "happy"state – viz. endogenous growth. Our discussion has shown that the existence of the state of equilibrium growth depends, for any degree of neoclassical substitution ϑ, on the ratio G/z. As will be shown later, the same parameter also critically tunes the local stability, and the following remark describes the effects due to ϑ on the equilibrium wage share V_1 (we omit the easy proof).

REMARK 1.
The function $V_1(\vartheta)$ is strictly decreasing for $G > z$ and strictly increasing for $G < z$.

3.2. Stability and bifurcation of the balanced growth path: the role of neoclassical substitution

Following some easy computations, the local stability analysis of the balanced growth path P_1 leads to the following expressions for the trace and determinant of the Jacobian J_1 at P_1 (for ease of notation we suppress the suffix 1 in the variables E, V, n):

$$\begin{aligned} Tr\,(J_1) &= -\left(\frac{1}{1+\theta}\frac{\rho}{(1-V)} + \frac{\partial n}{\partial E}\right)E\,, && (24)\\ Det\,(J_1) &= (-1)\,cs_p z^{\frac{-1}{\theta}}\frac{\vartheta}{1+\vartheta}\rho f_1'(V)\,EV, && (25)\end{aligned}$$

It is easy to check that $Det\,(J_1) > 0$ independently of the sign of ϑ, so possible switches of the balanced growth state from stability to instability are only governed by $Tr\,(J_1)$. The condition of local stability of P_1 $(Tr\,(J_1) < 0)$, yields

$$\frac{1}{1+\theta}\frac{\rho}{(1-V(\vartheta))}-\delta>0 \tag{26}$$

where

$$\delta=\frac{1-a}{qq_u}(qh-q_u)\ , \tag{27}$$

and we have written $V=V(\vartheta)$ to stress our interest in a one-parameter discussion focusing on the role played by ϑ. Since the quantity

$$H(\vartheta)=(1+\theta)(1-V_1(\vartheta)) \tag{28}$$

is non-negative in the whole admissible set of ϑ ($\vartheta\geq -1$), the positive equilibrium is always stable for $\delta<0$, so only the case $\delta>0$ corresponding to rather large unemployment benefits is of interest. (Fanti and Manfredi [4] study in detail the dynamical role of unemployment benefits.) When $\delta>0$, the fertility rate of those who are unemployed exceeds that of those employed. This case is far from being trivial, for it can be the consequence of economic environments where the rearing costs of children are higher for employed individuals and, thanks to an achieved high level of wellbeing or social protection, the society can afford high rates of unemployment benefit. Thus a growing structural unemployment may cause the unemployed individuals' fertility to grow excessively, which has a destabilising effect (see later). In this case, condition (26) may be rewritten as

$$H(\vartheta)<\frac{\rho}{\delta}\ , \tag{29}$$

which leads to the following proposition (proof given in the Appendix).

PROPOSITION 2.
(A) For $G>1$ (i.e., $G>z$), the growth steady state P_1 exists in the domain $\Omega_1=\{-1<\vartheta<\vartheta_{**}\}$, and is always locally asymptotically stable (LAS) when the degree of neoclassical substitution is maximal ($\vartheta\to -1^+$); and stability may be lost only when the degree of substitution is decreased, but not necessarily.
(B) For $0<G<1$ and $G>z$, the state of balanced growth exists in the domain $\Omega_2=\{\vartheta_*<\vartheta<+\infty\}$ and always becomes unstable for sufficiently large ϑ.
(C) For $0<G<1$ and $z>G$, the state of balanced growth exists in the domain $\Omega_2=\{\vartheta_*<\vartheta<+\infty\}$. In this event, three sub-cases are possible — i.e., the balanced growth state is:

(C1) always unstable; or
(C2) unstable for either relatively high (i.e., in a region $\vartheta_*<\vartheta<\vartheta_{H_1}$) or relatively low (i.e., in a region $\vartheta>\vartheta_{H_2}$ degrees of neoclassical substitution, and LAS in between; or
(C3) LAS for (relatively) high (i.e., close to ϑ_*) degrees of neoclassical substitution, and unstable for large ϑ.

Figures 1 and 2 illustrate Proposition 2 by showing the regions of ϑ where stability or instability respectively prevails (Figure 1 shows case (A), whereas Figure 2 illustrates case (C)).

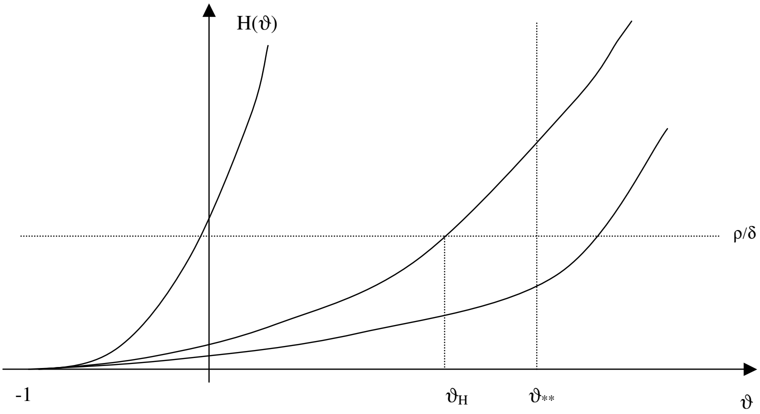


FIGURE 1. Regions of stability and instability of the balanced growth path P_1 in case A of Proposition 2 ($G > 1 > z$).

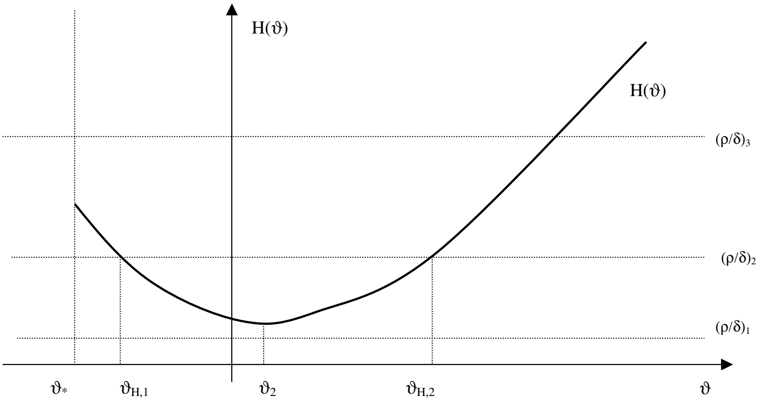


FIGURE 2. Regions of stability and instability of the balanced growth path P_1 in case C of Proposition 2 ($0 < G < z$).

It is notable how the stability of the balanced growth path is affected by the degree of neoclassical substitution, tuned by ϑ. For $G > z$, as in cases (A) and B) of Proposition 2, the balanced growth path may lose stability only by a decreased degree of neoclassical substitution (or by moving towards the Harrod–Domar world). Thus the stability may be reinforced by augmenting the degree of neoclassical substitution. However, case (B) differs from (A) in that too high a degree of neoclassical substitution (i.e., a too low ϑ) may prevent the existence of the balanced growth path (cf. Proposition 1). On the other hand, for $0 < G < z$ (case C) this is not necessarily so. In particular, there are cases where increasing the degree of neoclassical substitution leads to instability, which has never been pointed out before in the neoclassical growth literature. Thus high degrees of neoclassical substitution may trigger instability in circumstances which, although dependent on the model parameters in a complex way, necessarily require: (i) a sufficiently high weight of capital in CES technology (a high z); (ii) sufficiently low population and productivity growth rates; (iii) a high level of the technology index (c); and (iv) a high propensity to save by the capitalists (s_p).

Our results demonstrate the importance of stability of the balanced growth equilibrium, which is most relevant for both the Harrod–Domar and Solow model viewpoints. If the growth equilibrium is stable, at least locally, Solow's claim that neoclassical technology (and therefore a variable capital-output ratio) solves Harrod–Domar's knife-edge is obviously correct. Unfortunately, we have shown that this does not need to be the case. When a Solow-type economy is perturbed by sticky wage adjustments and therefore by structural unemployment, which in turn may feed back onto fertility, there are cases when the neoclassical model may suffer instability. In some of these cases (cf. again Figure 2), too high an elasticity of substitution may cause instability. The underlying mechanism is that a very large elasticity of substitution quickly increases unemployment when wages start increasing, and therefore increases the contribution to fertility by unemployed individuals; and larger fertility among unemployed individuals, compared to those who are employed, is a destabilising factor [4, 5, 14, 15].

Indeed, fertility may be very high when employment is low, so an increasing pace of population increase, further reducing employment, provides an instability mechanism if the demand for labour is unchanged. This recalls the old Malthusian intuition that the fertility of unemployed individuals may be destabilising since it is unrelated to other economic variables, and so hardly controllable by internal feedback in the economic system. It therefore emerges that instability arises not only in a neoclassical world, for there are cases where a too strong "neoclassicism" may itself be the cause of instability, as the spark triggering the Malthusian mechanism. Under Malthus' focus on the fertility of the "poor", it is possible to completely reverse the role that neoclassical technology has played in growth theory. In this case, a way to restore a stable economy is, paradoxically, to reduce the flexibility in production that Solow believed to be the "panacea" to remove the Harrod–Domar instability.

To close this section, let us now consider what happens in the special but pervasive case of Cobb–Douglas technology, arising in the limit for $\vartheta \to 0$), using the following remark.

REMARK 2.
In the Cobb–Douglas case, the stability condition of the balanced growth path

$$\frac{\rho}{1 - V_1(0)} - \delta > 0 \quad \Rightarrow \quad z < \frac{\rho}{\delta}$$

leads to the observation that stability does not necessary occur, but requires a balance between (a) the weight of capital in the CES technology (z); (b) the speed of adjustment of the labour market (ρ); and (c) the set of parameters influencing fertility (δ).

3.3. Bifurcation of the balanced growth path

In reappraising the neoclassical solution of the knife-edge problem, let us now consider the nature of the instability of the balanced growth path discussed in Proposition 2. It is a trivial matter to check that all cases of instability described in Proposition 2 occur by a Hopf bifurcation of the growth steady state (cf. Guckenheimer and Holmes [24]). Indeed, all possible cases of instability occur when there is a change of sign of the trace $Tr\,(J_1)$, but the determinant maintains a strictly positive sign. Thus all the points in Proposition 2 at which stability is lost are ϑ_H points — i.e., they correspond to a transition from stable eigenvalues (i.e., with negative real parts) to unstable eigenvalues. In addition, it is easy to check that the test for nonzero speed is fulfilled — i.e.,

$$\left(\frac{d}{d\vartheta}\left(Tr\,(J)\right)\right)_{\vartheta=\vartheta_H} \neq 0$$

at all ϑ_H points, thereby completing the proof of the existence of a Hopf bifurcation at the ϑ_H points. The implication is that the kind of instability appearing in our extended neoclassical model is of oscillatory type. The simulations of the next section show in addition that the emerging cycles are always at least locally asymptotically stable, and we comment further about this at the end of the next section.

4. Simulation and working of the system: steady oscillating balanced growth paths

We now illustrate the stability properties of the model, and the onset of oscillations detected in the previous section, by a concrete example in which we focus only on the dynamical effects of the parameter θ, keeping all other economic parameters fixed.

The Hopf bifurcation theorem only predicts the onset of oscillations, and it does not say whether the bifurcation is supercritical or subcritical — i.e., whether the emerging periodic orbit is locally stable or unstable. For this reason, we undertook a simulation to investigate the stability properties of the periodic orbits emerging via Hopf bifurcation of the balanced growth state P_1, and more generally the global behaviour of the model. In these numerical experiments, we used the following parameter values (in appropriate units): $a_0 = \mu = 0$, $z = 0.5$, $s_p = 1$, $c = 5$, $h = 0.45$, $\gamma = 0.009$, $\rho = 0.01$, $a = 0.99$, $q = 0.2, q_u = 0.05$.[6] The inequality $z > G$ holds in this example, so it falls under the assumption of (C b) of Proposition 2, which predicts the possibility of instability and oscillations as a consequence of too strong a degree of neoclassicity in production. The system is initialised from values very close to the balanced growth state P_1.

[6] We chose $a_0 = \mu = 0$ because these two parameters, obviously necessary to fit real data, do not affect the qualitative features of the model.

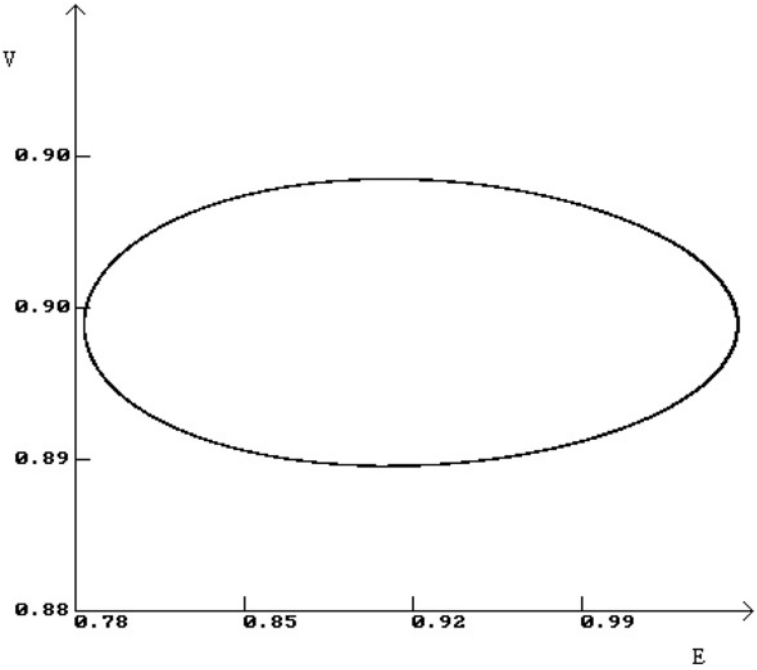


FIGURE 3. A phase-plane view of a stable limit cycle for θ close to $\theta_H = 0.775$, (parameter set as in the text); initial conditions: $V(0) = 0.95, E(0) = 0.91$.

