Mathematics in Industry 33
The European Consortium for Mathematics in Industry

Ewald Lindner
Alessandra Micheletti
Cláudia Nunes *Editors*

Mathematical Modelling in Real Life Problems

Case Studies from ECMI-Modelling Weeks

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Springer

Mathematics in Industry

The European Consortium for Mathematics in Industry

Volume 33

The *ECMI* subseries of the *Mathematics in Industry* series is a project of *The European Consortium for Mathematics in Industry*. *Mathematics in Industry* focuses on the research and educational aspects of mathematics used in industry and other business enterprises. Books for *Mathematics in Industry* are in the following categories: research monographs, problem-oriented multi-author collections, textbooks with a problem-oriented approach, conference proceedings. Relevance to the actual practical use of mathematics in industry is the distinguishing feature of the books in the *Mathematics in Industry* series.

More information about this subseries at http://www.springer.com/series/4651

Ewald Lindner • Alessandra Micheletti •
Cláudia Nunes
Editors

Mathematical Modelling in Real Life Problems

Case Studies from ECMI-Modelling Weeks

Springer

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Editors
Ewald Lindner
Institute of Computational Mathematics
Johannes Kepler University of Linz
Linz, Austria

Alessandra Micheletti
Department of Environmental Sciences and
Policy and Data Science Research Center
Milano
Università degli Studi di Milano
Milano, Italy

Cláudia Nunes
Center for Computational and Stochastic
Mathematics
Instituto Superior Técnico (IST),
Universidade de Lisboa
Lisboa, Portugal

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
The European Consortium for Mathematics in Industry
ISBN 978-3-030-50387-1 ISBN 978-3-030-50388-8 (eBook)
https://doi.org/10.1007/978-3-030-50388-8

Mathematics Subject Classification: 97Mxx, 97M10, 00A69

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

Preface

In recent years, numerous reports and studies have demonstrated that Mathematics is an essential tool to improve industrial innovation, and mixed academic–industrial consortia and networks, like ECMI (European Consortium for Mathematics in Industry—http://ecmi-indmath.org) and MI-Net (Mathematics in Industry Network—https://mi-network.org/) are working to foster the recognition of Mathematics as an enabling technology.

In this framework, an increasing need of mathematicians trained to work in an industrial environment has been observed and has pushed the academic world to provide novel training formats, able to respond to industrial needs. ECMI in particular has established an educational programme, offered by the ECMI Educational Centres, which is aimed to provide such training. Its main ingredients are mathematical modelling activities, in particular the International Modelling Weeks and the modelling seminars.

The International Modelling Weeks format is an international workshop where master students and/or early-career investigators (PhD students and postdocs) receive hands-on training in problem-solving, teamwork, and in learning to exploit their different skills to model efficiently non-mathematical problems. During modelling Weeks are training workshops where students from different countries spend a week working in small multinational groups on projects which are based on real-life problems. Each group is led by an instructor who introduces the problem, usually formulated in non-mathematical terms, on the first day and then helps to guide the students to a solution during the week. The students present their results to the other groups on the last day and then write up their work as a report. This format allows to train students in mathematical modelling and stimulate their collaboration and communication skills, in a multinational environment. The instructors "emulate" the figures of real industrial delegates, thus pushing the students to start working in a non-academic environment.

Modelling seminars are also offered locally by the ECMI Educational Centres. They have a structure similar to the modelling weeks, but are usually spread over one semester, and are attended only by the students enrolled in the offering university.

This book is a collection of real-world problems that have been assigned to students during the ECMI International Modelling Weeks. The problems are first described, and then a possible solution is proposed. The aim of this book is thus to provide a set of examples, in different fields of application, and faced with different mathematical techniques, to support teachers and instructors to organize future modelling activities.

Linz, Austria — Ewald Lindner
Milano, Italy — Alessandra Micheletti
Lisboa, Portugal — Cláudia Nunes
February 2020

Contents

Chapter 1
Inverse Problems in Diffuse Optical Tomography Applications

Paola Causin and Rada-Maria Weishaeupl

1.1 Introduction

Aim of this material is to provide a guideline to the modeling and fast numerical solution of the inverse problem arising in the context of Diffuse Optical Tomography (DOT), an innovative imaging technique which finds application in several clinical settings. DOT applications extend to a wide ensemble of diagnostic/monitoring purposes, ranging from cancer screening—object of this work, and in particular for breast cancer screening—to monitoring of brain function in newborns or stroke patients, to seizure detection in real time (see [11] for a comprehensive review).

Background In the seventeenth century, the French painter Georges de La Tour (1593–1652) portrayed in his work *St. Joseph the carpenter* the light of a candle transmitted through the thin fingers of the child Jesus (see Fig. 1.1). The painter's observation represents a common experience, which can be easily reproduced also with present means: if a flashlight is shone onto one's hand, it is clearly apparent that light can travel through tissue and be detected on the other side with respect to the source. This fact motivates the use of light to image the inside of the body, with the benefit of a non-invasive and non ionizing technique. Moreover, in a broader sense, physicians have always tried to diagnose health conditions from the appearance of a patient. However, the first attempts to use light as a quantitative diagnostic tool were impractical. The depth of penetration of light at visible frequencies is too low to allow for investigations in tissues/organs thicker than a few millimeters. In the context of breast screening, optical characterization was attempted in 1929

P. Causin (✉)
Department of Mathematics, University of Milano, Milano, Italy
e-mail: paola.causin@unimi.it

R.-M. Weishaeupl
Department of Environmental Science and Policy, University of Milano, Milano, Italy

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_1

Fig. 1.1 In the painting *St. Joseph the carpenter* (1640), the French artist Georges de La Tour represented the effect of candlelight crossing the thin fingers of the child Jesus (source: Wikipedia)

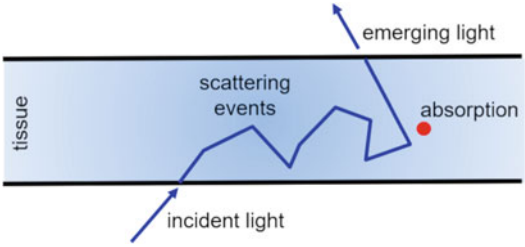

Fig. 1.2 Light crossing a biological tissue undergoes elastic scattering and, in a smaller portion, absorption. The trajectories of emerging photons are for the greatest part diffusive and not ballistic. This causes an inherent difficulty in reconstructing the optical coefficients that lead to such photon paths

by Cutler [6] who performed a "transillumination" analysis, which requires illumination of the breast on one side and examination of the shadowgraph type image viewed from an opposite side. This exam depends upon sufficient light being transmitted through the breast along straight paths. The expected result was the ability to detect vascularization and tissue patterns that could lead to the diagnosis of cancer, based on the evidence that translucence and opalescence is different for malignant and benign tissue. Despite the attempts, at the time the technique was abandoned since it was found difficult to produce the necessary light intensity without overheating the tissue. It was only in the 1970s that techniques

emerged to improve the quality of the results and the patient comfort during the examination. Optical breast imaging emerged as a novel imaging technique that uses near-infrared light (NIR, 600–900 nm) to assess optical properties of tissues. NIR wavelengths represent the so–called optical window, where absorption is minimal and thus deeper penetration in the tissue is obtained. At the same time, since the existing absorption is mainly affected by the tissue hemoglobin content and only secondarily by other components, this information can be used for diagnostic purposes. Higher absorption for carcinomas than for the surrounding tissue is indeed to be expected due to increased blood content associated with tumor-related angiogenesis [17]. However, imaging of tissue structure and function is greatly blurred by the simultaneous presence of elastic scattering, which can be 100x larger than absorption and is caused by small differences in refractive index at the microscopic level, mainly due to fat droplets and structures inside the cell nucleus [15]. Scattering makes impossible to reconstruct the deterministic path of a single photon from its emission to the exiting point (see Fig. 1.2). Several years were required to gain the necessary knowledge and technology to overcome this difficulty. Significant improvements were obtained with the use of tomographic methods, crucial for recovery of spatial information about optical properties. The combination of these technologies gave rise to the DOT approach. In DOT several challenging topics emerge: in this project we will focus on the mathematical aspects arising from the reconstruction of the optical coefficients, which represent the core of the DOT problem. To start with, we report here below the clear explanation of the DOT principle provided in the review paper by Boas et al. [2].

DOT Principle "The basic idea of DOT imaging is to illuminate the tissue with an array of light sources and to measure the light leaving the tissue with an array of detectors. For each source location, one records an image of the light reaching each detector from that particular source. A model of the propagation of light in tissue is developed and parametrized in terms of the unknown absorption and scattering as a function of position in the tissue. Then, using the model together with the ensemble of images over all the sources, one attempts to "invert" the propagation model to recover the parameters of interest, or, in other words, to estimate the optical parameters out of the data, using the model."

The present work is organized as follows: in Sect. 1.2 we introduce the mathematical model to compute light propagation in the medium with prescribed optical coefficients and we briefly discuss its derivation and properties; in Sect. 1.3 we introduce the inverse problem and its treatment with the Rytov perturbation approach, detailing its derivation and we discuss the important issue of problem regularization; in Sect. 1.5 we propose a scheme to implement the DOT algorithm in a computer code; in Sect. 1.6 we propose two numerical test cases to validate the algorithm and, eventually, in Sect. 1.7 we draw the conclusions and we propose some topics for further work.

1.2 Mathematical Models of Light Propagation in Tissue

As stated above, a model of light propagation is a necessary tool in DOT imaging. The mathematical reconstruction of the light distribution in the medium, given its optical properties, amounts to solving the so–called *forward model*. Here below we provide details of existing models and, among them, we propose to use the diffusive approximation, a reasonable balance between physical accuracy and computational cost in tomographic image reconstruction.

Beer-Lambert Law The most basic model of light propagation is obtained under the hypothesis that the only acting phenomenon is light absorption. The Beer-Lambert law is derived from an approximation for the absorption coefficient in which molecules are described by opaque disks whose cross-sectional area represents the effective area seen by an incoming photon [16]. Considering a distribution of molecules in a slab and integrating the photon balance along the propagation distance r [cm] yields

$$I = I_0 \mathrm{e}^{-\mu_a r}, \tag{1.1}$$

where I_0 is the unperturbed light intensity and μ_a $[\mathrm{cm}]^{-1}$ is the absorption coefficient representing the probability of absorption per unit length. The ratio I/I_0 deducible from Eq. (1.1) is called light transmittance and is an exponentially decreasing function of the distance r. This model is adequate to study light propagation only when the sample of biological tissue is sufficiently thin (less than a few millimeters) and a good fraction of photons with ballistic, unscattered, trajectory reaches the detectors. A conventional image reconstruction algorithm based on relation (1.1) for x-ray computer tomography can then be applied to DOT. When dealing with thicker samples "as is often the case" Eq. (1.1) fails in correctly representing light transmission and one must resort to more complex models.

The most general model for light propagation stems from the *Radiative Transfer Equation* (RTE) [1], an integro–differential equation which ensures the conservation of energy of the light radiance [$\mathrm{Wcm}^{-2}\ \mathrm{sr}^{-1}$], which is the light power per unit area traveling in a certain angular direction. Solving the RTE is computationally very intensive, so that many simplifications have been sought in literature.

Diffusion Equation A physically reasonable and cost–effective approach is to perform an expansion in spherical harmonics of the radiance in the RTE, the so–called P_N expansion. This leads to the introduction of the quantity called *photon fluence* $U = U(r, t)$ [Wcm^{-2}], which represents the integral of the radiance over the solid angle. Considering the expansion truncated to first order and under the additional hypotheses of isotropic sources and slow variation of the photon fluence, the following diffusive equation (DE) is obtained for the photon fluence (see, e.g. [8]

for a detailed derivation): find $U = U(r, t)$ such that

$$\frac{1}{c}\frac{\partial U(r,t)}{\partial t} - \nabla \cdot (D(r)\nabla U(r,t)) + \mu_a(r)U(r,t) = S(r,t), \tag{1.2}$$

where $r \in \Omega \subset \mathbb{R}^3$ is the position in the domain Ω (the tissue sample), $t > 0$ [s] is time, $S(r,t)[\mathrm{Wcm}^{-3}]$ is the volumetric source strength and c [cm/s] is the light speed. Observe that light is not physically present in the body as a volumetric source, but rather it should be represented as a flux a boundary condition in the region of application. However, the present approach presents many advantages (as will be clear in the following) and is sufficiently accurate already a few millimeters far from the source itself (see [1] for a discussion of this topic). Moreover, in (1.2) we let $D(r)$[cm] be the diffusion coefficient defined as

$$D(r) = \frac{1}{3(\mu_a(r) + \mu_s'(r))},$$

where $\mu_s'(r) = \mu_s(r)(1-g)[\mathrm{cm}]^{-1}$ is the reduced scattering coefficient, μ_s being the scattering coefficient and g an anisotropy scattering factor. Observe that, from the microscopic viewpoint, the DE modeling approach supposes that the photons move throughout the sample along random paths. Each photon travels along straight segments with a sudden discontinuity when the photon changes direction or is absorbed. The average length of rectilinear tracts is the mean free path, $\ell_{tr} \simeq 1/\mu_s$.

Equation (1.4) is endowed with the following Fresnel air–sample interface law

$$AU(\xi) + \nabla U(\xi) \cdot n = 0, \ \ \forall \xi \in \partial\Omega, \tag{1.3}$$

where n is the unit normal vector with respect to the boundary $\partial\Omega$ of Ω at point ξ. The accomodation coefficient A takes into account the differences in the refractive indices of the scattering medium and the surrounding medium and is defined as (see e.g. [8])

$$A = \frac{1 + R_{\mathrm{eff}}}{1 - R_{\mathrm{eff}}}, \quad R_{\mathrm{eff}} = -1.440n^{-2} + 0.71n^{-1} + 0.668 + 0.0636n,$$

$n = n_{in}/n_{out}$ being the ratio of the index of refraction inside (tissue) and outside (air).

For the purposes of the following discussion, the stationary case ($\partial U/\partial t = 0$) of the DE is considered. This hypothesis implies that the light distribution within the object is instantaneous: this is justified when $\mu_a \ll \mu_s'$, as in the application at hand. This latter observation also justifies the further assumption $D(r) \simeq 1/(3\mu_s'(r))$, which in the present context is reduced to $D = const$ since we will consider a homogeneous scattering coefficient throughout the domain. Then, Eq. (1.4) reduces

to: find $U = U(r)$ such that

$$-D\Delta U(r) + \mu_a(r)U(r) = S(r). \tag{1.4}$$

For a given μ_a distribution, problem (1.4) shall be solved in this project in two different contexts:

(P1) to model fluence in the points of the domain where required in the solution of the inverse problem (see next section). A fundamental solution method with Green's functions will be used in this context. This approach is chosen in order to dispose of cheap evaluations of the solution without the need of meshing the patient–specific volumetric geometry and is adequate to provide, despite the approximation, "almost real time" diagnostic results

(P2) to generate in silico substitutes of experimental data. A finite element method will be used to discretize (1.4) and fluence will be sampled in chosen locations on the boundary of the domain, simulating fluence measurements at locations of the light detectors.

1.3 Inverse Problem for the Reconstruction of the Optical Coefficients

The inverse problem constitutes the procedure in which optical coefficients are sought given a set of measurements of the fluence on the body surface. The most general approach should consist in an iterative method where at each step the unknown parameters are updated while trying to minimize an error functional depending on the difference between the measured values and the computed values. However, in practice, this procedure turns out to be exceedingly expensive for the application at hand. It is thus convenient to adopt a linearized approach as described below.

Rytov Approximation (Perturbation Approach)

Due to the nature of the present work, we deem useful to report in detail the derivation of the perturbation approach we adopt, known as Rytov approximation. Such derivation is inspired by Ishimaru [12, Vol. II, Ch. 17], with a modification to include the presence of a volumetric light source. Starting from Eq. (1.4), we suppose that the absorption coefficient μ_a may be written as $\mu_a = \mu_{a,0} + \delta\mu_a$, where $\mu_{a,0}$ is a background value and $\delta\mu_a$ a (small) perturbation term. We now let

$$U(r) = \mathrm{e}^{\psi(r)}, \tag{1.5}$$

and we write $\psi = \psi_0 + \psi_1$, where:

- the exponent ψ_0 corresponds to the homogeneous solution, i.e. it is such that $U_0 = \mathrm{e}^{\psi_0}$ satisfies the background field equation

$$-D\Delta U_0 + \mu_{a,0} U_0 = S \tag{1.6}$$

- the exponent ψ_1 represents the *logarithmic amplitude fluctuation* of the light intensity with respect to the background. As a matter of fact, since $U = U_0 \mathrm{e}^{\psi_1}$, it results $\psi_1 = \log(U/U_0)$.

Inserting in Eq. (1.4) the expression for U and the expansion for μ_a yields

$$-D\Delta \mathrm{e}^{\psi} + (\mu_{a,0} + \delta\mu_a)\mathrm{e}^{\psi} = S. \tag{1.7}$$

We now observe that

$$\Delta \mathrm{e}^{\psi} = \mathrm{e}^{\psi}[\nabla\psi \cdot \nabla\psi + \Delta\psi],$$

so that Eq. (1.7) becomes

$$-D[\nabla\psi \cdot \nabla\psi + \Delta\psi] + (\mu_{a,0} + \delta\mu_a) = S\mathrm{e}^{-\psi}, \tag{1.8}$$

and, correspondingly, Eq. (1.6) becomes

$$-D[\nabla\psi_0 \cdot \nabla\psi_0 + \Delta\psi_0] + \mu_{a,0} = S\mathrm{e}^{-\psi_0}. \tag{1.9}$$

We take now the difference of Eqs. (1.8) and (1.9), yielding

$$-[\Delta\psi_1 + 2\nabla\psi_1 \cdot \nabla\psi_0] = -\frac{\delta\mu_a}{D} + \nabla\psi_1 \cdot \nabla\psi_1 + \frac{S}{D}(\mathrm{e}^{-\psi} - \mathrm{e}^{-\psi_0}). \tag{1.10}$$

Observing that

$$\Delta(U_0\psi_1) = \psi_1 \Delta U_0 + 2U_0 \nabla\psi_1 \cdot \nabla\psi_0 + U_0 \Delta\psi_1 \tag{1.11}$$

we may rewrite the left hand side of Eq. (1.10) as

$$-\frac{1}{U_0}[\Delta\psi_1 + 2\nabla\psi_1 \cdot \nabla\psi_0] = -\frac{1}{U_0}[\Delta(U_0\psi_1) - \psi_1 \Delta U_0] = \\ -\frac{1}{U_0}[\Delta - \frac{\mu_a}{D}](U_0\psi_1) - \frac{\psi_1}{U_0}\frac{S}{D}, \tag{1.12}$$

where we have also used the fact that U_0 solves problem (1.6). We eventually obtain the following equation for $(U_0\psi_1)$

$$[\Delta - \frac{\mu_a}{D}](U_0\psi_1) = \left(\frac{\delta\mu_a}{D} - \nabla\psi_1 \cdot \nabla\psi_1\right) U_0, \tag{1.13}$$

where we have used the approximation (valid for small ψ_1 and recalling that $U_0 = e^{\psi_0}$)

$$\frac{S}{D}\psi_1 + U_0\frac{S}{D}\left(e^{-\psi} - e^{-\psi_0}\right) = \frac{S}{D}\psi_1 + \frac{S}{D}(U_0/e^{\psi_0})\left(e^{-\psi_1} - 1\right) \quad \approx$$
$$\frac{S}{D}\psi_1 + \frac{S}{D}(1 - \psi_1 - 1) = 0.$$

Equation (1.13) is of modified Helmholtz type and its solution can be expressed as the convolution integral

$$U_0\psi_1 = \int_\Omega G(r - r')\left(\frac{\delta\mu_a}{D} - \nabla\psi_1 \cdot \nabla\psi_1\right) U_0\, dr', \tag{1.14}$$

where $G(r - r')$ is the Green's function for the operator $[\Delta - \frac{\mu_a}{D}]$ (see Appendix for details). Neglecting the second order term in $\nabla\psi_1$ and dividing by U_0, we obtain the following expression for $\psi_{1,0} \sim \psi_1$

$$\psi_{1,0}(r) = \frac{1}{U_0(r)}\int_\Omega G(r - r')\frac{\delta\mu_a(r')}{D}U_0(r')\, dr', \tag{1.15}$$

where we have made explicit all the dependencies on the integration and position variables. Gathering Eq. (1.15) and the definition of ψ_1, we eventually obtain

$$\frac{1}{U_0(r)}\int_\Omega G(r - r')\frac{\delta\mu_a(r')}{D}U_0(r')\, dr' = \log\frac{U(r)}{U_0(r)}. \tag{1.16}$$

Important Relation (1.16) is the core of the inverse problem: the left-hand side is built according to the DE model of light propagation (this corresponds to its use as detailed in P1), while the right-hand side is measured experimentally recording the transmitted light in the perturbed and homogeneous field, respectively. The inverse problem consists in finding the perturbation field $\delta\mu_a = \delta\mu_a(r)$ such that Eq. (1.16) holds (in a mathematical sense that we will specify later on).