The simulation shows that, when the technology displays a high degree of factor substitution (i.e., a relatively high θ: $\theta < 0.775$), the balanced growth state P_1 is unstable. The instability is not only maximal with a perfect substitution technology, but the system also continues to be strongly unstable with a Cobb–Douglas technology. On gradually reducing the elasticity of substitution, the system shows progressively less wild unstable oscillations, until the threshold value $\theta = 0.775$ where the oscillations end in a stable cycle. From the economic point of view, this implies patterns of growth with steady cycles around the path of balanced growth. Further reductions in θ stabilise the system. However, further reduction in θ leads to a second Hopf bifurcation (at $\theta = 7.7$), and again to steady oscillations; and by further approaching the fixed coefficents world, to purely unstable oscillations resembling the classical knife-edge behaviour. Morover, on starting from a "fully" neoclassical technology (e.g., the perfect substitution case $\theta = -1$) and progressively reducing the degree of factor substitution, we find the phase portrait of the system undergoes the following transformations: no balanced growth state at first → unstable balanced growth state P_1 (initially monotonic and subsequently oscillatory) → *unstable* P_1, but with convergence to a stable limit cycle → P_1 as a stable focus, then a stable node, then a stable focus again) → *destabilised*

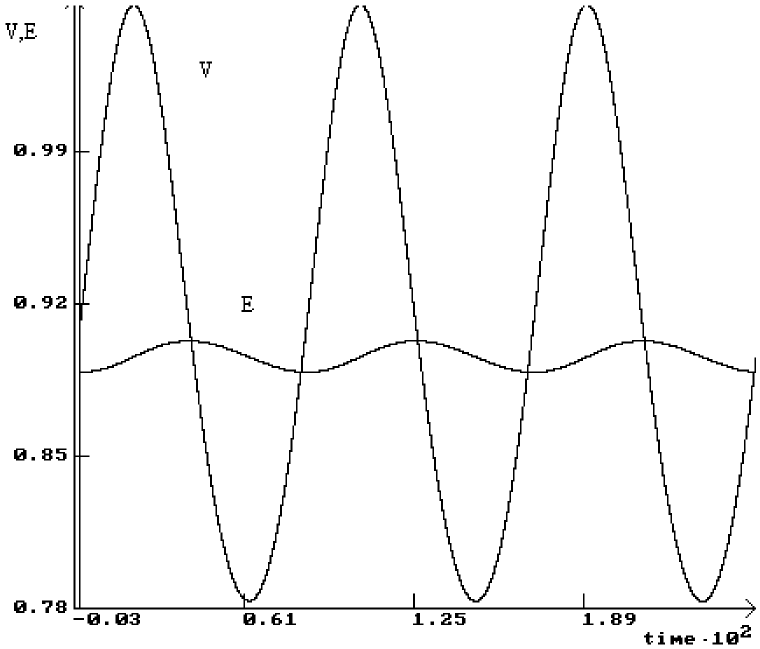


FIGURE 4. Time paths of the rate of employment and wage share in the limit cycle; initial conditions: $V(0) = 0.95, E(0) = 0.91$.

P_1, with convergence to a stable limit cycle → the cycle is destabilised, and full instability occurs.

It is worthwhile remarking on at least three aspects that emerged from the simulation. Firstly, both of the limit cycles that emerged, one for a high degree and one for a small degree of substitution, were locally stable (and in all of our simulations appeared to be globally stable), indicating the possibility that there are two different cyclical growth paths, depending on quite different regimes of flexibility in production. Secondly, a limit cycle exists in a significantly wide range of θ, implying that fluctuating rather than stationary growth is the rule for wide ranges of technologies. (From this standpoint, our model predicts the stylised fact of cyclical growth.) Thirdly, the oscillatory time paths of the unemployment rate, and of the distributive shares, are consistent with another stylised fact of economic growth — viz. that the rate of unemployment shows sharp fluctuations, whereas distributive shares oscillate only slightly.

Another remark, more in the vein of the classical debate on growth, concerns the nature of the instability occurring in our model. Compared to the basic Harrod–Domar's model, where the knife-edge represented an elegant definition for the pure instability of a steady state, our model includes a much more reasonable instability of the steady state in the form of steady oscillations. This is the consequence of the fact that the model has a Goodwinian feature — viz. oscillatory potential due to sluggish labour markets. In brief, the synthesis of the Goodwin and Solow models presented in this paper has produced a very fruitful approach to growth.

5. Conclusions

There is a widespread belief that all growth models based on a non-neoclassical production theory, such as the Harrod–Domar or Goodwin models, exhibit instability of the balanced growth path. It is also believed that a steady state can always be "stabilised" by introducing neoclassical technology and the behaviour of firms. This need need not be so.

This paper shows that neoclassical Solow models with sluggishly adjusting real wages and endogenous fertility (extensions suggested by Solow himself), may be unstable. The surprising result is that, under fertility behaviour as postulated here, the existence of too strong a "neoclassicism" triggers instability in the model economy. Paradoxically, one way to restore a stable economy is then to reduce the flexibility in production, which Solow believed to be the "panacea" for removing the Harrod–Domar knife-edge instability. From this standpoint, we believe that the model developed in this paper represents a significant extension of the neoclassical growth model, and sheds new light on the relation between stability and neoclassical production theory in a growing economy.

The present model also classifies the nature of the instability, showing when it appears as steady oscillations as a consequence of the oscillatory potential embedded in the hypothesis of sluggish wages (the heritage from Goodwin), and when it degenerates into knife-edge instability. Moreover, we have shown that cycles: (a) can be caused by changes in the degree of flexibility in production: and (b) can occur not only in non-neoclassical situations with scarce substitution but also in strongly neoclassical ones with very high degrees of factor substitution.

References

[1] Ahn, H., Mira, P., 2001. Job bust, baby bust? Evidence from Spain. Journal of Population Economics. 14, 3, 505–521.

[2] Barro R. J., Sala-y-Martin X., 1995. Economic Growth. New York: McGraw-Hill.

[3] Fanti L., Manfredi P. 1998, A Goodwin-type Growth Cycle Model with Profit-sharing, Economic Notes 3, 183–214.

[4] Fanti L., Manfredi P., 2003a. Population, Unemployment, and Economic Growth Cycles: a further explanatory perspective, Metroeconomica 54, 2, 179–207.

[5] Fanti L., Manfredi P., 2003b. The Solow's Model with Endogenous Population: A Neoclassical Growth Cycle Model, Journal of Economic Development 28, 2, 103–115.

[6] Fanti L, Manfredi P., 2006. Neoclassical labour market dynamics, chaos and the real wage Phillips curve, Journal of Economic Behaviour and Organisation 62, 3, 470–483.

[7] Fazzari S, Hubbard R., Petersen B., 1988. Financing Constraint and Corporate Investment, Brookings Papers on Economic Activity 1, 141–195.

[8] Goodwin R. M., 1972. A growth cycle. In Hunt E.K., Schwartz J.G. (eds.). A critique of Economic Theory. Harmondsworth.

[9] Jones C. I., 1999, Was an Industrial Revolution Inevitable? Economic Growth Over the Very Long Run. Cambridge, MA: NBER, WP 7375.

[10] Holt C. C., 1970. Job Search, Phillips Wage Relation and Union Influence. In Phelps E. S. (ed.), Microeconomic Foundations of Employment and Inflation Theory. New York: Norton & Co.

[11] Lipsey R. G., 1960, The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the U.K. 1861-1957: a Further Analysis. Economica 27, 1–31.

[12] Manfredi P., Fanti L., 2000a. Long term effects of the efficiency wage hypothesis in a Goodwin-type model, Metroeconomica 51, 4, 454–481.

[13] Manfredi P., Fanti L., 2000b. Long term effects of the efficiency wage hypothesis in a Goodwin-type model: a reply. Metroeconomica 51, 4, 488–491.

[14] Manfredi P., Fanti L., 2006a. The complex effects of demographic heterogeneity on the interaction between the economy and population. Structural Change and Economic Dynamics 17, 148–173.

[15] Manfredi P., Fanti L., 2006b. Demography in Macroeconomics Models: When Labour Supply Matters for Economic Cycles. Metroeconomica 57, 4, 536–563.

[16] Momota A., Futagami K., 2000. Demographic transition pattern in a small country, Economics Letters 67, 231–237.

[17] Solow R. M., 1956. A model of balanced growth, Quarterly Journal of Economics 70, 65–94.

[18] Solow R. M. 2000. Growth Theory. New York: Oxford University Press.

[19] Solow R. M., Tobin J., von Weizsacker C. C., Yaari M. 1966. Neoclassical Growth with Fixed Factor Proportions. Review of Economic Studies, 33, 2.

[20] Southall, H., Gilbert, D. 1996. A good time to wed? Marriage and economic distress in England and Wales, 1883–1914. Economic History Review 49, 1, 35–57.

[21] Staiger, D., Stock, J. H.,Watson, M. W., 2001. Prices,Wages and the US NAIRU in the 1990s. Cambridge, MA: NBER, Working Paper 8320.

[22] Van der Ploeg F., 1985. Classical Growth Cycles, Metroeconomica, pp. 221–230.

[23] Kaldor N., 1957. A model of economic growth, Economic Journal, **57**, 591–624.

[24] Guckenheimer J., Holmes, P., 1983. Nonlinear oscillations, dynamical systems, and bifurcation of vector fields. New York, Tokyo: Springer-Verlag

Appendix

Proof of Proposition 1 (Existence of a nontrivial growth steady state)

Let us reconsider equation (21) — viz.

$$(1-V)^{1+\frac{1}{\theta}} = Gz^{\frac{1}{\theta}} \,. \tag{A.30}$$

Since a discontinuity occurs for $\vartheta = 0$, we distinguish two cases — viz. $\vartheta > 0$ and $-1 < \vartheta < 0$.

Case 1: $\vartheta > 0$**.** When $\vartheta > 0$, the function $f_1(V) = (1-V)^{\frac{1+\theta}{\theta}}$ is non-negative, monotonically decreasing and convex for $0 \leq V \leq 1$, as

$$\frac{df_1(V)}{dV} = (-1)\,\frac{1+\theta}{\theta}(1-V)^{\frac{1}{\theta}} < 0 \quad ; \quad \frac{d^2 f_1(V)}{dV^2} = \frac{1+\theta}{\theta^2}(1-V)^{\frac{1}{\theta}-1} > 0 \tag{A.31}$$

with $f_1(0) = 1$, $f_1(1) = 0$, so $f_1(V)$ is bounded in $[0,1]$. Moreover,

$$\frac{\partial f_1(V)}{\partial \vartheta} = (1-V) * f_1(V)\,\frac{-1}{\vartheta^2}\ln(1-V) > 0 \tag{A.32}$$

as $0 \leq 1-V < 1$, implying that for $\vartheta > 0$ an increase in ϑ produces an upward shift in f_1. When $\vartheta \to +\infty$, $f_1(V)$ approaches the line $(1-V)$, as in the Goodwin model. Moreover, as $0 < z < 1$, the quantity $f_2(\vartheta) = z^{1/\vartheta}$ is monotonically increasing in ϑ for $\vartheta > 0$, with $\lim\limits_{\vartheta \to 0^+} f_2(\vartheta) = 0$ and $\lim\limits_{\vartheta + \infty} f_2(\vartheta) = 1$. This in turn implies that, as ϑ increases between 0 and $+\infty$, the right-hand side of (21) $Gz^{\frac{1}{\theta}}$ increases between 0 and G. Thus if $0 \leq G \leq 1$ a unique positive equilibrium exists and is always admissible. On the other hand, if $G > 1$ a positive equilibrium exists and is admissible for $Gz^{\frac{1}{\theta}} < 1$, or $z^{\frac{1}{\theta}} < G^{-1}$ for

$$\vartheta < \vartheta_{**} \tag{A.33}$$

where $\vartheta_{**} = (-1)\,\frac{\ln z}{\ln G} > 0$ — i.e., it is a meaningful bound. We summarise our results as follows:

RESULT 1a. For $\vartheta > 0$, the system (19) admits a unique equilibrium that is always admissible if $G < 1$, while if $G > 1$ a unique admissible equilibrium exists in the set $0 < \vartheta < \vartheta_{**}$ where

$$\vartheta_{**} = (-1)\,\frac{\ln z}{\ln G} > 0. \tag{A.34}$$

Case 2: $-1 < \vartheta < 0$**.** When $-1 < \vartheta < 0$, the function $f_1(V) = (1-V)^{\frac{1+\theta}{\theta}}$ is monotonically increasing and convex in $0 \leq V < 1$, with $f_1(0) = 1$, $\lim\limits_{V \to 1} f_1(V) = +\infty$. Since for $-1 < \vartheta < 0$ the quantity $z^{1/\vartheta}$ is monotonically increasing and convex in ϑ between $z^{-1} > 1$ and $+\infty$, so for $G > 1$ the function $Gz^{\frac{1}{\theta}}$ is greater than one and a unique admissible equilibrium always exists. On the other hand, for $G \leq 1$ we need $Gz^{\frac{1}{\theta}} > 1$ — i.e., after some algebra, we have

$$\vartheta > (-1)\,\frac{\ln z}{\ln G} = \vartheta_{*} \tag{A.35}$$

where $\vartheta_* < 0$ but not necessarily greater than -1. Thus if $(-1)\frac{\ln z}{\ln G} < -1$ (i.e., $0 < z < G < 1$), the threshold ϑ_* is not meaningful and a positive equilibrium exists for all ϑ values — i.e., for $-1 < \vartheta < +\infty$.

We may therefore summarise our results on the role of ϑ for the existence and admissibility of the positive equilibrium in the case $-1 < \vartheta < 0$ as follows:

RESULT 1b. For $G > 1$, the system (19) admits a unique admissible equilibrium for all $\vartheta < 0$. For $z < G \leq 1$, there is again a unique admissible equilibrium for all $\vartheta < 0$. Finally, for $0 < G < z$ an admissible equilibrium exists in the set $\vartheta_* < \vartheta < 0$ where

$$\vartheta_* = (-1)\frac{\ln z}{\ln G} \; .$$

By combining RESULTS 1a and 1b, and distinguishing the three cases (A), (B) and (C) on the basis of the mutual values of G and z, we obtain Proposition 1 in the main text.

Other useful results

We prove here some results that are useful for discussing the stability of states of balanced growth. At the positive equilibrium P_1, the following relation holds:

$$1 - V = [G]^{\frac{\theta}{1+\theta}} z^{\frac{1}{1+\theta}} \; .$$

The function

$$\pi(\vartheta) = 1 - V(\vartheta) \tag{A.36}$$

is strictly increasing in ϑ for $G > z$, and strictly decreasing for $G < z$. Indeed, by some simple algebra we have

$$\frac{d}{d\vartheta}\pi(\vartheta) = \frac{d}{d\vartheta}(1 - V(\vartheta)) = (1 - V(\vartheta))\left(\frac{1}{1+\vartheta}\right)^2 \ln\frac{G}{z} \; ,$$

so in particular the assumption $G > z$ leads to the economic condition

$$\frac{n(E_1) + a_0}{zcs_p} > 1 \; .$$

Let us consider the behaviour of $\pi(\vartheta)$ for $\vartheta \to -1$. For $G > 1$, the P_1 equilibrium exists and is admissible in $(-1, \vartheta_{**})$. Thus since $G > 1$ implies $G > z$, we have

$$\lim_{\vartheta\to -1^+}\pi(\vartheta) = \lim_{\vartheta\to -1^+}(1 - V(\vartheta)) = G\lim_{\vartheta\to -1^+}\left(\frac{z}{G}\right)^{\frac{1}{1+\theta}} = 0 \; .$$

For $G < 1$, the P_1 equilibrium exists only for $(\vartheta_*, +\infty)$, where from (A.35) we have $\vartheta_* = (-1)\ln z/\ln G$. Consequently,

$$\lim_{\vartheta\to\vartheta_*}\pi(\vartheta) = \lim_{\vartheta\to\vartheta_*}(1 - V(\vartheta)) = G\lim_{\vartheta\to\vartheta_*}\left(\frac{z}{G}\right)^{\frac{1}{1+\theta}} = k \; , \tag{A.37}$$

where k is positive and smaller than 1.

Proof of Proposition 2

From (29), we need to consider the behaviour of $H(\vartheta)$ for $\theta \in (-1, +\infty)$. Now $H(\vartheta)$ is always positive for admissible equilibrium values of V. For $G > 1$, the De l'Hopital theorem yields

$$\begin{aligned} \lim_{\vartheta \to -1^+} H(\vartheta) &= \lim_{\vartheta \to -1} (1+\vartheta)\,\pi(\vartheta) = \lim_{\vartheta \to -1} \frac{\frac{d}{d\vartheta}(1 - V(\vartheta))}{\frac{d}{d\vartheta}\frac{1}{(1+\vartheta)}} \\ &= (-1)(1+\vartheta)^2 (1 - V(\vartheta)) \ln \frac{G}{z} = 0 \,. \end{aligned} \tag{A.38}$$

For $G < 1$,

$$\lim_{\vartheta \to \vartheta_*^+} H(\vartheta) = (1 + \vartheta_*)\,k > 0 \,. \tag{A.39}$$

Since the function $1 + \vartheta$ is monotonically increasing, we can distinguish the three cases (A), (B) and (C) of the text:

Case (A):
For $G > 1$ (so that $G > z$), the function $1 - V(\vartheta)$ is increasing, implying that $H(\vartheta)$ is also increasing in the whole domain $\Omega_1 = \{-1 < \vartheta < \vartheta_{**}\}$ in which P_1 exists (and is meaningful). From (A.38) $\lim_{\vartheta \to -1^+} H(\vartheta) = 0$, the P_1 equilibrium is always locally stable when $\vartheta \to -1$. By a continuity argument, stability continues to prevail in a neighbourhood of (-1), and in particular instability will occur (through a Hopf bifurcation at the point where $H(\vartheta) = \rho/\delta$) only if $H(\vartheta_{**}) > \rho/\delta$.

Case (B):
For $0 < G < 1$ but $G > z$, $H(\vartheta)$ is strictly increasing in the domain $\Omega_2 = \{\vartheta_* < \vartheta < +\infty\}$ in which P_1 exists (and is meaningful). As

$$\lim_{\vartheta \to +\infty} (1 - V(\vartheta)) = G \,,$$

it follows that $\lim_{\vartheta \to +\infty} H(\vartheta) = +\infty$. Let $H_* = H(\vartheta_*)$. If $H_* < \rho/\delta$, the balanced growth state P_1 is locally stable in a neighborhood of ϑ_*; but as $H(\vartheta)$ grows unbounded with ϑ, instability will necessarily arise (through a Hopf bifurcation) for large ϑ values. However, if instead $H_* > \rho/\delta$, then P_1 is always unstable.

Case (C):
For $0 < G < z$ ($z/G > 1$), P_1 still exists and is admissible on $\Omega_2 = \{\vartheta_* < \vartheta < +\infty\}$. In this case, unlike cases (A) and (B), $H(\vartheta)$ is the product of an increasing function with a decreasing one. Again, $H(\infty) = +\infty$. Moreover,

$$\frac{dH}{d\vartheta} = \frac{1 - V(\vartheta)}{1 + \vartheta}\left(1 - \ln\frac{z}{G} + \vartheta\right) \,. \tag{A.40}$$

Consequently, since $G < z$ we have $\ln\frac{z}{G} > 0$ and $1 - \ln\frac{z}{G} < 1$, so that a value ϑ_2 exists ($\vartheta_2 = \ln\frac{z}{G} - 1$) such that $H(\vartheta)$ is decreasing for $\vartheta < \vartheta_2$ and increasing thereafter ($\vartheta = \vartheta_2$ represents a minimum for $H(\vartheta)$). Thus stability depends on the mutual position of the line ρ/δ with respect to the value $H(\vartheta_*)$ and the minimum of the curve $H(\vartheta)$, so the three cases in Figure 2 emerge and lead straightforwardly to cases (C1), (C2) and (C3). Again, when instability arises it does so through a Hopf bifurcation of the steady state.

Luciano Fanti
Dipartimento di Scienze Economiche
Via Ridolfi 10, 56124 Pisa
Italy

Piero Manfredi
Dipartimento di Statistica e Matematica Applicata all'Economia
Via Ridolfi 10, 56124 Pisa
Italy
Tel. 0039-050-2216317
Fax 0039-050-2216317
e-mail: manfredi@ec.unipi.it

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 343–362

A Bass-type Model for a Dynamic Market with Logistic Growth

Franscesca Centrone and Ernesto Salinelli

Abstract. A model of the diffusion of a new product into a market with two segments but following a logistic demographic is considered. The consequent diffusion and adoption curves obtained are compared with previous results from two related models that assume a fixed population and an exponential dynamic market, respectively.

Mathematics Subject Classification (2000). Primary 90B60; Secondary 34C11, 91B62.

Keywords. Adoption, Bass model, Diffusion, Logistic dynamics, Relative sales.

1. Introduction

Since the seminal work of Fourth and Woodlock [7] and Mansfield [14], and particularly the celebrated model introduced by Bass [1], the penetration of consumer durables into a population of potential customers has been the subject of many contributions in the field of marketing.

The *Bass model* describes the "first purchase" diffusion process of a new product launched into a market of fixed composition and size N, consisting of two segments at any time t — viz. the *unawares* $U(t)$, those uninformed about the new product, and *adopters* $A(t)$, those who are informed. The latter are called adopters because it is explicitly assumed that they decide to actually buy the product at the same time they are informed. From this perspective, the Bass model is a "diffusion model", focusing on the diffusion of information and leaving aside details of the economic process of purchase. The information-adoption channels between the two segments are considered to be the *external influence* due to the mass media and advertising, and the *internal influence* due to interpersonal communication (*word-of-mouth*). The market population is assumed to mix homogeneously and be homogeneously exposed to the external influence, and there are no social or economic differences inducing a different rate of adoption of the new product in a fixed time interval. At any time $t > 0$, a percentage $p \in (0,1)$ is taken to be reached by the marketing message, where p is called the *coefficient of innovation* or *of external influence*, and contact between unawares and adopters to produce new adopters in a percentage $q \in (0,1)$, where q is called the *coefficient of imitation* or *of internal influence.*

The Bass model is thus described by the system of ordinary differential equations

$$\begin{aligned} U' &= -pU - q\frac{UA}{N}\ , \\ A' &= pU + q\frac{UA}{N}\ , \end{aligned} \tag{1}$$

where the prime denotes the time derivative. Since $U(t) + A(t) = N$ at any time t, the elimination of $U(t)$ yields the equation

$$A' = p(N - A) + q\frac{A}{N}(N - A) \tag{2}$$

for the adopters that, subject to the initial condition $A\,(0) = 0$, has the solution

$$A(t) = N\,\frac{1 - e^{-(p+q)t}}{1 + (q/p)e^{-(p+q)t}}\ . \tag{3}$$

The graphs of the function A and its derivative A' (the rate of increase of adopters) are called the *diffusion* and *adoption* curves, respectively. Indeed, if every adopter acquires the product just once and there are no succeeding generations of the product, new adoptions may be identified with the current sales and the adopters with cumulative sales. The diffusion and adoption curves are illustrated in Figure 1, where the "relative adopters" ratio $\mathcal{A}(t) = A(t)/N$ is plotted against time t. When word-of-mouth is more effective than advertising $(q > p)$, sales peak at $t_* = (p+q)^{-1}\ln\left(q/p\right)$ when $A\left(t_*\right) = N\left(q - p\right)/2q$ and then decrease. In particular, if

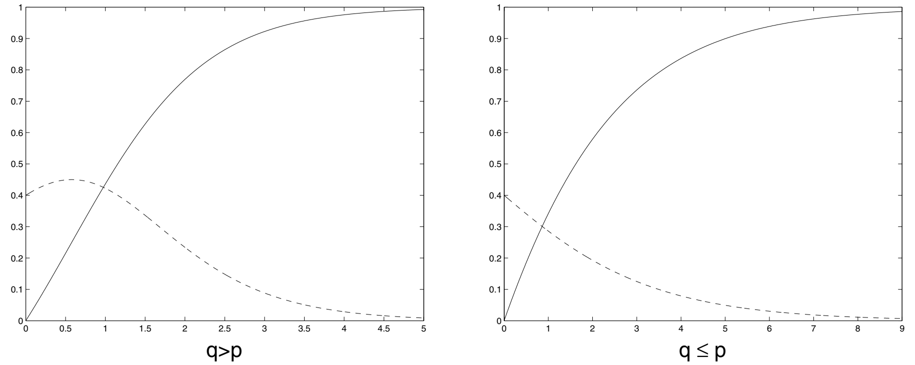


FIGURE 1. Diffusion (solid line) and adoption curves (dotted line) in the Bass model.

$q \gg p$ the sales attain their maximum value at about the time that cumulative sales are approximately one-half of N. When advertising is more effective than word-of-mouth ($q \leq p$), sales strictly decrease over time. In any case, in the long run the market is saturated — i.e. $A(t) \to N$ as $t \to +\infty$. The current sales produced by the external influence decrease as the adopters increase, whereas sales induced by the internal influence display a "logistic" dynamics, with an initially fast and then slower growth — cf. the relevant terms in (2).

Despite its simplicity, the Bass model (2) does describe the basic mechanisms of the diffusion process and forecasts sales in various consumer durable sectors. A number of extensions have been proposed (e.g. [2, 6, 8]), but most assume a *fixed market population* — although in many markets there is growth due to demographic or economic factors such as pricing, government action or marketing. Previous contributions to account for a market of variable size ([9, 11–13, 16]) simply assume that the variable N in the Bass model depends on the time and possibly such economic variables, without considering how the demographic component influences fluxes in market segments or the diffusion dynamics. More recently however, the following model of first purchase in an exponential dynamic market was introduced [4]:

$$\begin{aligned} U' &= bN - pU - q\frac{UA}{N} - \mu U \ , \\ A' &= pU + q\frac{UA}{N} - \mu A \ , \\ N' &= (b - \mu) N \ , \end{aligned} \tag{4}$$

where $b > \mu$. In this exponential model, the "birth" process involves only the unawares segment, but the "death" process involves more than one compartment at the common rate μ. Thus the parameters b and μ capture not only a purely demographic birth-death process but also socioeconomic aspects — each individual can enter or leave the market according to age, wealth, or personal preference for example. Since $U(t) + A(t) = N(t)$, the resulting differential equation governing the evolution of the relative adopters ratio $\mathcal{A} = A/N$ then obtained from (4) is

$$\mathcal{A}' = p + (q - p - b)\mathcal{A} - q\mathcal{A}^2 \ , \tag{5}$$

containing the additional term with coefficient b. Properties of the exact solution $\mathcal{A}_E$ of (5), given the initial condition $\mathcal{A}_E(0) = 0$ and its dependence on the demographic-economic parameters, have previously been discussed [4]. Major features in this model can be summarised as follows:

- a relative study of the diffusion and adoption curves is necessary to account for variation in the market size;
- adoptions and sales (both relative and not) do not coincide because of the mortality process;
- the market is not saturated in the long run; and
- the relative diffusion and adoption path scenarios are essentially the same as in the Bass model, and strictly linked to the relation between q and $p + b$:

indeed, adoptions peak if and only if the internal influence overcompensates for the combined effects of word of mouth and the birth rate ($q > p + b$).