1.4 Numerical Approximation of the Inverse Problem

To solve the inverse problem, we start from Eq. (1.16) and we recall that in the tomographic configuration light is successively generated by different sources and collected by the detectors (which lay on the boundary of the domain). We suppose thus to dispose of M experimental measurements, that is, M source–detector couples with $M = N_s \times N_d$, N_s being the number of light sources and N_d the number of detectors. We choose successively in (1.16) r as the detector positions $r_d^l, l = 1, \ldots, N_d$ and we consider in turn each light source of position

$r_s^k, k = 1, \ldots, N_s$. Then, for source–detector couple, we discretize the integral by breaking up the domain into N discrete volume elements (called voxels). The size of each voxel may be variable, depending on the local thickness of the sample, even if in a first approximation simple equally–sized square (cubic) voxels are an acceptable choice. Using a midpoint quadrature rule (with voxel centroid r_j, voxel volume ΔV_j, $j = 1, \ldots, N$), we then obtain

$$\sum_{j=1}^{N} \frac{\Delta V_j}{U_0(r_s, r_d)} G(r_d - r_j) \frac{\delta\mu_a(r_j)}{D} U_0(r_s, r_j) \simeq \log \frac{U(r_s, r_d)}{U_0(r_s, r_d)} \tag{1.17}$$

where we have made explicit all the dependencies. Relation (1.17) is conveniently written in matrix form by introducing the *sensitivity matrix* $J \in \mathbb{R}^{(M \times N)}$

$$J = J_{ij} = \left[\frac{\Delta V_j}{U_0(r_s^k, r_d^l)} G(r_d^l - r_j) \frac{1}{D} U_0(r_s^k, r_j) \right], \tag{1.18}$$

and the right-hand side $y \in \mathbb{R}^{(M \times 1)}$

$$y = y_i = \log \frac{U(r_s^k, r_d^l)}{U_0(r_s^k, r_d^l)}, \tag{1.19}$$

where the index i stands for the detector/source pair, i.e. $i = 1, \ldots, M \rightarrow \{l, k\}$, $l = 1, \ldots, N_d, k = 1, \ldots, N_s$ and the index j for the voxel number, $j = 1, \ldots, N$. Letting (with a slight abuse of notation) $\delta\mu_a \in \mathbb{R}^{(N \times 1)}$ be the vector of unknown variations in absorption coefficient, we then obtain

$$J\delta\mu_a = y. \tag{1.20}$$

System (1.20) can be under or over–determined, depending on the fact that M is greater or smaller than N. This aspect is strictly related to the specific DOT configuration one considers. In applications where light is collected via a (limited number) of optical fibers the first case is obtained, while the use of CCD cameras to collect emerging light leads to the second case (and possibly to $M >> N$, even after image quality check procedures and saturated pixel pruning). Let us focus on the case $M > N$: system (1.20) is overdetermined and has to be solved in the least square sense. Namely, one has to minimize the residual norm (here $\|x\|_2^2 = \sum_i |x_i|^2$ denotes the l_2-norm of the vector x)

$$\min_{\delta\mu_a} \mathscr{L}(\delta\mu_a) = \min_{\delta\mu_a} ||y - J\delta\mu_a||_2^2, \tag{1.21}$$

which represents the discrepancy between modeled data and measured data. We can formally derive the minimizer for the functional $\mathscr{L}(\delta\mu_a)$ by performing the Fréchet

derivative, and setting it to zero

$$\frac{d}{d\varepsilon}\mathscr{L}(\delta\mu_a + \varepsilon v)\bigg|_{\varepsilon=0} = \frac{d}{d\varepsilon}\|y - J(\delta\mu_a + \varepsilon v)\|_2^2\bigg|_{\varepsilon=0} = 0, \tag{1.22}$$

where v is a vector that prescribes the "direction" in which the derivative is computed, and (1.22) should be valid for all directions. The minimizer satisfies the system:

$$[J^T J]\delta\mu_a = J^T y. \tag{1.23}$$

Due to the nature of J, the matrix $[J^T J]$ is badly conditioned (close to singular). The ill-conditioning implies that a standard numerical method for solving linear systems will produce an unacceptably large error. A regularization procedure must thus be carried out.

An Ill–Conditioned System To have an idea of the degree of ill-conditioning of the system, you may want to compute and plot the singular values of J for one of the simulation tests suggested below. To do this, in Matlab®, you can use the command `svd`. Notice that:

- the singular values of J gradually decay to zero (without really being zero)
- the singular values span a range of several orders of magnitude.

Since the condition number of a rectangular matrix can be related to the ratio between the largest and smallest singular value, it appears that it is very large for the DOT inverse problem.

In order to regularize problem (1.23) a robust and commonly used approach is to perform a Tikhonov regulation [10]. Its purpose is to dampen or filter out the contributions from the smallest singular values of J. The idea of Tikhonov regularization is to add a penalization term to the problem:

$$\min_{\delta\mu_a} \mathscr{L}^R(\delta\mu_a) = \min_{\delta\boldsymbol{\mu}_a}\left(\|y - J\delta\mu_a\|_2^2 + \lambda\|\delta\mu_a\|_2^2\right). \tag{1.24}$$

Again one can derive the equation for the minimizer of the regularized functional $\mathscr{L}^R(\delta\mu_a)$ and obtain the regularized system:

$$(J^T J + \lambda I)\delta\mu_a^R = J^T y. \tag{1.25}$$

The regularization parameter λ controls the weight given to minimization of the side constraint relative to minimization of the residual norm. Its choice is not trivial and several approaches exist (see [10] for more details).

Choice of λ and numerical solution of the regularized system To start with, we suggest to simply set λ to a constant. We have found that in several test cases $\lambda = 10^{-3}$ is a reasonable value. However, you may want to play around a bit and see

the effect of choosing different values. Then, you can solve system (1.25) by a linear solver of your choice. Notice that more sophisticated regularization approaches exist, of which we provide some details in Sect. 1.7.

1.5 Details of the Computer Implementation

Many programming languages can be used to implement a computer code for the solution of the DOT problem and to construct in silico substitutes of the measured data (these latter are strongly suggested over "real" experimental data to start with!). We propose here to use Matlab®, in order to take advantage of its several built-in algorithms and functions in addition to its efficient linear systems solver. The following steps must be carried out (you may want to refer to Fig. 1.3 for an example in 2D):

(I) first, you need to implement the solution of the forward problem (as detailed in point P2 of Sect. 1.2) to generate the in silico data for photon fluence. It is convenient (but not compulsory!) to use the Matlab® PDE Solver. To do this, you must perform the following steps:

- I.1 define and mesh the domain geometry
- I.2 define the position of sources and detectors
- I.3 define the background optical coefficients. You can use literature values as suggested in Sect. 1.6
- I.4 when computing the perturbed solution, define the position and shape of inclusions (see Sect. 1.6 for details); define their optical coefficients
- I.5 define sources as Gaussian bells centered around the application point. The intensity of the source is not relevant since you will perform a normalization (see Eq. (1.19)). You can play a bit with the variance of the Gaussian (something of the order of 0.01 usually works fine)
- I.6 enforce boundary conditions (see Eq. (1.3) and, more specifically, Eq. (1.26))
- I.7 solve the problem with finite elements. Matrix assembly and system solution is automatically performed by the Matlab® PDE solver
- I.8 collect the values of the solution (fluence) at the detector locations.

You must execute the above procedure twice: once with an homogeneous background value for the absorption coefficient (corresponding to $\mu_{a,0}$), and once with the presence of inclusions (corresponding to perturbation $\delta\mu_a$ in some regions of the domain). Save the respective data on two separate files, along with the locations of sources and detectors and background optical coefficients

(II) you now have to implement the solution of the inverse problem by performing the following steps (in this phase you must pretend you do not know the

location and absorption strength of the inclusions and you must reconstruct them!):

- II.1 read the files you saved in the previous point
- II.2 build the voxelization of the domain (see Fig. 1.3(right) for an example); compute the voxel centroids and volumes
- II.3 build the sensitivity matrix (a delicate point for code efficiency!). To do this:
  - i. write a function which, given source position and evaluation point, returns the value of the corresponding Green's function for the modified Helmholtz problem (free-space and dipole versions, respectively, refer to the Appendix)
  - ii. build the sensitivity matrix using Eq. (1.18). Pay attention: it is convenient to compute the value U_0 at the numerator with dipole Green's function expansion and use instead data generated at point (I) for the value U_0 at the denominator. In addition, use for the Green's function appearing at the numerator the free-space version
- II.4 build the right-hand side y by normalizing corresponding perturbed data over background data read from the files
- II.5 solve the resulting overdetermined system for the variation of the absorption coefficient by Tikhonov regularization (see Sect. 1.3). To solve the regularized system you can simply use the \ command of Matlab®

(III) visualize the results representing the solution value in each voxel and superpose to the plot the exact position of the inclusions. This procedure is known as "mapping" of the results. To perform it, you can use the Matlab® command `patch`. Use a colorbar to check the magnitude of the values!