Despite a good fit with data and its usefulness for short to medium term analysis, a major drawback of the exponential model is the unbounded market growth assumption. The aim of this paper is to continue with a Bass-type model, but to assume a market with a general logistic (and hence bounded) growth that we believe to be more realistic — since markets are more likely to exhibit an initial exponential growth phase followed by a deceleration, leading to an eventual near-stationary level. In brief, the logistic model appears flexible enough to capture the possible expansions that various types of markets display. We focus on analysing and interpreting the forms that the corresponding adoption and diffusion curves then take for different feasible parameters, highlighting the similarities and differences compared with the exponential model. The direction is similar to that in our previous work, except that a qualitative analysis and simulation is pursued because an exact solution is not available. We find once again that saturation does not occur, despite the bounded market growth; but we find a greater variety of possible relative adoption and sales pattern configurations, dependent upon market "maturity"— i.e., on how far the initial population size is from the eventual asymptotic level. In particular, there is a parameter configuration that produces multiple peaks, not found in either the Bass or the exponential model.

Our presentation is as follows. In Section 2, we introduce the model, discuss its assumptions and present a stability result. In Section 3, we consider some properties of the adopters equation solution from a qualitative point of view, in comparison with the solution in the exponential case. In Section 4, we analyse the relative and absolute sales dynamics, and discuss our simulations to support the theoretical analysis and interpret the various possible phenomena. Conclusions are drawn in Section 5.

2. Logistic growth model

Let us consider the introduction of a new product at time $t = 0$ into a marketplace of size $N(t)$ that evolves according to the *logistic law* [18]

$$N' = [b(N) - m(N)]\, N, \tag{6}$$

where the birth rate $b(N) > 0$ is differentiable with $b'(N) < 0$ and the mortality rate $m(N)$ is differentiable with $m'(N) > 0$. Assuming $b\,(N_0) > m(N_0)$ where $N_0 = N\,(0)$ and that $m(+\infty) > b(+\infty)$, there is a unique globally asymptotically stable (GAS) equilibrium N^* such that

$$b(N^*) = m(N^*). \tag{7}$$

Given its greater economic importance, we only consider the case $N_0 < N^*$.

Among the possible birth and mortality rate behaviours, we will sometimes refer to the *linear logistic* case

$$b(N) = b - k_1 N \quad \text{and} \quad m\,(N) = \mu + k_2 N \quad \text{where} \quad k_1, k_2 > 0\ , \tag{8}$$

and to the *density-independent fertility (DIF)* case

$$b\,(N) = b > m(N). \tag{9}$$

As indicated in the Introduction, both b and μ are intended to be demographic-economic parameters.
The market is composed of the unawares $U(t)$ and adopters $A(t)$ such that $N\,(t) = U\,(t) + A\,(t)$ at any time t, and

$$A\,(0) = 0, \qquad\qquad U\,(0) = N_0 > 0.$$

As in the exponential model, let us suppose that the contact process between unawares and adopters is captured by the so called *true mass action form* [4] — i.e., at any time t, the rate of transition of individuals from the first segment to the second is proportional to UA/N. Thus the model is described by the system of differential equations

$$\begin{aligned} U' &= b\,(N)\,N - m\,(N)\,U - pU - q\frac{UA}{N}\ , \\ A' &= -m\,(N)\,A + pU + q\frac{UA}{N}\ , \\ N' &= (b\,(N) - m\,(N))\,N\ . \end{aligned} \tag{10}$$

It is convenient to work with the relative ratios $\mathcal{U} = U/N$ and $\mathcal{A} = A/N$, so that the system becomes

$$\begin{aligned} \mathcal{U}' &= b\,(N) - [b\,(N) + p]\mathcal{U} - q\mathcal{U}\mathcal{A}\ , \\ \mathcal{A}' &= -b\,(N)\,\mathcal{A} + p\mathcal{U} + q\mathcal{U}\mathcal{A}\ , \\ N' &= [b\,(N) - m\,(N)]\,N\ . \end{aligned} \tag{11}$$

for the solution set

$$\Gamma = \{(\mathcal{U}, \mathcal{A}, N) : \quad 0 < N \leq N^*,\ \mathcal{U} \geq 0,\ \mathcal{A} \geq 0,\ \mathcal{U} + \mathcal{A} = 1\}\ .$$

Since $\mathcal{U}\,(t) + \mathcal{A}\,(t) = 1$, we obtain the corresponding relative adopters equation

$$\mathcal{A}' = p + (q - p - b\,(N))\,\mathcal{A} - q\mathcal{A}^2\ . \tag{12}$$

Remark 1. *The dynamics of the adopters described by* (12) *does not depend on the mortality function* $m\,(N)$, *a direct consequence of the assumption that in each compartment the mortality rate is the same — cf. also (5). Furthermore, the variation of the relative adopters over time is increased by the economic component* $p\,(1 - \mathcal{A}) + q\mathcal{A}\,(1 - \mathcal{A}) > 0$ *similar to the Bass and exponential models, and is decreased by the demographic term* $-b(N)\mathcal{A} < 0$ *similar to the exponential model.*

Remark 2. *The sign of the coefficient $\beta(N) = q - p - b(N)$ is not determined a priori, neither at N_0 nor at N^*. From the assumptions on $b(N)$ it follows that*

$$\beta'(N(t)) = -b'(N)\,N'(t) > 0 \tag{13}$$

with

$$\lim_{t\to+\infty} \beta'(N(t)) = 0^+ \,. \tag{14}$$

Hence if $\beta(N_0) \geq 0$, then $\beta(N) > 0$ for all $t > 0$, and if $\beta(N^*) \leq 0$ then $\beta(N) < 0$ for all $t > 0$.

Unlike the original Bass model, in the exponential model (5) the market never saturates — i.e., $\mathcal{A}_E(+\infty) \neq 1$. At first glance, this may appear to be caused by the continual demographic growth, but this is not the case (cf. [4]). Indeed, the birth of new unawares generally slows the diffusive process, and the magnitude of $A_E(+\infty)$ depends upon the fertility rate b. The following proposition collects the relevant results from a stability analysis of the logistic model.

Proposition 3. *System (11) admits the unique GAS equilibrium $(\mathcal{U}^*, \mathcal{A}^*, N^*)$ on Γ, where*

$$\mathcal{A}^* = \frac{\beta^* + \eta^*}{2q} \qquad \textit{and} \qquad \mathcal{U}^* = \frac{q + p + b(N^*) - \eta^*}{2q}\,, \tag{15}$$

with $\beta^* = \beta(N^*)$ and $\eta^* = \sqrt{(\beta^*)^2 + 4pq}$.

Proof. The equation $N' = 0$ has the nontrivial solution $N^* > 0$. Substituting in (12) and setting $\mathcal{A}' = 0$, we obtain

$$q\mathcal{A}^2 - \beta^*\mathcal{A} - p = 0$$

and hence the unique admissible solution

$$\mathcal{A}^* = \frac{1}{2q}\left(\beta^* + \sqrt{(\beta^*)^2 + 4pq}\right)\,.$$

Since $\mathcal{U}^* + \mathcal{A}^* = 1$, we also have $\mathcal{U}^*$. Then using results for asymptotically autonomous dynamical systems (see e.g. [18]), we integrate the differential equation

$$\mathcal{A}' = p + \beta^*\mathcal{A} - q\mathcal{A}^2$$

subject to the initial condition $\mathcal{A}(0) = \mathcal{A}_0 < 1$, to obtain

$$\mathcal{A}(t) = \frac{\eta^* + \beta^*}{2q}\,\frac{1 + \psi\;\phi\;\exp\{-\eta^* t\}}{1 - \phi\;\exp\{-\eta^* t\}} \tag{16}$$

and

$$\phi = \frac{2q\mathcal{A}_0 - (\eta^* + \beta^*)}{2q\mathcal{A}_0 + (\eta^* - \beta^*)}$$

where

$$\psi = (\eta^* + \beta^*)\,/\,(\eta^* - \beta^*)\,.$$

Since for all $\mathcal{A}_0$ we have $\mathcal{A}(t) \to \mathcal{A}^*$ as $t \to +\infty$, the result follows. □

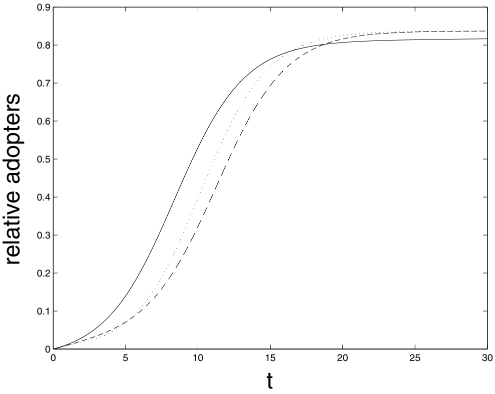


FIGURE 2. Different diffusion processes in the linear logistic population (8) with the same N^* but different values of b: $b = 0.1$ (solid line), $b = 0.5$ (dashed line), $b = 0.9$ (dotted line).

From (15), it is evident that $\mathcal{A}^* \leq 1$ and $\mathcal{A}^* = 1$ if and only if $b\,(N^*) = 0$, so the logistic market does not saturate. This is not so surprising, for eventually the population approaches a stationary level characterised by replacement of individuals, as in a stationary model with $b(N^*) = \mu(N^*) \neq 0$. That is why the adoption mechanism never fades.

Furthermore, note that the negativity of the partial derivative

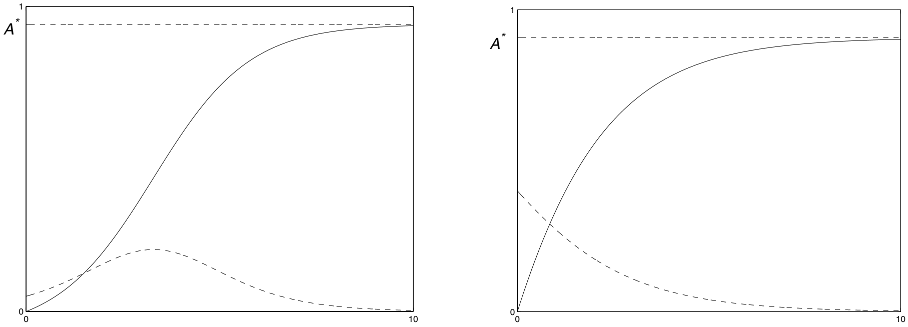
$$\frac{\partial \mathcal{A}^*}{\partial b\,(N^*)} = -\frac{\eta^* + \beta^*}{2q\eta^*}$$

shows that the maximum diffusion decreases as the asymptotic fertility rate grows. Obviously, a different fertility rate $b\,(N)$ is sufficient to produce different diffusion histories (for the same N^* and other parameters fixed), as illustrated in Figure 2 for the case of a linear logistic market.

3. The diffusion and adoption curves

In the Bass model fixed population framework, we recall that the diffusion curve coincides with that of cumulative sales, and the adoption curve with current sales (at time t). Analysis of the diffusion, adoption and sales patterns is more complex in our logistic model from two standpoints — viz. (i) the mortality process, as adoptions and sales do not coincide; and (ii) it is opportune to distinguish between the relative and absolute form of the relevant quantities. Indeed, analysis of the relative adoption and diffusion characterises the "market penetration" of the product. This is the objective of the present section, and the sales are analysed in the next section.

We start by observing that the relative adopters equation (12) in the density-independent fertility case (9) coincides with that in the exponential model (5), so we have an explicit solution that allows us to determine relevant properties of the



$q > p + b$ $\qquad$ $q \le p + b$

FIGURE 3. Relative diffusion (solid line) and adoption (dotted line) curves in the logistic-DIF model.

diffusion and adoption curves (cf. [4]), as collected in the following proposition and illustrated in Figure 3.

Proposition 4. *For the logistic-DIF model, the percentage of adopters*

$$\mathcal{A}(t) = \frac{\eta+\beta}{2q}\frac{1-e^{-\eta t}}{1+\psi e^{-\eta t}} \tag{17}$$

- *is strictly increasing for any admissible configuration of the parameters;*
- *does not saturate the market if* $b > 0$ *(i.e.,* $\mathcal{A}^* = 1$ *if and only if* $b = 0$*);*
- *is concave if and only if* $q \le p + b$*; and*
- *is S-shaped if and only if* $q > p + b$*, with a flex point* $t_* = (\ln\psi)/\eta$ *where the adoption curve reaches its peak — viz.* $\mathcal{A}(t_*) = \beta/2q$ *.*

Proposition 4 shows that in the logistic-DIF case, as in the Bass model, there are only two possible adoption and diffusion scenarios. The first case is characterised by an initial increase in current adoptions supported by word-of-mouth until the peak is reached, and the corresponding diffusion curve is S-shaped. In the second, the current adoptions decline from the initial peak caused by the external influence, and the corresponding diffusion curve is concave. This is similar to the Bass model behaviour, but now the shift from one scenario to the other is determined by the magnitude of q relative to $p + b$. Thus the diffusion curve is S-shaped if and only if the greater effectiveness of word-of-mouth relative to advertising compensates for the rate at which the relative adopters decreases for demographic reasons (cf. also Remark 1).