1.6 Examples of Numerical Simulations

To assess your implementation, we suggest to consider a 2D domain represented by a semicircle of radius 4 cm (see Fig. 1.3(left)). We consider a set of $N_s = 19$ pointwise light sources uniformly located on the horizontal boundary at a distance

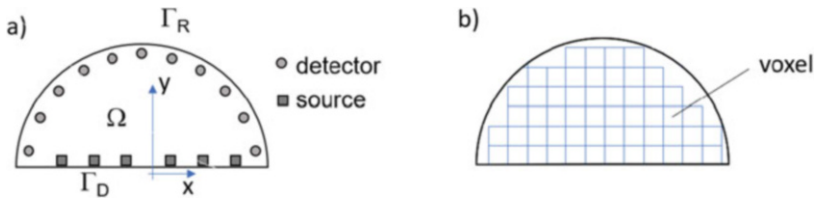

Fig. 1.3 (**a**) Common geometry and source/detector positioning for the numerical experiments; (**b**) discretization of the domain in voxels

of 1 mm Γ_D inside the domain and a set of $N_d = 200$ detectors radially disposed 1 mm inside the domain along the boundary Γ_R.

Generate fluence "observations" in the detectors by using your code implemented as in point (I) of Sect. 1.2. We advise to use a fine enough mesh, for example we used a mesh consisting in 37,858 triangles, with an approximation of degree 2 on each element.

Set the following boundary conditions (slightly different from the general ones given in Eq. (1.3)):

$$\begin{aligned} U &= 0 && \text{on } \partial\Gamma_D, \\ U + 2AD\frac{\partial U}{\partial n} &= 0 && \text{on } \Gamma_R, \end{aligned} \tag{1.26}$$

where $\partial\Omega = \Gamma_D \cup \Gamma_R$, n is the outward normal vector. Notice that the first condition represents the fact light does not escape from the solid plate. In all the numerical tests, we adopted the following physically reasonable background values (see e.g. [14]): $\mu_s' = 0.1\ [\text{cm}^{-1}]$, $\mu_{a,0} = 0.01\ [\text{cm}^{-1}]$. Moreover, we used $N = 588$ voxels, which correspond to a spatial resolution of 0.25 cm (see Fig. 1.3b for an example of voxelization).

Perform the Following Studies

- Test case 1
  Consider a single circular inclusion placed at $x = 2, y = 2$ cm with radius 0.3 cm. Set the absorption coefficient of the inclusion to $\mu_{a,inc} = 2 \times \mu_{a,0}$ (see Fig. 1.4a).
- Test case 2
  Place two circular inclusions at $x = -2.5$, $y = 2.5$ cm and $x = 2$, $y = 1$ cm, respectively, both with radius 0.3 cm. Set the absorption coefficient of the leftmost inclusion to $\mu_{a,inc} = 5 \times \mu_{a,0}$, while for the rightmost to $\mu_{a,inc} = 2 \times \mu_{a,0}$ (see Fig. 1.5a).

The results we obtained for test case 1 are shown in Fig. 1.4b, while the results for test case 2 are shown in Fig. 1.5b. Observe that we correctly estimate the differential absorption coefficient of the inclusion(s) and we localize their positions, but the

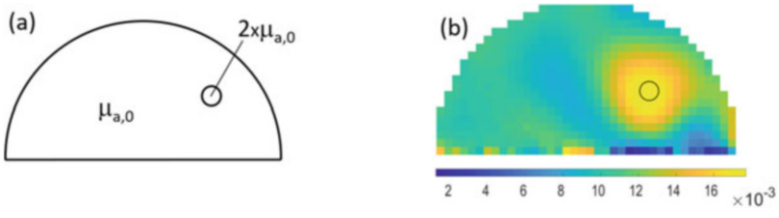

Fig. 1.4 Test case 1: (**a**) setting; (**b**) reconstructed absorption coefficient

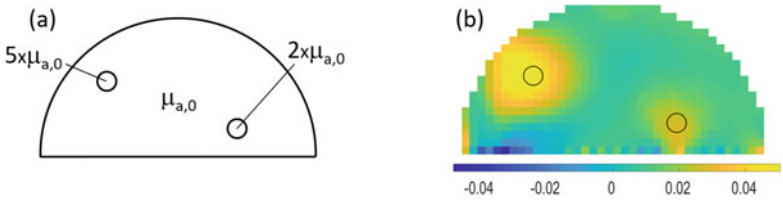

Fig. 1.5 Test case 2: (**a**) setting; (**b**) reconstructed absorption coefficient

affected region is largely over–estimated. This effect is in large part due to the use of the ℓ_2-norm regularization (see the next section for further comments). Moreover, in the case with two inclusions, the value of the weaker one is not computed with high precision.

1.7 Conclusions and Ideas for Further Work

This work was aimed at offering a "gentle introduction" to some of the mathematical and numerical issues related to DOT technology. This is a very promising field, capable of offering a fast, cost–effective and unharmful screening tool. However, DOT resolution is a mathematically challenging problem. Here, we have provided a self-contained material to carry out a complete DOT solution in a simplified setting. The results obtained with the techniques that we presented allow to obtain an estimation of the position and strength of the inclusion, but much space is left to improvements. For example, as you can observe from the rightmost plots in Figs. 1.4 and 1.5 that the position of the estimated variation in the absorption coefficient is not exactly coincident with the true one, as well as its spatial extension. This latter results to be somewhat blurred, a typical effect of the Tikhonov regularization. To obtain better and sharper results, more sophisticated tools are required. We suggest to consult the different literature contributions in this field, for example see [3–5, 8, 9, 13]. In addition, we list below a series of points that represent issues that deserve further attention in a deeper analysis of the problem.

Further Work You may want to investigate the following points (some of them not trivial at all!):

- do the results change using a finer voxelization? Is there a good compromise? Try to use, instead of a voxelization as suggested here, a Delaunay triangular mesh. How do the results change?
- what is the role of the regularization parameter? Can we find an automatic way to establish a good value for it? To start with, you can explore the possibilities offered in the `regtool` suite which is a freely available package for regularization in Matlab®
- try to add some noise to the data (for example 1–5% of the local value, with an independent Gaussian distribution at each detector). What is the effect?

Appendix: Green's Function Solution of the DE Model

The DE model and its successive manipulations in the Rytov procedure yield relation that represent instances of the inhomogeneous modified Helmholtz equation of the form:

$$[\Delta - \alpha^2]\phi(r) = f(r). \tag{1.27}$$

We are interested in finding an analytic solution to this partial differential equation. It is clear that this will be possible only on simple geometries. In particular, we will refer to the solution on a infinite or semi–infinite domain and we will "pretend" that it can be used as it is for our finite domain. With this aim, for a linear operator $\mathscr{L}$, we introduce the Green's function $G = G(r - r')$, such that

$$\mathscr{L}G(r - r') = \delta(r - r')$$

where δ is the Dirac delta centered in r'. It holds the following (see e.g., [7])

Lemma 1.1 *Given the partial differential equation $\mathscr{L}\phi(r) = f(r)$, $r \in \Omega$, if $G(r - r')$ is the Green's function with respect to the linear partial differential operator $\mathscr{L}$, then a solution to the PDE is given by the convolution between the source term f and the Green's function*

$$\phi(r) = \int_{\Omega} G(r - r')\, f(r')\, dr' \tag{1.28}$$

The Green's function for our operator at hand $\mathscr{L} = (\Delta - \alpha^2)$ has the following expression for $n = 2$ or 3 (see e.g. [7])

$$G(r - r') = \begin{cases} -\dfrac{1}{2\pi} K_0(\alpha||r - r'||) & n = 2, \\ -\dfrac{1}{4\pi} \dfrac{\mathrm{e}^{-\alpha||r-r'||}}{||r - r'||} & n = 3, \end{cases} \tag{1.29}$$

where K_0 is the modified Bessel function of the second kind of order zero (in Matlab® you can compute it with the command `besselk`). Observe that the Green's functions in (1.29) refer to an infinite domain and only satisfy the radiation condition $|G| \to 0$ for $|r| \to \infty$. As a result, the corresponding solution obtained from the convolution procedure does not satisfy the proper boundary conditions on the finite domain Ω. In order to partially correct this fact, you can use the so–called dipole approximation which allows to enforce null solution over a plane, which is a convenient condition in our case for the part of the breast laying on the solid plate. To fix ideas, let us think this plane to correspond to $z = 0$. When a point source is placed at a small depth ℓ into the sample, an equivalent opposite "sink" is placed at

$-\ell$. The following dipole solution is obtained for a source located in $x' = (0, 0, \ell)$

$$G_D(x, x') = G(x, (0, 0, \ell) - G(x, (0, 0, -\ell)).$$

This expression guarantees that $\phi = 0$ on the plane.

References

1. Arridge, S.R., Hebden, J.C.: Optical imaging in medicine: II. modelling and reconstruction. Phys. Med. Biol. **42**(5), 841 (1997)
2. Boas, D.A., Brooks, D.H., Miller, E.L., DiMarzio, C.A., Kilmer, M., Gaudette, R.J., Zhang, Q.: Imaging the body with diffuse optical tomography. IEEE Signal Proc. Mag. **18**(6), 57–75 (2001)
3. Cao, N., Nehorai, A., Jacob, M.: Image reconstruction for diffuse optical tomography using sparsity regularization and expectation-maximization algorithm. Opt. express **15**(21), 13695–13708 (2007)
4. Causin, P., Naldi, G., Weishaeupl, R.M.: Elastic net regularization in Diffuse Optical Tomography applications. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice (to appear, 2019)
5. Choe, R.: Diffuse optical tomography and spectroscopy of breast cancer and fetal brain. Ph.D. thesis, University of Pennsylvania (2005)
6. Cutler, M.: Transillumination of the breast. Ann. Surg. **93**(1), 223 (1931)
7. Duffy, D.G.: Green's Functions with Applications. Chapman and Hall/CRC (2018)
8. Durduran, T., Choe, R., Baker, W., Yodh, A.G.: Diffuse optics for tissue monitoring and tomography. Rep. Prog. Phys. **73**(7), 076701 (2010)
9. Gibson, A., Hebden, J., Arridge, S.R.: Recent advances in diffuse optical imaging. Phys. Med. Biol. **50**(4), R1 (2005)
10. Hansen, P.C.: Regularization tools, a matlab package for analysis and solution of discrete ill-posed problems (2008). Http://www.imm.dtu.dk/~pcha/Regutools/
11. Hoshi, Y., Yamada, Y.: Overview of diffuse optical tomography and its clinical applications. J. Biomed. Opt. **21**(9), 091312 (2016)
12. Ishimaru, A.: Wave propagation and scattering in random media, vol. 2. Academic press New York (1978)
13. Konovalov, A.B., Genina, E.A., Bashkatov, A.N.: Diffuse optical mammotomography: state-of-the-art and prospects. "J. Biomed. Photonics Eng. **2**(2) (2016)
14. Sun, Z., Wang, Y., Jia, K., Feng, J.: Comprehensive study of methods for automatic choice of regularization parameter for diffuse optical tomography. Optical Eng. **56**(4), 041310 (2016)
15. Swartling, J., Andersson-Engels, S.: Optical mammography–a new method for breast cancer detection using ultra-short laser pulses. DOPS-NYT p. 19 (2001)
16. Swinehart, D.: The Beer-Lambert law. J. Chem. Educ. **39**(7), 333 (1962)
17. Taroni, P.: Diffuse optical imaging and spectroscopy of the breast: a brief outline of history and perspectives. Photochem. Photobiol. Sci. **11**(2), 241–250 (2012)

Chapter 2
1D Models for Blood Flow in Arteries

Alexandra Bugalho de Moura

2.1 Introduction

Cardiovascular diseases remain one of the major causes of death in developed countries, with great social and economic impact. The simulation of the cardiovascular system helps understanding the physiology of blood circulation and enables non-invasive based clinical predictions. In the past years a large research activity has been devoted to complex 3D models of blood flow, using patient-specific cardiovascular geometries obtained through medical imaging, see [1–5] for some examples. The 3D simulations provide great detail on the blood flow patterns and allow to quantify a number o clinical indices. However, in many situations, the detailed information of the 3D model is not crucial, and the analysis of average quantities, such as flow rate and pressure, suffices to make clinical predictions and decisions [6, 7]. One of the features of blood circulation that is best captured by 1D simplified models in large arterial networks is its pulsatility. The elastic deformations in large arteries, such as the aorta or the carotid, are very important, helping to regularize blood flow during the cardiac cycle and leading to the pulse propagation that characterizes the arterial tree. This pulsation feature of blood flow in arteries has been observed and used in medical practices for hundreds of years. For example, the superposition of the waves reflected by medical devices, such as prosthesis or stents, with those produced by the heart can generate anomalous pressure peaks [8].

Several approaches can be followed to derive 1D models for blood flow in arteries, and different 1D models can be obtained depending on the level of

A. B. de Moura (✉)
REM—Research in Economics and Mathematics; CEMAPRE—Center for Applied Mathematics and Economics, ISEG—Lisbon School of Economics and Management, Universidade de Lisboa, Lisbon, Portugal
e-mail: amoura@iseg.ulisboa.pt

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_2

simplification and on the characteristics of blood circulation kept during the simplification process. Here, the 1D model is derived by integrating the 3D Navier-Stokes equations for fluid flow coupled with a model for the vessel compliance, considering some simplifying assumptions [8, 9]. The resulting mathematical model consists of an hyperbolic system of partial differential equations (PDE's). This means that it has *wave-like* solutions, with characteristic propagation speed and wave length. The numerical discretization of the 1D hyperbolic model is briefly discussed and numerical results are presented by considering an application to the study of blood circulation in the human brain. The purpose of the application here introduced is to answer the question "What are the effects of anatomical changes in the main arteries of the arterial system of the human brain?". Regarding this subject and other clinical applications of 1D blood flow models, see for instance [6–8].

Following the same methodology of integrating over a generic axial section, more complex 1D models are derived, namely accounting for vessel curvature. This is achieved by relaxation of some simplifying assumptions. The inclusion of curvature means more complexity on the model, and the resulting system of PDEs reflects that extra complexity. In this context we discuss on the balance between simplicity and accuracy when doing mathematical modeling.

2.2 The 1D Model for Blood Flow in Arteries

The 1D model dates back to Euler [10], that already in 1775 introduced a 1D model of the human arterial system, yet claiming "the incredible difficulties for its solution". Here the 1D model for blood flow in arteries is derived from the 3D model. We start by considering the Navier-Stokes equations for Newtonian incompressible fluids [11]:

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} + \dfrac{1}{\rho} \nabla P - \nu \Delta \mathbf{u} = \mathbf{f}, \\ \operatorname{div} \mathbf{u} = 0, \end{cases} \quad \text{in } \Omega \tag{2.1}$$

where the unknowns are the fluid velocity $\mathbf{u} = (u_x, u_y, u_z)$ and pressure P, depending on space $\mathbf{x} = (x, y, z)$ and time t, with Ω the 3D vascular district of interest (see Fig. 2.1). Here $\mathbf{f}$ is a given function representing the volume forces exerted on the fluid, as e.g. the gravity, ρ is the constant blood density, and ν is the constant blood viscosity. We will neglect body forces, $\mathbf{f} = \mathbf{0}$.

The first equation of (2.1) describes the momentum conservation, while the second is the continuity equation and represents the conservation of mass.

The Navier-Stokes equations (2.1) are coupled with a model for the vessel wall displacement. Due to their complex structure, it is very difficult to devise appropriate and accurate models describing the mechanical behaviour of the artery walls. We will not go into detail on this subject, but we will consider that the walls of the vessel can move as the result of the fluid pressure. Equations (2.1), together with a

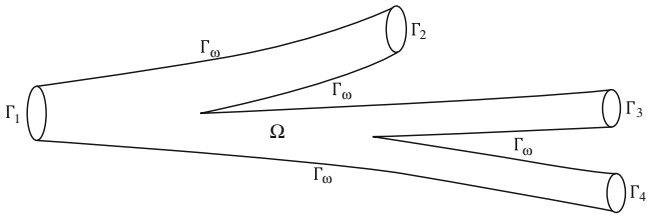

Fig. 2.1 Generic vascular district Ω

model for the vessel wall, constitute a 3D FSI (fluid-structure interaction) model for blood flow in vascular districts.

2.2.1 Deriving the 1D Model

To derive the 1D model, we assume some simplifying hypothesis and then we integrate the 3D FSI model in each cross section $S(z)$ of the vessel [9]. Applying this procedure, the only spatial coordinate remaining is the axial direction, denoted z. The simplifying assumptions are as follows:

H1 *Axial symmetry*. All quantities are independent from the angular coordinate, implying that each axial section $S(z)$ remains circular during the wall motion. Hence, the tube radius R is a function of time t and axial direction z, $R = R(t, z)$.

H2 *Radial displacements*. Wall displacements occur only in the radial direction. Defining R_0 as the reference radius, the wall displacement is $\mathbf{d} = d\mathbf{e}_r$, with $\mathbf{e}_r$ the outward unit vector in the radial direction and $d(t, z) = R(t, z) - R_0(t, z)$. The reference radius R_0, usually the radius of the vessel at rest, may depend on the axial direction z. Indeed, one characteristic of arteries is its tapering geometry.

H3 *Fixed cylinder axis*. The axial axis is fixed in time and the vessel expands and contracts around it.

H4 *Constant pressure on each axial section*. Pressure is assumed constant in each section, depending only on z and t, $P = P(t, z)$. This is reasonable, since the pressure field of the fluid flow in 3D straight tubes is mainly constant in each section.

H5 *No body forces*. External forces are neglected. This is often considered already at the 3D model level.

H6 *Axial velocity dominance*. The velocity components orthogonal to the axial direction are neglected, since they are considered negligible when compared to the axial velocity: $\mathbf{u} = u_z$. In cylindrical coordinates we have $\mathbf{u} = (u_r, u_\theta, u_z)$

and

$$u_z(t,r,\theta,z) = u_z(t,r,z) = \bar{u}(t,z) \times s\left(\frac{r}{R(t,z)}\right) \tag{2.2}$$

where $s(\cdot)$ is the velocity profile, assumed constant in time, t, and space, z, which is in fact in contrast with the observations and 3D models. In this simplifying setting, $s(\cdot)$ may be though of as a profile representative of an average flow configuration. In practice, it will be considered flat or parabolic.

The unknown variables of the 1D model will be averaged quantities. The area is related with wall displacement and is given by $A(t,z) = \int_{S(t,z)} ds = \pi R^2(t,z)$; the flow rate, $Q(t,z) = \int_{S(t,z)} u_z ds = A(t,z)\bar{u}(t,z)$, and mean velocity, $\bar{u}(t,z) = \frac{1}{A(t,z)} \int_{S(t,z)} u_z ds = \frac{Q(t,z)}{A(t,z)}$, are related with fluid velocity. Due to H4, the mean pressure is $\bar{p}(t,z) = \int_{S(t,z)} P(t,z) ds = P(t,z)A(t,z)$. All these quantities depend on t and z. In the notation, we will usually omit, unless needed, this dependence. From H6, H1, and (2.2) we have

$$\bar{u}(t,z) = \frac{1}{\pi R^2} \int_0^{2\pi} \int_0^R \bar{u}(t,z) s\left(\frac{r}{R}\right) r dr d\theta = \frac{\bar{u}(t,z)}{\pi R^2} 2\pi \int_0^R s\left(\frac{r}{R}\right) r dr$$

meaning that $\frac{R^2}{2} = \int_0^R s\left(\frac{r}{R}\right) r dr$, and $\int_0^1 s(y) y dy = 0.5$ by doing the change of variable $r = Ry$. We also define the momentum-flux coefficient or Coriolis coefficient, related with the fluid velocity profile:

$$\alpha = \frac{\int_S u_z^2 ds}{A\bar{u}^2} = \frac{\int_S \bar{u}^2 s^2 ds}{A\bar{u}^2} = \frac{\int_S s^2 ds}{A}$$

It is easy to see that $\alpha \geqslant 1$. In general α varies with time, t, and space, z. Here it is considered to be constant as a consequence of H6, since α is related with u_r. For steady flow in circular rigid tubes the Navier-Stokes equations have the very well known Poiseuille solution, consisting of a parabolic velocity profile. For a Poiseuille profile we have $s(y) = 2(1 - y^2)$ and $\alpha = 4/3$. For blood flow it has been found that the velocity profile is rather flat [11], corresponding to $s(y) = 1$ and $\alpha = 1$.