Remark 5. *From the last bullet point in Proposition* 4*, for* $q \gg p + b$ *the relative current adoptions peak when the diffusion is approximately one-half of the market size.*

Relative adoptions in the logistic-DIF case do not depend upon the structure of the market in which the new product is launched — i.e. the change does not

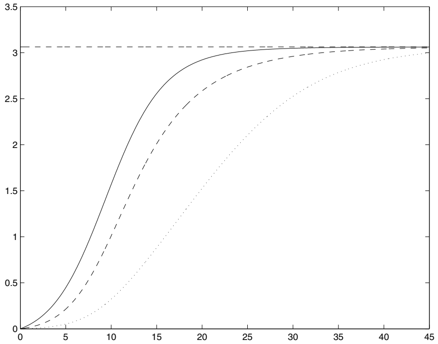


FIGURE 4. Absolute diffusion dynamics in a logistic-DIF market depending on the maturity of the market: $N_0 = 3N^*/4$ (solid line), $N_0 = N^*/20$ (dashed line), $N_0 = N^*/4$ (dotted line).

depend on the difference $N^* - N_0$. This is obviously not the case for the absolute diffusion curve $A(t) = \mathcal{A}(t)\, N(t)$. In the linear case where $m(N) = \mu + k_2(N)$, and with all other parameters fixed, the smaller the value of N_0 the slower the diffusion process (cf. Figure 4). It emerges that when b is not constant, this dependence also alters the relative diffusion curve.

Returning to the general setting of the logistic dynamics (6), we have to consider the Riccati equation (12) with variable coefficient β. We begin by proving a result on the monotonicity of the diffusion curve.

Proposition 6. *For the logistic model (11), the relative diffusion function* $\mathcal{A}$ *is strictly increasing with*

$$\lim_{t\to+\infty} \mathcal{A}'(t) = 0^+ \ . \tag{18}$$

Proof. Let

$$y_\star(t) = \frac{1}{2q}\left(\beta(N) + \sqrt{\beta^2(N) + 4pq}\right)$$

be the time dependent positive solution of the equation $p + \beta y - qy^2 = 0$. It is easy to verify that $y_\star$ is not a solution of (12), for a simple computation shows that

$$y_\star'(t) = \frac{\beta'(N)\, y_\star(t)}{2qy_\star(t) - \beta(N)} > 0$$

since $\beta'(N) > 0$. From $0 = \mathcal{A}(0) < y_\star(0)$, we deduce the existence of $\delta > 0$ such that $\forall t \in I_\delta(0)$

$$\mathcal{A}(t) < y_\star(t) \ , \tag{19}$$

hence $\mathcal{A}'(t) > 0$ for all $t \in I_\delta(0)$. If there is a value $t_0 \geq \delta$ such that $\mathcal{A}'(t_0) = 0$, then the two conditions

$$\mathcal{A}(t_0) = y_\star(t_0) \qquad \text{and} \qquad y_\star'(t_0) > \mathcal{A}'(t_0) = 0 \tag{20}$$

simultaneously hold. From the second condition and the regularity of $\mathcal{A}$ and $y_\star$, in a left neighbourhood of t_0 we would have $y_\star' > \mathcal{A}'(t)$. However, from (19) the first relation in (20) then could not be satisfied. Hence $\mathcal{A}'(t) > 0$ for every $t > 0$. Finally, since $\mathcal{A} \to \mathcal{A}^*$ as t increases, then

$$\lim_{t\to+\infty} \mathcal{A}'(t) = p + \beta^* \mathcal{A}^* - q(\mathcal{A}^*)^2 = 0^+. \qquad \square$$

Thus the percentage of adopters grows over time for any parameter configuration, although it never saturates (recall Proposition 3). This is also enough to sustain the adoptions mechanism indefinitely, as the entrance of unawares slows down and the exit of adopters increases, unlike the exponential model where the population growth rate is fixed. This is not surprising, since the mortality rate of adopters is the same as for the population and at any time there are newcomers to the market.

We would expect the more complex population dynamics of the logistic model to be reflected in a richer set of possible adoption curve patterns. Let us now obtain some more information on the shape of the adoptions curve, in order to further examine similarities and differences between results for the exponential and logistic models. We consider the second derivative

$$\mathcal{A}''(t) = \beta'(N)\mathcal{A}(t) + \beta(N)\mathcal{A}'(t) - 2q\mathcal{A}(t)\mathcal{A}'(t). \tag{21}$$

Proposition 7. *The following conclusions on $\mathcal{A}''$ hold:*

1.
$$\lim_{t\to+\infty} \mathcal{A}''(t) = 0;$$

2. *if $\beta(N_0) > 0$ (respectively < 0) there exists a value $t_* > 0$ such that*
$$\mathcal{A}''(t) > 0 \ (\textit{respectively } < 0) \qquad \forall t \in (0, t_*).$$

Proof. 1. Since β and $\mathcal{A}$ are bounded, from (18) and (14) we immediately obtain $\lim_{t\to+\infty} \mathcal{A}''(t) = 0$. Note that the value of $\mathcal{A}''$ cannot approach zero from above, since this would imply $\mathcal{A}''(t) > 0$ asymptotically such that $\mathcal{A}'(t)$ would be eventually positive and strictly increasing, contradicting (18).

2. It is sufficient to observe that $\mathcal{A}''(0) = p\beta(N_0)$, for the conclusion follows by a continuity argument. Thus when $\beta(N_0) > 0$, we have $t_* > t_{**}$, where t_{**} is the smallest zero of $\beta(N) - 2q\mathcal{A}(t)$. Indeed, since $\beta(N_0) > 0 = \mathcal{A}(0)$, the conclusion follows by the positivity of β' and $\mathcal{A}'$. Note also from (15) that

$$\lim_{t\to+\infty} \beta(N) - 2q\mathcal{A}(t) = \beta^* - 2q\frac{\beta^* + \eta}{2q} = -\eta < 0. \tag{22}$$

$\square$

If we could exclude the possibility of an infinite number of flexes, it would follow from Proposition 7 that $\mathcal{A}''(t) \to 0^-$. Hence if $\beta(N_0) > 0$, the diffusion function $\mathcal{A}$ would have an odd number of flexes, whereas if $\beta(N_0) < 0$ it would have an even number. Moreover, the “regularity” of $N(t)$ and of the Bass and exponential model results (at most one flex point) leads us to conjecture that the logistic model can display at most three flexes. This number is linked to the relevant demographic dynamics, as shown below.

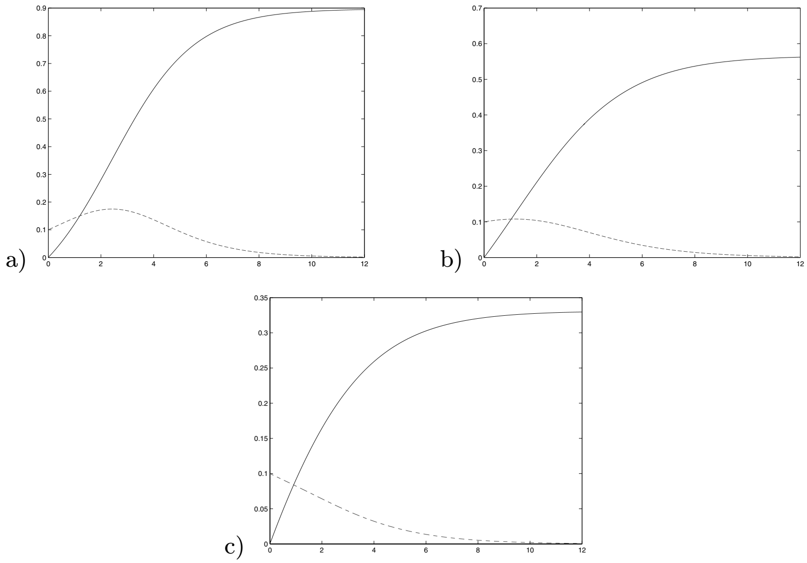


FIGURE 5. Relative diffusion (solid line) and adoption (dashed line) curves for a linear logistic dynamics in a "mature" market with $N_0 = 3N^*/4$, $q = 0.6$, $p = 0.1$, $\mu = 0.01$, $k_1 = 0.01$, $k_2 = 0.02$, for different values of b : (a) $b = 0.1$; (b) $b = 0.5$; (c) $b = 0.9$.

When the demographic component is negligible, more precisely when N_0 is sufficiently close to the maximum expansion level of the market N^*, the dynamics is qualitatively similar to that of the Bass model — i.e., characterised by only two possible histories, depending upon the level of the internal influence parameter. Different fertility and mortality rates (cf. Figure 5) do not seem to substantially change these outcomes, except for the sign of the coefficient β associated with one or the other, and the eventual diffusion level $\mathcal{A}^*$.

The situation is more complex when $N_0 \ll N^*$. Then the evolving market population, which directly affects the unawares and indirectly the adopters, can interact in a more complex way with the influence mechanism.

Figure 6 illustrates this for the linear logistic model with $\beta\,(N) < 0$. Case (a) resembles Bass-like dynamics; but case (b) is more complex. Since $\beta\,(N_0) < 0$, the adoption curve that starts from the level p at time zero initially decreases during the exponential phase in the market population, thereby increasing the relative size of the unawares segment. When the recruitment of new individuals slows, the influence mechanism temporarily prevails, allowing a period of growth in the adoption curve. Finally, on approaching the near-stationary phase, the current adoptions (at any time t) again decrease.

When $\beta\,(N_0) > 0$, the previous dichotomy is maintained (cf. Figure 7), with $\mathcal{A}'$ initially increasing. The additional first flex in $\mathcal{A}$ is due to the internal influence in the exponential phase of the logistic market dynamics.

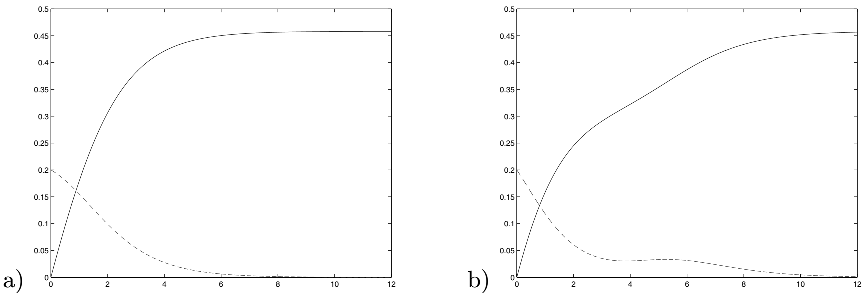


FIGURE 6. Diffusion (solid line) and adoption (dashed line) curves of the two-flexes case for the parameters $b = 1$, $\mu = 0.01$, $p = 0.2$, $q = 0.8$, $k_1 = 0.01$, $k_2 = 0.02$, depending on the initial population level. (a) $N_0 = 3N^*/4$; (b) $N_0 = N^*/60$.

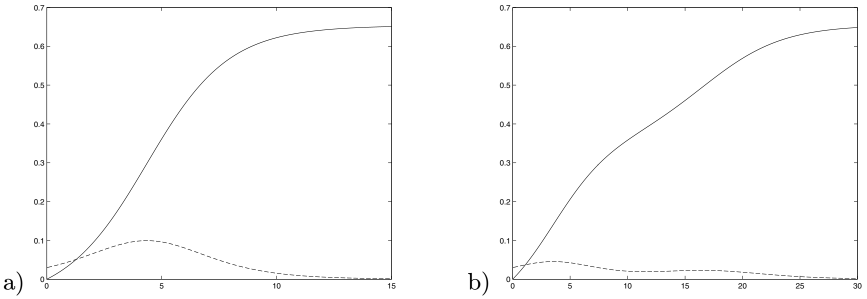


FIGURE 7. Diffusion (solid line) and adoption (dashed line) curves of the three-flexes case for the parameters $b = 0.6$, $\mu = 0.3$, $p = 0.03$, $q = 0.85$, $k_1 = 0.6$, $k_2 = 0.02$, depending on the initial population level. (a) $N_0 = 3N^*/4$; (b) $N_0 = N^*/80$.

It is notable that in both Figure (6b) and Figure (7b) the adoption curve has two peaks, suggesting a marketing strategy to sustain the adoption process until the second peak is reached.

We conclude this section by providing a comparison result illustrated in Figure 8, which shows lower and upper bounds for the relative adoption curve of a general logistic model in the logistic-DIF adoption case.

Proposition 8. *For any p and q,*

$$\mathcal{A}_{\min}(t) \leq \mathcal{A}(t) \leq \mathcal{A}_{\max}(t) \qquad \forall t \geq 0$$

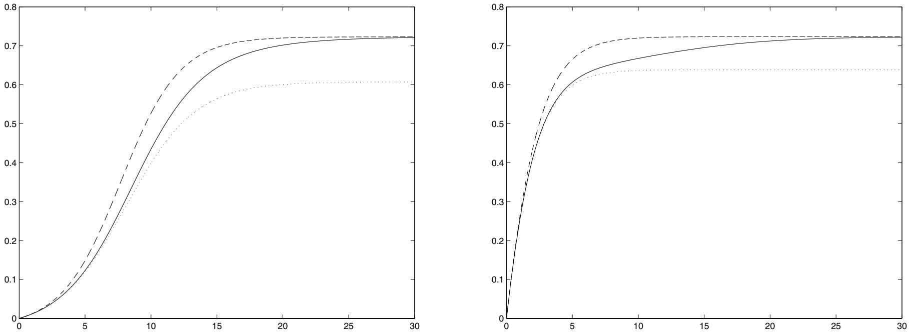
where $\mathcal{A}_{\min}$ and $\mathcal{A}_{\max}$ are the solutions of (12), subject to $\mathcal{A}(0) = 0$ when $b(N) = b(N_0)$ and $b(N) = b(N^*)$, respectively. Furthermore

$$\lim_{t\to+\infty} \mathcal{A}(t) = \lim_{t\to+\infty} \mathcal{A}_{\max}(t). \tag{23}$$

Proof. Since $b\,(N)$ is assumed to be strictly decreasing, for all $(\mathcal{A}, N) \in \Gamma$ it follows that

$$\beta\,(N_0)\,\mathcal{A} \leq \beta\,(N)\,\mathcal{A} \leq \beta^*\mathcal{A}\ ,$$

whence the result from $\mathcal{A}_{\min}\,(0) = \mathcal{A}\,(0) = \mathcal{A}_{\max}\,(0) = 0$, on applying a classical comparison result for ordinary differential equations (cf. [3], Theorem 8, page 23). The equality in (23) can be obtained by comparing $\mathcal{A}^*$ given in (15) with the limit of (17), on setting $b = b\,(N^*)$. □



$q > p + b - kN_0$ $q < p + b - kN^*$

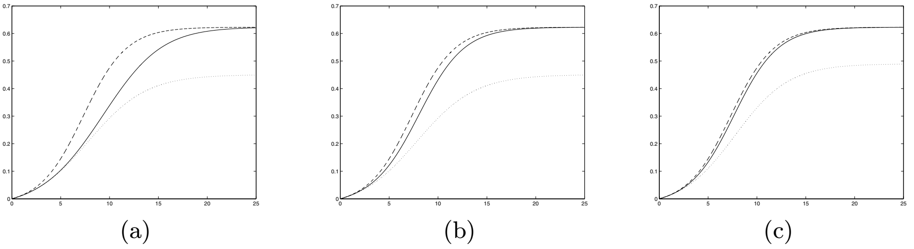
FIGURE 8. Comparison between the diffusion dynamics for $\mathcal{A}$ (solid line), $\mathcal{A}_{\min}$ (dotted line) and $\mathcal{A}_{\max}$ (dashed line) under a linear logistic dynamics.

Figure 8 illustrates some differences in the diffusion process, depending on different demographics. With fixed influence parameters, the eventual diffusion of a logistic market agrees with the logistic-DIF case with fertility rate $b\,(N^*)$, while in its early phase the behaviour depends upon the distance between N_0 and N^*. This is due to the initial exponential but eventual near-stationary dynamics of a logistic evolution. The difference evident in the intermediate evolutionary phase increases with increasing values of N_0.

Figure 9 shows that the lower $\mathcal{A}_{\min}$ curve better approximates the logistic case in the exponential phase when the value of N_0 is small relative to N^*; whereas when N_0 is near N^*, the upper curve $\mathcal{A}_{\max}$ provides a better approximation in both the exponential and near-stationary phase.

4. Sales

In the Bass model where $b = \mu = 0$, the current and cumulative sales are given by the adoption and diffusion curves, respectively. In our logistic model, the situation is more complex for three reasons: (i) current absolute sales S do not coincide with the variation of the adoptions; (ii) one has to distinguish between relative and absolute sales; and (iii) relative cumulative sales make no sense. Let us consider these three points further.



(a) (b) (c)

FIGURE 9. Approximation by minimal (dotted line) and maximal (dashed line) exponential bounds of $\mathcal{A}$ (solid line) in a linear logistic market where $q > p + b - kL$, with fixed parameters other than N_0: viz. (a) $N_0 = N^*/40$, (b) $N_0 = N^*/4$, and (c) $N_0 = N^*/2$.

In order to link S to the time variation of the adopters A', it is necessary to take into account the demographic influence of the mortality term $m\,(N)\,A$. The adopters at time $t + dt$ are

$$A(t + dt) = A(t) - m\,(N\,(t))\,A(t)dt + S(t)dt$$

whence

$$S\,(t) = A'\,(t) + m\,(N)\,A\,(t) = p\,(N - A) + q\frac{(N - A)\,A}{N}. \tag{24}$$

From (27) and since $S > 0$, the corresponding (absolute) cumulative sales given by $\int_0^t S\,(\tau)\,d\tau$ diverge for $t \to +\infty$, corresponding to the long term sustained diffusion due to population replacement, and therefore are of little interest. To take market growth into account, the study of relative sales $\mathfrak{S}\,(t) = S\,(t)\,/N\,(t)$ is relevant, as pointed out empirically in [5] and theoretically in [4]. On noting from (24) that

$$\mathfrak{S}\,(t) = p + (q - p)\,\mathcal{A}(t) - q\mathcal{A}^2(t) \tag{25}$$

we can summarise the main properties of the relative sales curve as follows.

Proposition 9.

1. $\lim\limits_{t\to+\infty} \mathfrak{S}\,(t) = b(N^*)\mathcal{A}^*$;
2. $\mathfrak{S}$ is strictly decreasing if and only if $p \geq q$;
3. there exists a unique $\tilde{t} > 0$ such that $\mathfrak{S}$ is strictly increasing on $\left[0, \tilde{t}\right]$ and then strictly decreasing (i.e. $\mathfrak{S}$ peaks at $\tilde{t}$) if and only if $q > p$ and $\eta^* > b\,(N^*)$;
4. $\mathfrak{S}$ is strictly increasing if and only if $q > p$ and $\eta^* \leq b\,(N^*)$; and
5. when in the logistic-DIF model with $b = b\,(N^*)$ the relative sales function $\mathfrak{S}_{\max}$ peaks at a point $t_{\max}$, then $\mathfrak{S}$ also peaks at time $\tilde{t} > t_{\max}$.

Proof. Note first that $\mathfrak{S}'(t) = (q - p - 2q\mathcal{A})\mathcal{A}'$, whence from $\mathcal{A}' > 0$ we have

$$\mathfrak{S}' \geq 0 \qquad \Leftrightarrow \qquad \mathcal{A} \leq (q - p)\,/2q\;. \tag{26}$$

1. This follows by noting that $\mathfrak{S}\,(t) = \mathcal{A}'(t) + b(N(t))\mathcal{A}(t)$ and recalling the results of the previous sections.

2. If $p \geq q$, then the result follows from the non-negativity of $\mathcal{A}$. Conversely, if $\mathfrak{S}$ is strictly decreasing, then by (26) it must be $\mathcal{A}(t) \geq (q-p)/2q$ for all t; but $\mathcal{A}(0) = 0$ implies $q - p \leq 0$.

3. If $q > p$ and $\eta^* > b(N^*)$, then $\lim_{t\to+\infty} \mathcal{A}(t) = \mathcal{A}^* > (q-p)/2q$. By $\mathcal{A}(0) = 0$, the regularity and the strict increasing monotonicity of $\mathcal{A}$, it follows that there exists a unique $\tilde{t}$ (with $\mathcal{A}(\tilde{t}) = (q-p)/2q$) such that $\mathcal{A}(t) < (q-p)/2q$ for $t < \tilde{t}$, and for $t > \tilde{t}$ we have $\mathcal{A}(t) > (q-p)/2q$. Hence $\mathfrak{S}$ attains its maximum at $\tilde{t}$.

Conversely, the existence of a unique $\tilde{t} > 0$ such that $\mathfrak{S}$ is strictly increasing on $\left[0, \tilde{t}\right]$ and then strictly decreasing implies $\mathfrak{S}'\left(\tilde{t}\right) = 0$, and from (26)

$$0 < \mathcal{A}\left(\tilde{t}\right) = \frac{q-p}{2q} \qquad \Rightarrow \qquad q > p.$$

Furthermore, $\mathcal{A}\left(\tilde{t}\right) < \mathcal{A}^*$ implies $\eta^* > b(N^*)$.

4. If $q > p$ and $\eta^* \leq b(N^*)$ then $\mathcal{A}^* \leq (q-p)/2q$, and since $\mathcal{A}(t) < \mathcal{A}^*$ it follows that $\mathfrak{S}'(t) > 0$ for every $t > 0$.

Conversely, the monotonicity of $\mathfrak{S}$ implies $\mathfrak{S}' \geq 0$ — i.e., $\mathcal{A} \leq (q-p)/2q$. From the positivity of $\mathcal{A}$ we deduce $q > p$, and passing to the limit we obtain $\mathcal{A}^* \leq (q-p)/2q$, whereby the result follows.

5. Suppose that the logistic-DIF relative sales function $\mathfrak{S}_{\max}$ peaks at a (unique) point $t_{\max}$ — i.e. $q > p$ and $\eta_{\max} > b(N^*)$, a consequence of (17). Then from item 3, $\mathfrak{S}$ attains its maximum at a point $\tilde{t}$ such that

$$\mathcal{A}_{\max}(t_{\max}) = \frac{q-p}{2q} = \mathcal{A}\left(\tilde{t}\right).$$

From Proposition 8, $\mathcal{A} \leq \mathcal{A}_{\max}$, and from the strict increasing monotonicity of $\mathcal{A}$ and $\mathcal{A}_{\max}$ it follows that $\tilde{t} > t_{\max}$. □

Some comments on Proposition 9 are in order. Item 1 says that, unlike the static case where sales eventually vanish, a positive birth rate guarantees that advertising and word-of-mouth are sufficient to support the sales as in the exponential model, although the population growth is bounded. Relative sales also behave monotonically, as in the exponential case. Recalling that as adopters grow advertising is less effective as time passes, independently of the market size, item 2 shows that a prevalence of advertising over interpersonal communication causes a decrease in relative sales. On the other hand, when word-of-mouth is stronger than advertising, from 3 and 4 the market growth rate counts. Under 3 there is a bound for the equilibrium growth rate that depends on the magnitude of p and q, and which forces sales to decrease after a certain point. Under 4 a high population growth rate combined with the prevalence of the word-of-mouth parameter supports an increase of relative sales over time. All these results are illustrated in Figure 10.

The logistic assumption is reflected in a richer series of cases in the behaviour of the second-order derivative; as for relative adoptions, this is again due to the magnitude of the difference $N^* - N_0$, as shown in Figure 11.

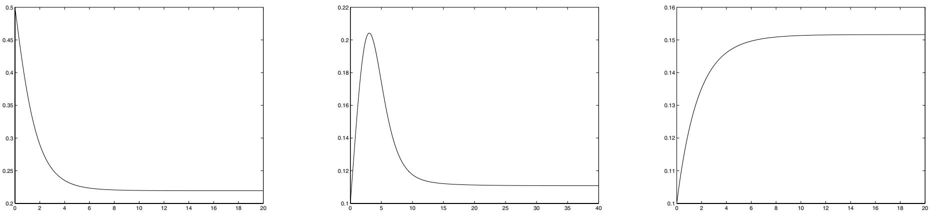


FIGURE 10. Different dynamics of the relative sales in dependence on the values determined in Proposition 9 when $N_0 = 3N^*/4$.

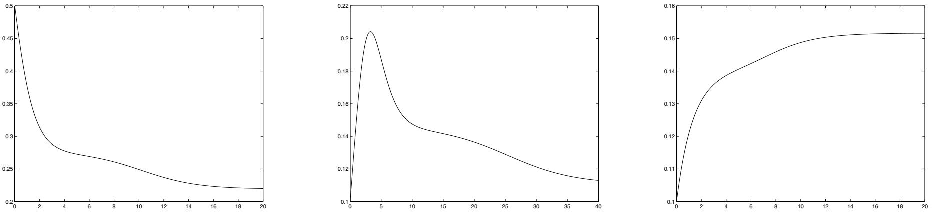


FIGURE 11. Different dynamics of the relative sales depending upon the values determined in Proposition 9 when $N_0 = N^*/80$, with other parameters as in Figure 10.

Turning now to analyse absolute sales, unfortunately only a partial study of their pattern seems possible, due to the complexity of the parameter configurations and the difficulties in treating $\mathcal{A}''$ analytically. Consequently, we illustrate various cases numerically, confining the theoretical interpretation to consequent observations. To this end, let us first note that

$$\lim_{t\to+\infty} S\,(t) = \lim_{t\to+\infty} N(t)(\mathcal{A}'\,(t) + b(N)\mathcal{A}(t)) = N^* b(N^*)\mathcal{A}^* > 0\ . \tag{27}$$

Thus contrary to the exponential model, the absolute sales do not eventually "explode", due to the limited growth of the population.

We have $S(0) = pN_0$, and $S'(0) = pN_0(b(N_0) - m(N_0) + q - p) > 0$ if and only if $q > p - (b(N_0) - m(N_0))$. Thus the absolute sales are initially strictly increasing if the internal influence is more effective than the external one decreased by a demographic term. Recalling that increasing sales are possible under the more restrictive condition $q > p$ in the Bass model, we can appreciate that the presence of a growing potential market increases the effectivity of word-of-mouth; and if this growth takes place at an initial rate $b(N_0) - m(N_0) > p$, each positive level of q initially produces increasing sales.

Moreover, from $S\,(t) = \mathfrak{S}\,(t)\,N(t)$ we have

$$S'(t) = N(t)\,[(b(N) - m(N))\mathfrak{S}\,(t) + \mathfrak{S}'\,(t)]\ ,$$

whence $S'(t) > 0$ if and only if $(b(N) - m(N))\mathfrak{S}\,(t) + \mathfrak{S}'\,(t) > 0$. Thus recalling our previous results about relative sales, we immediately conclude that the absolute sales are strictly increasing (at any time t) if $q > p$ and $\eta^* \leq b(N^*)$.

A deeper analysis of the path of absolute sales would require a more complete theoretical knowledge of the diffusion function. Nevertheless, from the previous section and the partial results just obtained, we are able to provide some insights for the linear case in Figure 12. Those on the left illustrate what happens for different parameter configurations when the market has almost reached the level N^*, and those on the right illustrate the case of a market starting far from its stable level. The different behaviour can be ascribed to the background of relative adoptions and sales already mentioned. For example, Figure 12(a) recalls the Bass model — the market is mature, hence for $p > q$ the sales decrease but do not approach zero in the long run. In Figure 12(b), there is an initial increase during the fast-growth phase of the population, despite the influence parameters. Afterwards, in the near-stationary phase, the parameters p and q again become important, but the sales also approach a steady state. Figure 12(c) shows that, after an initial decrease, the sales can increase due to a stronger contribution from the birth rate. Indeed, there is an immediate sales growth, when the population expands more rapidly (cf. Figure 12(d)).