From assumptions H4, H5 and H6, the Navier-Stokes equations (2.1) become:

$$\begin{cases} \dfrac{\partial u_z}{\partial t} + \operatorname{div}(u_z \mathbf{u}) + \dfrac{1}{\rho}\dfrac{\partial P}{\partial z} - \nu \Delta u_z = 0, \\ \operatorname{div} \mathbf{u} = 0, \end{cases} \quad \text{in } \Omega \tag{2.3}$$

On the wall of the vessel, Γ_w, we have the kinematic condition $\mathbf{u} = \frac{\partial \mathbf{d}}{\partial t} = \frac{\partial d}{\partial t}\mathbf{e}_r$, meaning that the wall moves at the same velocity as the fluid.

We will integrate equations (2.3) on the generic cross section $S(z)$ term by term. Consider the portion V of the vascular tube Ω around point z, comprising the axial

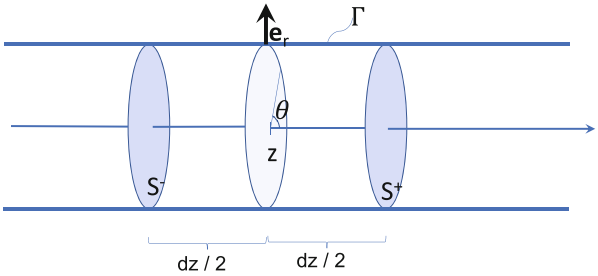

Fig. 2.2 Portion of the tube between $z - \frac{dz}{2}$ and $z + \frac{dz}{2}$

region $\left(z - \frac{dz}{2}, z + \frac{dz}{2}\right)$, see Fig. 2.2. We denote the boundary of V by $\partial V = S^- \cup S^+ \cup \Gamma$, with Γ the part of the boundary of V intercepting the vessel wall. The 1D model is derived by integrating equations (2.3) in V and doing the limit as $dz \to 0$, assuming all quantities are smooth enough.

Using the Reynolds transport theorem [12], taking into account that the border of V intercepting the vessel wall, Γ, moves with time t, it can be shown that $\frac{\partial A}{\partial t} = 2\pi R \frac{\partial d}{\partial t}$. Using this expression, the divergence theorem and the mean-value theorem for integrals, we have

$$\begin{aligned} 0 = \int_V \operatorname{div}(\mathbf{u}) dv = \int_{\partial V} \mathbf{u} \cdot \mathbf{n} ds = -\int_{S^-} u_z ds + \int_{S^+} u_z ds + \int_\Gamma \frac{\partial \mathbf{d}}{\partial t} \cdot \mathbf{n} ds \\ = Q\left(z + \frac{dz}{2}\right) - Q\left(z - \frac{dz}{2}\right) + \frac{\partial A(z)}{\partial t} dz + o(dz) \end{aligned}$$

where $\mathbf{n}$ denotes the outward unitary normal. Dividing by dz and doing the limit as $dz \to 0$ we obtain $\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial z} = 0$ for the continuity equation.

From Reynolds theorem we have $\frac{d}{dt}\int_V u_z dv = \int_V \frac{\partial u_z}{\partial t} dv + \int_{\partial V} u_z \frac{\partial \mathbf{d}}{\partial t} \cdot \mathbf{n} ds$. Due to H2, boundaries S^- and S^+ do not move longitudinally, and due to H6, $u_z = 0$ on Γ. Thus $\int_{\partial V} u_z \frac{\partial \mathbf{d}}{\partial t} \cdot \mathbf{n} ds = 0$ and

$$\int_V \frac{\partial u_z}{\partial t} dv = \frac{d}{dt} \int_V u_z dv = \frac{\partial}{\partial t} \left(A(z)\bar{u}(z) dz + o(dz)\right) = \frac{\partial Q(z)}{\partial t} dz + o(dz) \quad (2.4)$$

Using the divergence theorem and again assumptions H2 and H6, we have:

$$\int_V \operatorname{div}(u_z \mathbf{u}) dv = \int_{\partial V} u_z \mathbf{u} \cdot \mathbf{n} dv = -\int_{S^-} u_z^2 ds + \int_{S^+} u_z^2 ds + \int_\Gamma u_z \frac{\partial \mathbf{d}}{\partial t} \cdot \mathbf{n} ds$$

Remembering that $u_z = 0$ on Γ and that $\alpha = \int_S u_z^2 ds/(A\bar{u}^2)$, we obtain

$$\int_V \text{div}(u_z \mathbf{u}) dv = \alpha \left[A\left(z + \frac{dz}{2}\right)\bar{u}^2\left(z + \frac{dz}{2}\right) - A\left(z - \frac{dz}{2}\right)\bar{u}^2\left(z - \frac{dz}{2}\right)\right]$$
$$\approx \alpha \frac{\partial (A\bar{u}^2)}{\partial z} \tag{2.5}$$

Considering now hypothesis H4 and again the divergence theorem, we obtain

$$\int_V \frac{\partial P}{\partial z} dv = -\int_{S^-} P ds + \int_{S^+} P ds + \int_\Gamma P \mathbf{n}_z ds \approx \frac{\partial (AP)}{\partial z}(z) + \int_\Gamma P \mathbf{n}_z ds$$

From the mean-value theorem for integrals $\int_\Gamma P \mathbf{n}_z ds = P(z) \int_\Gamma \mathbf{n}_z ds + o(dz)$. Also, since V is closed, $\int_{\partial V} \mathbf{n}_z ds = 0$ and $\int_\Gamma \mathbf{n}_z ds = -\left(\int_{S^+} ds - \int_{S^-} ds\right)$, thus

$$\int_\Gamma P \mathbf{n}_z ds = -P(z)\left(\int_{S^+} ds - \int_{S^-} ds\right) + o(dz)$$
$$= -P(z)\left[A\left(z + \frac{dz}{2}\right) - A\left(z - \frac{dz}{2}\right)\right] + o(dz) \approx -P(z)\frac{\partial A}{\partial z}(z)$$

Thus, we obtain

$$\int_V \frac{\partial P}{\partial z} dv \approx \frac{\partial (AP)}{\partial z} - P\frac{\partial A}{\partial z} = A\frac{\partial P}{\partial z} \tag{2.6}$$

Finally, we consider the viscous term Δu_z. Applying the divergence theorem and noticing that $\nabla u_z = \left(\frac{\partial u_z}{\partial r}, \frac{\partial u_z}{\partial \theta}, \frac{\partial u_z}{\partial z}\right)$, we have:

$$\int_V \Delta u_z dv = \int_{\partial V} \nabla u_z \cdot \mathbf{n} ds = -\int_{S^-} \frac{\partial u_z}{\partial z} ds + \int_{S^+} \frac{\partial u_z}{\partial z} ds + \int_\Gamma \nabla u_z \cdot \mathbf{n} ds$$

We neglect the term $\frac{\partial u_z}{\partial z}$ by assuming that the variation of axial velocity u_z along the axial direction z is small compared to the other terms. From H1 the gradient of u_z becomes $\nabla u_z = \left(\frac{\partial u_z}{\partial r}, 0, 0\right)$, and $\nabla u_z \cdot \mathbf{n}$ becomes $\frac{\partial u_z}{\partial r}$. Recalling expression (2.2), we have that $\frac{\partial u_z}{\partial r} = \frac{\bar{u}(t,z)}{R(z)} s'\left(\frac{r}{R(z)}\right)$. Noticing that $r = R$ on Γ, we obtain:

$$\int_V \Delta u_z ds = \int_\Gamma \frac{\bar{u}}{R} s'(1) ds = 2\pi \int_{z-\frac{dz}{2}}^{z+\frac{dz}{2}} \bar{u} s'(1) dz = 2\pi \bar{u}(z) s'(1) dz + o(dz) \tag{2.7}$$

Finally, putting together expressions (2.4), (2.5), (2.6) and (2.7), diving (2.4) and (2.7) by dz, and passing to the limit as $dz \to 0$, the momentum equation becomes

$$\frac{\partial Q}{\partial t} + \alpha \frac{\partial}{\partial z}\left(\frac{Q^2}{A}\right) + \frac{A}{\rho}\frac{\partial P}{\partial z} - 2\pi \nu s'(1)\frac{Q}{A} = 0$$

The friction parameter is defined as $K_r = -2\pi\nu s'(1)$ and depends on the velocity profile $s(\cdot)$. For a parabolic profile $K_r = 8\pi\nu$, which corresponds to the value commonly used in practice. In the case of a flat profile, $\alpha = 1$ and $K_r = 0$, meaning that there is no friction.