Interpretation of the other cases may be made in the same spirit and left to the interested reader. In every case, it is found that the further the initial population is from its equilibrium value, the longer the time necessary for the sales to approach a stationary state.

5. Conclusions and future work

In this paper, we have introduced and analysed a model of diffusion of a new product into a market with logistic demographics. We have shown that in a *binomial* model (i.e., with only two segments) the introduction of a dynamic potential market makes the adoption process richer and more complex. We compared this logistic model with results obtained previously for a less realistic exponential model [4]. There remains no possibility of a saturated market, important in the dynamics of the associated relative adoptions and sales, and the adoptions and sales do not coincide due to the mortality mechanism. These aspects clearly distinguish both of these models from the original Bass model, where a fixed population was assumed.

Several important differences in the results of the exponential and the logistic models have also emerged. Firstly, excluding the logistic-DIF case considered in Proposition 4, the results for the relative adoption process in the logistic model are further enriched. Thus the variable speed of the underlying demographics, interacting with the influence process of advertising and word-of-mouth, can generate more complex dynamics. Most strikingly, there is no longer a unique peak in the relative adoption curve. For those involved in marketing decisions, an understanding of the market behaviour from both the demographic and economic standpoints is important, since it evidently determines the sales path that can be expected (cf. the interesting and complex sales dynamics shown in Figure 12).

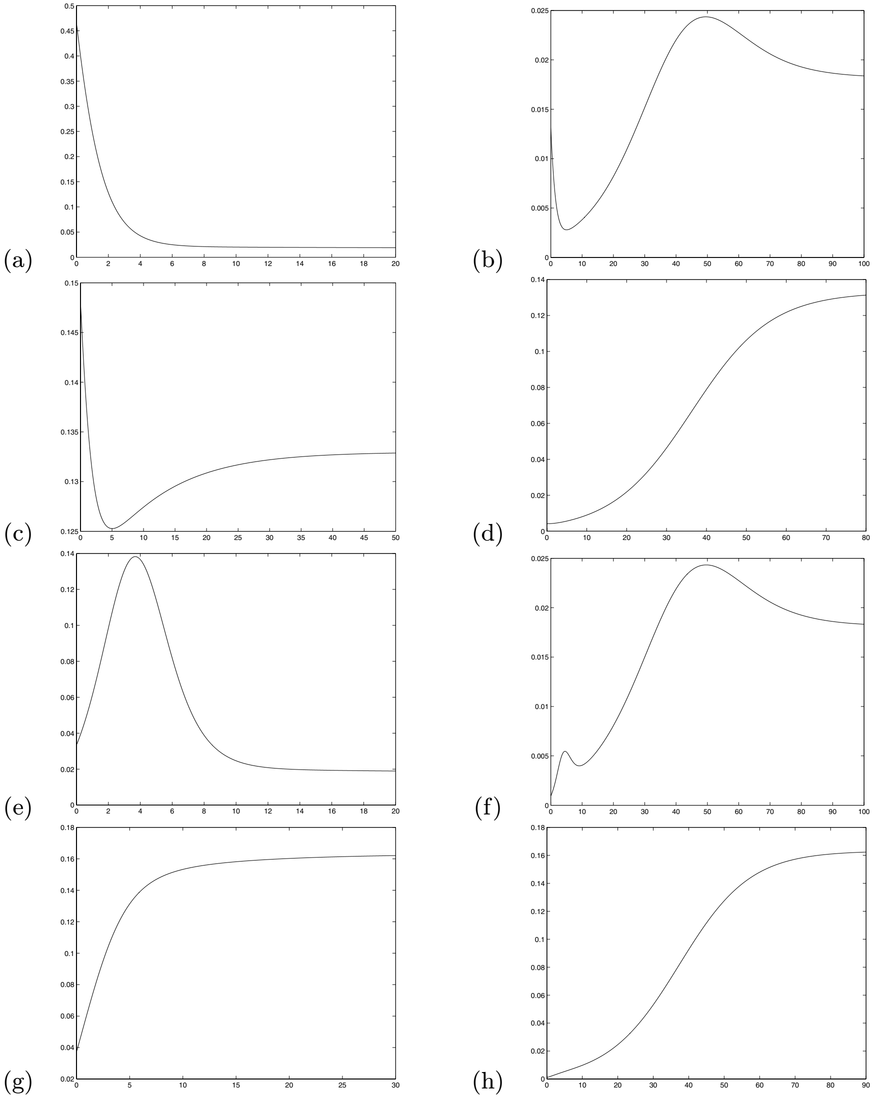


FIGURE 12. Various scenarios for absolute sales in a linear logistic market with $N_0 = 8N^*/9$ (on the left side) and $N_0 = N^*/40$ (on the right side). In all cases $k_1 = 0.1$ and $k_2 = 0.02$.
(a) and (b) $b = 0.1$, $\mu = 0.01$, $p = 0.7$, $q = 0.1$;
(c) and (d) $b = 0.6$, $\mu = 0.5$, $p = 0.2$, $q = 0.1$;
(e) and (f) $b = 0.1$, $\mu = 0.01$, $p = 0.05$, $q = 0.7$;
(g) and (h) $b = 0.6$, $\mu = 0.5$, $p = 0.05$, $q = 0.7$.

Although several aspects of our work seem similar to some modelling in the epidemiological literature (e.g. the SI models), most of the results we obtained are very different. In our approach, the stability analysis is only an introductory step, and the various solution behaviour we have found is obviously important in marketing.

This work also opens several different perspectives for investigation. Further clarification of the relationships and role of demographic and influence parameters on the adoption, diffusion and sales dynamics in our logistic model is warranted. Secondly, it would be worthwhile calibrating and testing this model on real data, although collecting data on the demographic market dynamics appears to be a demanding task (cf. also [4]). One could also consider embodying our demographic approach into some of the extensions of the Bass model introduced in the past in the stationary framework — e.g., the *polynomial* models [17], where the market is divided into more than two segments with different forms of interpersonal communication. Further compartments for example might include those who are aware of the existence of the product but have not bought it, rejectors of the new product, others who have forgotten about the product but return again to the market, or those who communicate an unfavorable judgement [6], [10], [15]. Of course, any of these possible extensions to render a more realistic model may reduce the mathematical tractability.

Acknowledgements

The authors thank Alberto d'Onofrio, Piero Manfredi and two anonymous referees for useful suggestions. Usual disclaimers apply.

References

[1] Bass, F. M., A new product growth for model consumer durables, *Management Science*, 15 (5), 215–227, (1969).

[2] Bass, F. M., and Mahajan, V., and Muller, E., New product diffusion models in marketing: a review and directions for research, *Journal of Marketing*, 54, 1–26, (1990).

[3] Birkhoff, G. and Rota G. C., Ordinary Differential Equations, Ginn and Company (1962).

[4] Centrone, F., Goia, A., Salinelli, E., Demographic processes in a model of innovation diffusion with a dynamic market, *Technological Forecasting & Social Change* 74, (2007), 247–266.

[5] Dekimpe, M., and Parker P. and Sarvary M.: Staged Estimation of International Diffusion Models. An Application to Global Cellular Telephone Adoption, *Technological Forecasting and Social Change*, 57, 105–132 (1998).

[6] Dodson, J. A., and Muller, E., Models of new product diffusion through advertising and word-of-mouth, *Management Science*, 24 (15), 1568–1578, (1978).

[7] Fourt, L. A., and Woodlock, J. W., Early prediction of market success for new grocery products, *Journal of Marketing*, 25, 31–38, (1960).

[8] Ho, T., and Savin, S. and Terwiesch, C., Managing demand and sales dynamics in new product diffusion under supply constraint, *Management Science*, 48 (2), 187–206, (2002).

[9] Kalish, S., A new product adoption model with pricing, advertising and uncertainty, *Management Science*, 31, 1569–1585 (1985).

[10] Mahajan, V., Muller, E. and Kerin, R. A., Introduction strategy for new products with positive and negative word-of-mouth, *Management Science*, 30 (12), 1389–1404, (1984).

[11] Mahajan, V., and Peterson, R.A., Innovation diffusion in a dynamic potential adopter population, *Management Science*, 24 (15), (1978).

[12] Mahajan, V., and Peterson, R. A., Erratum to "Innovation diffusion in a dynamic potential adopter population", *Management Science*, 28 (9), (1982).

[13] Mahajan, V., and Peterson, R. A., First-purchase diffusion models of new-product acceptance, *Technological Forecasting and Social Change*, 15, 127–146, (1979).

[14] Mansfield, E., Technical change and the rate of imitation, *Econometrica*, 29, 741–766, (1961).

[15] Midgley, D. F., A simple mathematical theory of innovative behavior, J. Consum. Res. 3 (1), 31–41, (1976).

[16] Sharif, M. N., and Ramanathan, K., Binomial innovation diffusion models with dynamic potential adopter population, *Technological Forecasting and Social Change*, 20, 63–87, (1981).

[17] Sharif, M. N., and Ramanathan, K., Polynomial innovation diffusion models, *Technological Forecasting and Social Change*, 21, 301–323, (1982).

[18] Thieme, H., Mathematics in population biology, Princeton University Press, Princeton, (2003).

Franscesca Centrone
Dipartimento di Scienze Economiche e Metodi Quantitativi
Universita del Piemonte Orientale
Via Perrone 18, 28100 Novara
Italy
e-mail: `francesca.centrone@eco.unipmn.it`

Ernesto Salinelli
Dipartimento di Scienze Economiche e Metodi Quantitativi
Universita del Piemonte Orientale
Via Perrone 18, 28100 Novara
Italy
e-mail: `ernesto.salinelli@eco.unipmn.it`

Aspects of Mathematical Modelling
ed. by R.J. Hosking and E. Venturino
Mathematics and Biosciences in Interaction, 363–369

A Wavelet Neural Network applied to Textile Spinning

Kanfeng F. Wang and Yongchun Zeng

Abstract. A wavelet neural network (WNN) model is applied to predict worsted yarn evenness in textile processing. A compound network predicts yarn CV (a statistic value of the yarn diameter distribution) and numbers of thin places, thick places and Neps — by analysing the spinning theory and choosing the correct input parameters. Excellent results are obtained with square correlation coefficients 0.9854, 0.9758, 0.9312 and 0.8474 for these four relevant indices, suggesting that the WNN offers a suitable control process for improved yarn spinning.

Mathematics Subject Classification (2000). Primary 99Z99; Secondary 00A00.

Keywords. Wavelet Neural Network, Yarn Unevenness, Yarn CV, Thin Places, Thick Places, Neps.

1. Introduction

In the textile industry, yarn evenness is generally recognised to be a most important property to consider in weaving, and in fabric and garment performance. Martindale [8] invoked the statistics of random processes to impose a limit on achievable yarn evenness, and found that its ideal coefficient of variation depends solely on the coefficient of variation of the fibre diameter and the average number of fibres in the cross-section. The index of irregularity I, the ratio of the measured evenness to the random limit, is of practical importance. The measured yarn evenness is always greater than the ideal yarn evenness, because the arrangement of the fibres is worse than random. Grishin [4] suggested that yarn unevenness has three contributing factors — viz. the roving irregular, ideal unevenness, and non-ideal unevenness. Many authors have considered non-ideal unevenness. Fujino et al. [3] showed that floating fibres are a dominant cause of additional yarn irregularity. Johnson [7] analysed sliver elasticity in simulating the roller-drafting of staple fibres, and suggested that it was responsible for a small degree of drafting irregularity. Lamb [9, 10] concluded that the three most commonly suggested causes of drafting irregularity were sliver elasticity, roller eccentricity and floating fibres. The factors influencing yarn evenness are so complex that it is difficult to build a universal model to take them all into consideration. Furthermore, even when using the same fibres to produce yarns of identical specification, different spinners usually produce yarns of varying quality. It is also difficult to accurately predict yarn quality in different mills using a mathematical model with descriptive rules, since machine and processing conditions vary from mill to mill. However, Neural Networks (NNs) offer an alternative for mill-specific descriptions. The highly parallel structure of NNs is also an advantage for parallel computer processing, which can lead to a better fault tolerance and faster overall processing in resolving related control processes. Moreover, the NN configuration automatically adjusts when new samples become available. Cheng and Adams [1], Maresh *et al.* [12], Pynckels *et al.* [11], Ethridge and Zhu [2, 17], and Zeng *et al.* [15] have employed NNs to successfully predict various yarn properties. This paper presents a new kind of neural network, a Wavelet Neural Network (WNN), to control yarn evenness. The structure and principles of the WNN are considered in Section 2. Important factors influencing the yarn unevenness are identified in Section 3, where appropriate input parameters are chosen accordingly. Finally, from comparison with extensive experiments as discussed in Section 4, we conclude that our WNN provides a suitable control process.

2. Wavelet Neural Network

Wavelet neural networks may be based on the theories of feed-forward neural networks and wavelet decomposition [16]. Although several theoretical studies have demonstrated the superiority of WNNs over more conventional NNs, very little has been reported on the application of WNN — especially in textile spinning.

The structure of the WNN employed in this study is shown in Figure 1. This WNN has S input nodes, T hidden nodes and only one output node. Here U and W are connecting weights, x_n is the input data, and V_n is the corresponding output, with a Morlet wavelet defined by

$$h(t) = \cos(1.75t)\exp(-t^2/2)\,. \tag{2.1}$$

The calculated output is

$$V_n = \sum_{t=1}^{T} w_t h\left(\frac{\sum u_n x_n(i) - b_t}{a_t}\right)\,, \tag{2.2}$$

where a_t is the dilation parameter and b_t is the translation parameter. Further, the objective function of the WNN is

$$E = \frac{1}{2}\sum_{n=1}^{N}(V_n^T - V_n^2)\,, \tag{2.3}$$

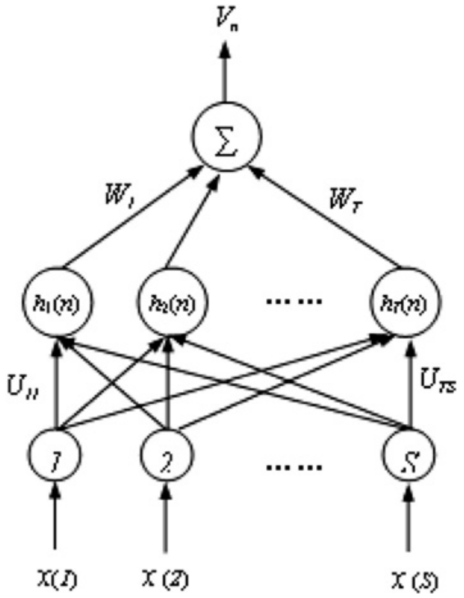


FIGURE 1. Structure of the Wavelet Neural Network.

where V_n^T is the output value corresponding to the input x_n, and the superscript T denotes the target output values.

A back-propagation algorithm used to optimise this objective is as follows:

- Initialise the dilation parameter at translation parameter b_t and connect weights u_{ti} and w_t to some random values;
- Input all training data and calculate the corresponding output V_n using equation (2.2);
- Reduce the objective function E by adjusting W, U, a and b using $\triangle w_t$, $\triangle u_{ti}$, $\triangle a_t$ and $\triangle b_t$ given by

$$\begin{aligned}\triangle w_t(j+1) &= -\eta\frac{\partial E}{\partial w_t(j)} + \alpha\triangle w_t(j)\ ,\\ \triangle u_{ti}(j+1) &= -\eta\frac{\partial E}{\partial u_{ti}(j)} + \alpha\triangle u_{ti}(j)\ ,\\ \triangle a_t(j+1) &= -\eta\frac{\partial E}{\partial a_t(j)} + \alpha\triangle a_t(j)\ .\\ \triangle b_t(j+1) &= -\eta\frac{\partial E}{\partial b_t(j)} + \alpha\triangle b_t(j)\ ,\end{aligned} \tag{2.4}$$

  where η is the learning rate and α is the momentum; and
- Return to Step 2 to provide another round of training, and continue until the network output V_n satisfies an adequate error criterion.

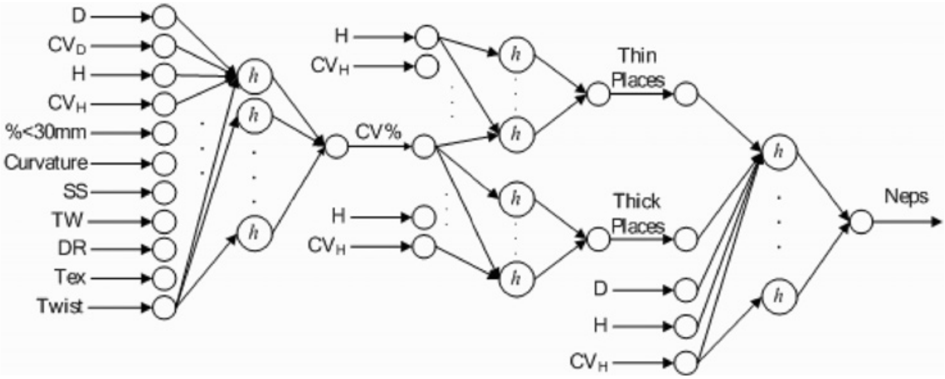


FIGURE 2. Structure of the compound WNN.

3. Input Parameters and Experimental Design

The predicted yarn unevenness is characterised by the following four indices:

1.
$$CV = \frac{\sqrt{\frac{1}{n}\sum(d-\overline{d})}}{\overline{d}} \,,$$
a statistic value of the yarn diameter distribution, where d_i is the ith measured value of the yarn diameter and $\overline{d}$ is the mean diameter;
2. Thin places, where the diameter is less than the mean diameter by 50%;
3. Thick places, where the diameter exceeds the mean diameter by 50%; and
4. Neps, the places where the diameter exceeds the mean diameter by 200%.

As already mentioned, Grishin [4] proposed that the yarn unevenness consists of three parts — viz. the roving irregular, ideal unevenness and additional or non-ideal unevenness. Thus we have

$$CV^2_{total} = CV^2_{roving} + \frac{d-1}{d} CV^2_{ideal} + CV^2_{non-ideal} \,,$$

where d is the draft ratio (DR). The ideal unevenness, determined by the Martindale [8] formula, is a function of the variation of mean fibre diameter (CV_D) and the number of fibres in the yarn cross-section. At the Commonwealth Scientific and Research Organisation (CSIRO) in Australia, it was found that the mean fibre length H, its distribution CV_H and the content of short fibre (fibres $<$ 30mm) each have a role in determining yarn unevenness. Hunter and Gee [6] concluded that, although trends are not always consistent, an increase in crimp frequency typically causes the yarn irregularity to increase. Indeed, the yarn linear density (Tex) particularly influences yarn unevenness, because there is a different fibre number in a yarn cross-section for a different yarn linear density. Further experiments at the CSIRO reported by Yang [18] also show that the yarn unevenness increases with increased spindle speed (SV). In addition, we believe that both the twist and traveller weight (TrW) affect the yarn irregularity. In summary, parameters such as D, CV_D, H, CV_H, fibre percentage < 30mm, crimp, Tex, twist, TrW, DR, and SV may all affect the yarn evenness, and should be included as inputs in the model to predict the yarn CV. Other research at the CSIRO has shown that thin places are primarily determined by yarn evenness. The mean fibre length and fibre length distribution also play a part, and that the circumstances are similar for thick places to occur. Yarn Neps are formed through fibre entanglement that occurs in scouring, carding, gilling and even combing. The mechanical settings and the condition of wires and pins are also involved. With wool fibres, the mean fibre diameter D, fibre length H and length distribution CV_H are three major factors contributing to Neps frequency in yarn. Generally, fine wool, with its less rigid fibres, produces more Neps than coarser wool [13]. Neps frequency also increases with mean fibre length and length distribution [5, 14]. Analysis of spinning trial data shows that yarn Neps are also linked to yarn evenness. Thus we use a compound WNN model to predict yarn CV, thin places, thick places, and Neps (cf. Figure 2). Our predictions involve three steps. In the first step, the yarn CV is determined by eleven parameters. Secondly, this CV value is combined with the H and CVH values to predict thin places and thick places. Thirdly, the predicted CV, thin places, thick places and D, H and CVH are used as input parameters to predict the Neps.

4. Wavelet Neural Network

From 1999 to 2003, a large-scale experiment was conducted in a Chinese top worsted mill, where 184 lots of top and yarn samples were collected and tested. Forty lots were chosen randomly as the testing set, and the rest as the training set. After training the WNN using the algorithm presented in Section 2, the testing set was input into the well-trained WNN and the predicted values obtained. Figure 3 shows the contrast between the predicted values and the measured values of the yarn CV, the number of thin places, the number of thick places, and the Neps.

It is evident that the WNN gives quite accurate results. The respective square correlation coefficients between the predicted values and measured values for the yarn CV, the number of thin places, the number of thick places, and the Neps are 0.9854, 0.9758, 0.9312 and 0.8474. However, we note that the prediction for the yarn Neps is not as good as the other three parameters — probably because the frequency of the Neps is also influenced by other factors such as the top wash, carding and machine conditions not included in this study.

5. Conclusion

The wavelet neural network (WNN) introduced in this paper successfully predicts several important properties of spun yarn.

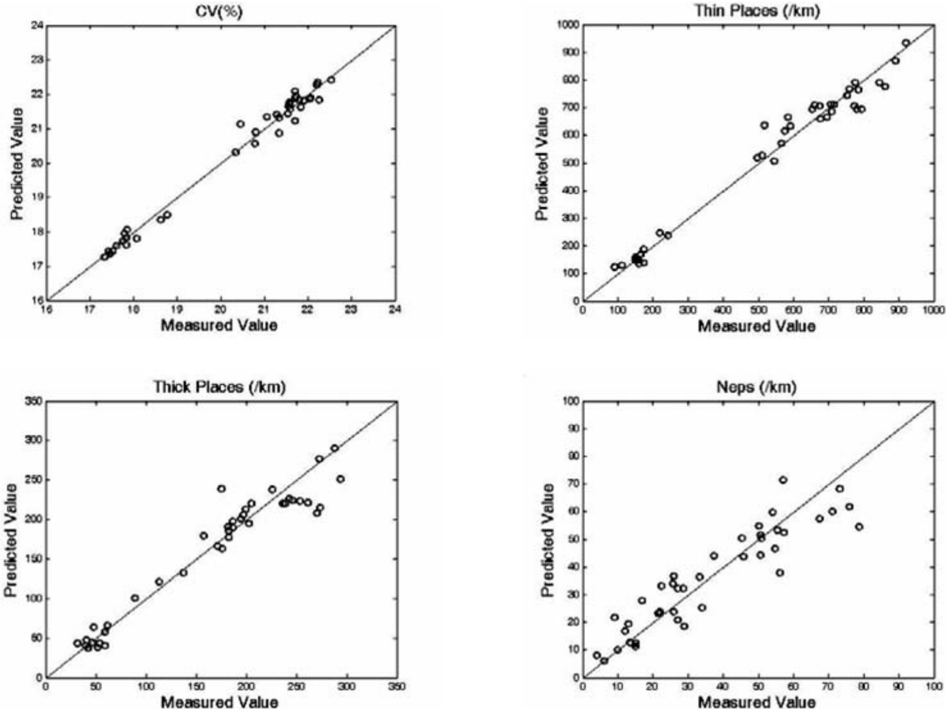


FIGURE 3. Comparison of Predicted and Measured Values.

References

[1] L. Cheng, and D. L. Adams, Yarn Strength Prediction Using Neural Networks, *Text. Res. J.* 65(9) (1995) 495–500.

[2] D. Ethridge and R. Y. Zhu, Rotor Spun Cotton Yarn Quality: A Comparison of Neural Network and Regression Algorithms, *Proc. of the Beltwide Cotton Conference* 2 (1996) 1314–1317.

[3] K. Fujino, Y. Shimotruma, and T. Fujii, A Study of Apron-drafting: Experimental Studies, *J. Text. Inst.* 68 (1977) 50–59.

[4] P. F. Grishin, A Theory of Drafting and Its Practical Applications, *J. Text. Inst.* 45(2) (1954) T167–266.

[5] L. Hunter and E. Gee, The relationship between certain yarn and fibre properties for wool worsted yarns. *Proc. 5th Int. Wool Text. Res. Conf. (Aachen)*, 1975, 259–267.

[6] L. Hunter and E. Gee, *Proc. 6th Int. Wool Text. Res. Conf. (Pretoria)*, 1980, 327–335.

[7] N. A. G. Johnson, A Computer Simulation of Drafting, *J. Text. Inst.* 72 (1981) 69–79.

[8] J. G. Martindale, New Method of Measuring the Irregularity of Yarns with Some Observations on the Origin of Irregularities in Worsted Slivers and Yarns, *J. Text. Inst.* 36 (1945) T35–T47.

[9] P. R. Lamb, The Effect of Spinning Draft on Irregularity and Faults PI: Theory and Simulation. *J. Text. Inst.* 57(2) (1987) 88–100.

[10] P. R. Lamb, The Effect of Spinning Draft on Irregularity and Faults PII: Experimental Studies, *J. Text. Inst.* 57(2) (1987) 101–111.

[11] F. Pynckels, P. Kiekens, S. Sette et al., Use of Neural Network for Determining the Spinnability of Fibres. *J. Text. Inst.* 86(3) (1995) 425–437.

[12] M. C. Maresh, R. Rajamanickam, and S. Jayaraman, Prediction of Yarn Tensile Properties using Artificial Neural Networks, *J. Text. Inst.* 86 (1995) 459–469.

[13] G. A. Robinson, M. W. Prins, and M. G. Haigh, The Importance of Entanglement and Neps. *Proc. 2nd China Int. Wool Textile Conf. (Xian)*, 1998, 98–112.

[14] D. W. F. Turpie and E. Gee, *Proc. 6th Int. Wool Text. Res. Conf. (Pretoria)*, 1980, 293–301.

[15] Y. C. Zeng, K. F. Wang, and C. W. Yu, Predicting the Tensile Properties of Air-Jet Spun Yarns. *Text.Res.J.* 74(8) (2004) 689–694.

[16] Q. Zhang and A. Benveniste, Wavelet Networks, *IEEE Transactions on neural networks* 3(6) (1992) 116–140.

[17] R. Y. Zhu and M. D. Ethridge, Predicting Hairiness for Ring and Rotor Spun Yarns and Analyzing the Impact of Fiber Properties, *Text. Res. J.* 67(9) (1997) 694–698.

[18] S. Yang, Spinning Speed Effect on Yarn Quality. CWT Report Number 106–136, 1998, CSIRO.

Kanfeng F. Wang
College of Textiles
Donghua University
1882 West Yan An Road
200021 Shanghai
P.R. China
e-mail: `kfwang@mail.dhu.edu.cn`

Yongchun Zeng
College of Textiles
Donghua University
1882 West Yan An Road
200021 Shanghai
P.R. China
e-mail: `yongchun@dhu.edu.cn`

Index