The 1D model for blood flowing in a cylindrical vessel is given, for all t, by

$$\begin{cases} \frac{\partial Q}{\partial t} + \alpha \frac{\partial}{\partial z}\left(\frac{Q^2}{A}\right) + A\frac{\partial P}{\partial z} = -K_r \frac{Q}{A}, & z \in (a,b) \\ \frac{\partial A}{\partial t} + \frac{\partial Q}{\partial z} = 0, & z \in (a,b) \end{cases} \tag{2.8}$$

where $L = b - a$ is the vessel length, and the unknowns are the cross section area $A(t,z)$, the flow rate $Q(t,z)$, and the constant cross sectional pressure $P(t,z)$. The friction parameter is here a source term on the momentum equation. We have three unknowns and two equations, meaning an extra equation is required. We close system (2.8) by providing a relation linking pressure and area. This is reasonable, since the tube wall moves as a response to the pressure inside the vessel. We consider the following simple pressure-area algebraic relation:

$$P = \beta \frac{\sqrt{A} - \sqrt{A_0}}{A_0} \tag{2.9}$$

where A_0 is the cross section area at rest and $\beta = \frac{\sqrt{\pi}hE}{1-\xi^2}$, with E and ξ the Young Modulus and the Poisson ratio of the wall material, and h is the wall thickness. Parameters A_0 and β may vary with z. We define $c = \sqrt{\frac{A}{\rho}\frac{\partial P}{\partial A}}$, which becomes $c = \sqrt{\frac{\beta}{2\rho A_0}}A^{1/4}$ for expression (2.9) and considering A_0 and β constant along z. In this case, system (2.8) can be diagonalized, meaning it is a strictly hyperbolic system of PDEs, describing very well wave propagation phenomenon. Indeed, system (2.8) with (2.9) has two eigenvalues, given by $\lambda_{1,2} = \alpha\bar{u} \pm \sqrt{c^2 + \bar{u}^2\alpha(\alpha-1)}$. If we choose $\alpha = 1$, corresponding to a flat profile, we obtain $\lambda_{1,2} = \bar{u} \pm c = \bar{u} \pm \sqrt{\frac{\beta}{2\rho A_0}}A^{1/4}$. Under physiological conditions, the mechanical properties of blood and of the arterial wall, reflected on c through β, are such that $c >> \bar{u}$, meaning that the two eigenvalues have opposite signs. Indeed, characteristic values of c are of the order of 10^3m/s [13], while $|\bar{u}|$ is of the order of 10^1 [14]. This means that system (2.8) describes two waves travelling along the cylindrical vessel, one moving forward and the other backwards.

Let $R = [r_1\ r_2]$ and $L = [l_1\ l_2]$ be the matrices of the right and left eigenvectors, respectively, such that $LR = I$, with I the identity matrix. Then, if there exist

quantities W_1 and W_2 such that $\frac{\partial W_1}{\partial U} = l_1$ and $\frac{\partial W_2}{\partial U} = l_2$, for $U = [A, Q]^T$, which is the case if $\alpha = 1$ and $K_r = 0$ (see for instance [9]), functions W_1 and W_2 are characteristic variables [15]. This means that, up to the additive source term, variables W_i, $i = 1, 2$ are constant along the characteristic lines, that is, along the lines satisfying the differential equation $\frac{d}{dt} y_i(t) = \lambda_i(t, y_i(t)), i = 1, 2$. Given the pressure-area relation (2.9) and $\alpha = 1$, W_1 and W_2 are given by $W_{1,2} = \bar{u} \pm \sqrt{\frac{8\beta}{\rho A_0}} \left(A^{1/4} - A_0^{1/4}\right)$.

Finally, to completely define problem (2.8), we must provide initial conditions $A(0, z) = A^0$, $Q(0, z) = Q^0$ and boundary conditions, which in general can be written as $\phi_1(A(t), Q(t)) = h_1(t)$, at $z = a$, and $\phi_2(A(t), Q(t)) = h_2(t)$, at $z = b$, where h_1 and h_2 are given functions. The number of boundary conditions to apply at each end is the number of incoming characteristics at that point. Thus, here we must impose exactly one boundary condition at $z = a$ and $z = b$, respectively [15]. Functions ϕ_i, $i = 1, 2$, produce admissible boundary conditions as long as they do not depend only on the exiting characteristic.

2.3 Numerical Approximation of the 1D Model

The numerical discretization of the 1D hyperbolic model is carried out using the finite element Lax-Wendroff method, [15, 16]. Being a second-order explicit scheme in time, it has excellent dispersion error properties and it is easily implemented.

Being explicit, the stability of the numerical scheme depends on the satisfaction of a CFL type condition (see [9]) relating the time step Δt with the space step h_i: $\Delta t \leqslant (\sqrt{3}/3) \min_{0 \leqslant i \leqslant N} \left[\frac{h_i}{\max_{k=1,2} \lambda_k(z_i)} \right]$, where z_i, $i = 0, \ldots, N$, are the mesh nodes, and $h_i = z_{i+1} - z_i$ is the measure of the i-th spacial element.

2.3.1 Boundary and Compatibility Conditions for the Discrete Problem

The numerical solution is defined only on the internal nodes. The solution values at the boundary points are computed from the boundary conditions. Yet, these data are not enough to close the numerical problem. Indeed, for the problem at hand we have to impose exactly one boundary condition at each end of the domain at the continuous level [15, 16]. However, at the discrete level we need to provide two boundary data at each end, corresponding to the unknowns $Q^{n+1} = Q(t^n)$ and $A^{n+1} = A(t^n)$. That is, at the numerical level we need an extra boundary condition at each extremity. To obtain the complete boundary data we need additional equations, which must be compatible with the original problem. Usually, the compatibility conditions are obtained by projecting the equation along the characteristic lines exiting the domain [15, 16]. The values of Q^{n+1} and A^{n+1}

at the boundaries are found by solving the following non linear systems:

$$\begin{cases} \phi_1(A_h^{n+1}(a), Q_h^{n+1}(a)) = h_1^{n+1} \\ W_2(A_h^{n+1}(a), Q_h^{n+1}(a)) = W_2^{n+1} \end{cases} \quad \begin{cases} \phi_2(A_h^{n+1}(b), Q_h^{n+1}(b)) = h_2^{n+1} \\ W_1(A_h^{n+1}(b), Q_h^{n+1}(b)) = W_1^{n+1} \end{cases} \tag{2.10}$$

If we consider $\alpha = 1$ and $K_r = 0$ (no source term in (2.8)), W_i^{n+1} is obtained following the characteristic line from $t = t^n$: $W_i^{n+1} = W_i^n(x_k)$, where x_k is the foot of the characteristic line of W_i at t^n, $i = 1, 2$.

For instance, if the user provides the pressure, at $z = a$, as $P(t, a) = h_1(t)$, from (2.9) and from the compatibility condition (2.10) at $z = a$ we need to solve the system

$$\begin{bmatrix} \frac{1}{A^{n+1}(a)}\beta\frac{\sqrt{A^{n+1}(a)}-\sqrt{A_0}}{A_0} & 0 \\ -\sqrt{\frac{8\beta}{\rho A_0}}\frac{(A^{n+1}(a))^{1/4}-A_0^{1/4}}{A^{n+1}(a)} & \frac{1}{A^{n+1}(a)} \end{bmatrix} \begin{bmatrix} A^{n+1}(a) \\ Q^{n+1}(a) \end{bmatrix} = \begin{bmatrix} h_1^{n+1} \\ C_{x_a}^n \end{bmatrix}$$

where $C_{x_a}^n = \frac{Q^n(x_a)}{A^n(x_a)} - \sqrt{\frac{8\beta}{\rho A_0}}\left((A^n(x_a))^{1/4} - A_0^{1/4}\right)$ and $x_a = a - \Delta t\lambda_2(Q^n(a), A^n(a))$ is the foot of the exiting characteristic W_2 at $z = a$. If, at $z = b$, a condition on the entering characteristic is imposed $W_2(t, b) = h_2(t)$, then the non-linear system to be solved at each time step is

$$\begin{bmatrix} -\sqrt{\frac{8\beta}{\rho A_0}}\frac{(A^{n+1}(b))^{1/4}-A_0^{1/4}}{A^{n+1}(b)} & \frac{1}{A^{n+1}(b)} \\ \sqrt{\frac{8\beta}{\rho A_0}}\frac{(A^{n+1}(b))^{1/4}-A_0^{1/4}}{A^{n+1}(b)} & \frac{1}{A^{n+1}(b)} \end{bmatrix} \begin{bmatrix} A^{n+1}(b) \\ Q^{n+1}(b) \end{bmatrix} = \begin{bmatrix} h_2^{n+1} \\ C_{x_b}^n \end{bmatrix}$$

where $C_{x_b}^n = \frac{Q^n(x_b)}{A^n(x_b)} + \sqrt{\frac{8\beta}{\rho A_0}}\left((A^n(x_b))^{1/4} - A_0^{1/4}\right)$ and $x_b = b - \Delta t\lambda_1(Q^n(b), A^n(b))$ is the foot of the exiting characteristic W_1 at $z = b$.

Very often the incoming characteristic is put equal to zero at the exiting point, $W_2(t, z) = 0$, corresponding to an absorbing boundary condition, meaning nothing enters the domain. This fact has been exploit to impose absorbing boundary conditions on 3D FSI models for blood flow [17].

2.3.2 Modular Simulation of Arterial Networks

So far we have the mathematical 1D model and corresponding numerical method to simulate blood flow in a single elastic tube. However, the interest is often the simulation of arterial networks. Hence, we need to couple two or more arteries. The most usual combination of arteries in the human arterial system are:

- *Coupling single tubes*: used to model, for instance, long tubes where physical characteristics vary.
- *Bifurcation of an artery into two arteries*: this is the most common, since almost all arteries enventually bifucarte to carry out blood for the whole body.
- *Merging of two arteries into one artery*: this situation is more rare, occurring for instance with the vertebral arteries (left and right), that merge into the basilar artery before reaching the Circle of Willis at the bottom of the brain.

Having the numerical scheme for one tube, couplings, bifurcations and mergings are possible by defining coupling conditions. These consist in imposing the continuity of the fluxes, Q, and of the total pressures, $P_t = P + \frac{\rho}{2}\bar{u}^2$, [7, 9].

Couplings lead to two (in the case of coupling) or three (in the case of bifurcating or merging) more boundary conditions. Thus, we also need more compatibility conditions, obtained again by means of the exiting characteristics. For instance, for the coupling of two tubes, the following coupling conditions are set

$$\begin{cases} Q_1(b_1) = Q_2(a_2) \\ P_{t,1}(b_1) = P_{t,2}(a_2) \\ W_1(t^{n+1}, b_1) = W_1(t^n, x_{b_1}) \\ W_2(t^{n+1}, a_2) = W_2(t^n, x_{a_2}) \end{cases}$$

where $a_i, b_i, i = 1, 2, 3$ are respectively the initial and final extremities of artery i and $x_{k_i}, k = a, b, i = 1, 2, 3$ are the foot of the outgoing characteristic lines passing in point (k_i, t^{n+1}). For the bifurcation (b) and merging (m) we have the following systems

$$b = \begin{cases} Q_1(b_1) = Q_2(a_2) + Q_3(a_3) \\ P_{t,1}(b_1) = P_{t,2}(a_2) \\ P_{t,1}(b_1) = P_{t,3}(a_2) \\ W_1(t^{n+1}, b_1) = W_1(t^n, x_{b_1}) \\ W_2(t^{n+1}, a_2) = W_2(t^n, x_{a_2}) \\ W_2(t^{n+1}, a_3) = W_2(t^n, x_{a_3}) \end{cases} \qquad m = \begin{cases} Q_1(b_1) + Q_2(b_2) = Q_3(a_3) \\ P_{t,1}(b_1) = P_{t,3}(a_3) \\ P_{t,2}(b_2) = P_{t,3}(a_3) \\ W_1(t^{n+1}, b_1) = W_1(t^n, x_{b_1}) \\ W_1(t^{n+1}, b_2) = W_1(t^n, x_{b_2}) \\ W_2(t^{n+1}, a_3) = W_2(t^n, x_{a_3}) \end{cases}$$

All these systems are non-linear. For each internal coupling point of the 1D arterial network, a system of this type must be solved, for instance using Newton's method, to set the discrete boundary conditions of all tubes in the network.

2.4 Simulating Anatomical Variations of the Circle of Willis

Here we illustrate the application of the 1D hyperbolic model to the study of anatomical variations of the Circle of Willis (CoW), see [1, 5, 7] for simulations at different levels of accuracy of the brain circulation. The CoW is a redundancy

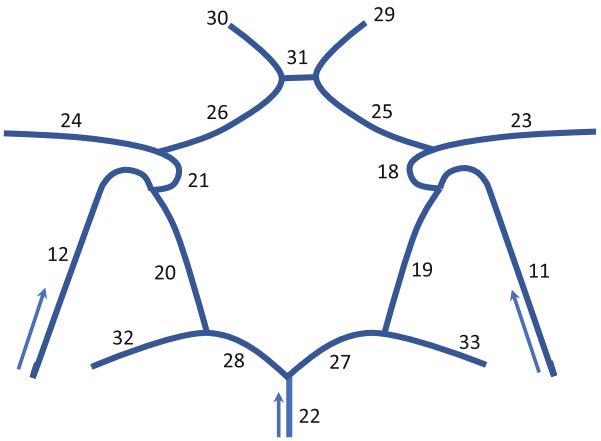

Fig. 2.3 Representation of the main arteries of the Circle of Willis (CoW)

set of arteries guaranteeing that blood reaches the brain by distributing it from the basilar and internal carotid arteries, see Fig. 2.3. We choose the parameters of each artery, namely length, radius, wall thickness and Young modulus (related to wall elasticity) as in [7]. We consider a sinusoidal impulse at the basilar artery (artery 22 in Fig. 2.3) and at the internal carotid arteries (ICAs, arteries 11 and 12 in Fig. 2.3):

$$Q(t,0) = \begin{cases} 2.5 \sin\left(\frac{\pi t}{0.003}\right), & \text{if } t \leqslant 0.003 \\ 0, \text{if } t > 0.003 \end{cases}$$

As customary, we consider that the system starts at rest $A(0,z) = A_0$ and $Q(0,z) = 0$.

Figure 2.4 represents the flow rate through the main arteries of the healthy CoW, represented in Fig. 2.3, at a particular time t. We can observe that as long as the CoW is vertically axis-symmetric, the flow rate through symmetric arteries is the same. Several people have anatomical changes in the CoW [7], including absence of some arteries. Figure 2.5 shows the flow rate when the right posterior cerebral artery is missing (right PCA, artery 28 in Fig. 2.3). It can be seen that the flow rate of symmetric arteries is not the same anymore. Indeed, the flow rate in arteries 32 and 33 differ. In both these cases, there is still blood perfusion to the brain through arteries 32 and 33, even though one artery is absent. In Fig. 2.6 we can observe the flow rate in the arteries going to the brain when both the right posterior communicating artery (PCoA, artery 20 in Fig. 2.3) and the left posterior cerebral artery (PCA, artery 27 in Fig. 2.3) are absent. We can see that in this case there is almost no blood perfusion in artery 32, that is, almost no blood perfusion on that part of the brain. This is thus a very serious anatomical variation of the CoW.

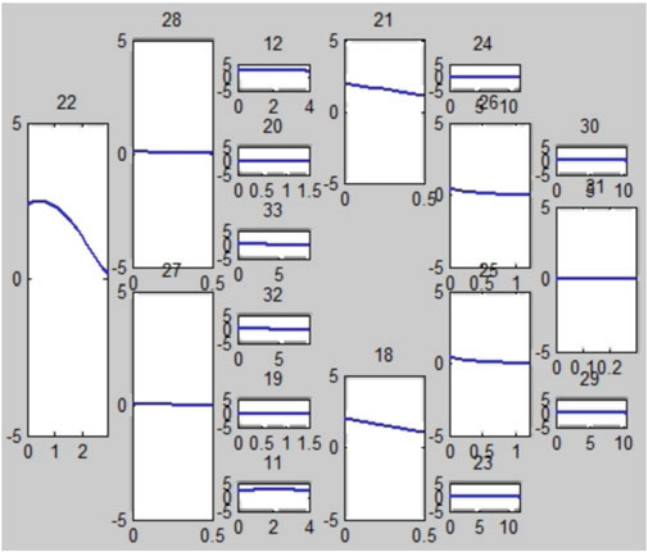

Fig. 2.4 Flow rate [cm^3/s] through the main arteries of the complete CoW at a particular time t

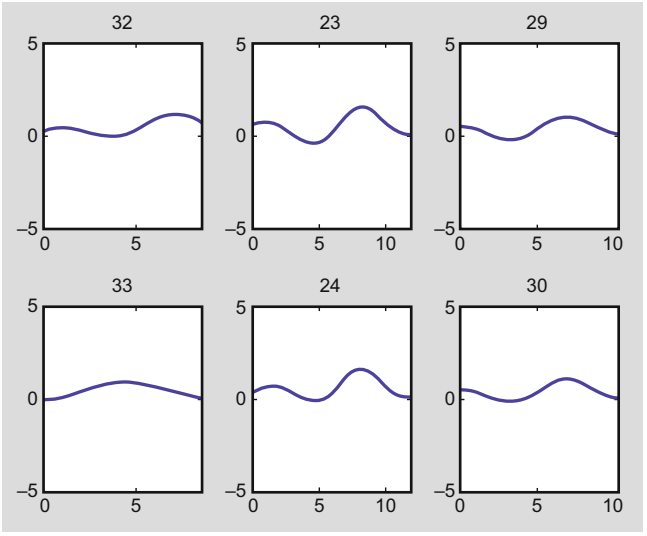

Fig. 2.5 Flow rate when the right posterior cerebral artery is missing (right PCA, artery 28 in Fig. 2.3)

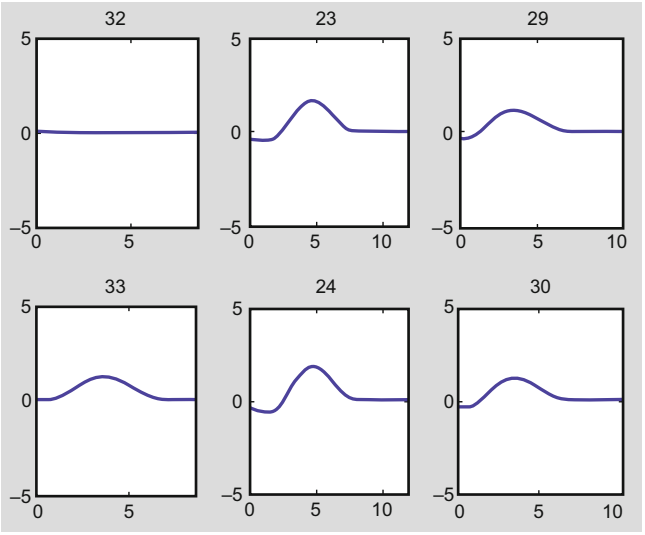

Fig. 2.6 Flow rate in the arteries when both the right posterior communicating artery (PCoA) and the left posterior cerebral artery (PCA) are absent (arteries 20 and 27 in Fig. 2.3)

2.5 Increasing the Complexity of the 1D Model: Including Curvature

The 1D hyperbolic model for blood flow in arteries studied in the previous sections is a simplified model with desirable mathematical properties, namely it is an hyperbolic system of PDEs. It may be considered one of the simplest 1D model for blood pulse. More complex models can be considered by accounting for variations of the radius or the wall properties along z, which introduce derivatives of the parameters A_0 and β, to be included on the source term. Also, more sophisticated pressure-area relations can be used, leading to the appearance of higher order derivatives. These derivatives alter the characteristic of the differential problem, making the numerical treatment and the identification of proper boundary conditions more problematic [9]. These effects also imply the inclusion of new parameters for which it is often difficult to obtain reasonable values.

On the other hand, some simplified assumptions can be relaxed. An example is to consider curvature. In this case hypothesis H3, of fixed cylinder axis, as well as H2, of only radial movements, are dropped. Now, the wall movements are not only radial, depending also on the x and y directions. Denoting $\boldsymbol{\eta}$ the wall displacement, then $\boldsymbol{\eta} = (\eta_x, \eta_y) + d\mathbf{e}_r$. In this setting, the fluid velocity depends not only on the axial direction through u_z, but also on the directions u_x and u_y:

$$u_x = \frac{\partial \eta_x}{\partial t} + \frac{1}{R}\frac{\partial d}{\partial t}x = \frac{\partial \eta_x}{\partial t} + \frac{1}{R}\frac{\partial d}{\partial t} r\cos\theta, \; u_y = \frac{\partial \eta_y}{\partial t} + \frac{1}{R}\frac{\partial d}{\partial t}y = \frac{\partial \eta_y}{\partial t} + \frac{1}{R}\frac{\partial d}{\partial t} r\sin\theta$$

Considering, for simplicity, a parabolic profile on the axial direction, from (2.2) we know that $u_z = \left(1 - \frac{r^2}{R^2}\right) a(t,z) = \bar{u}(t,z)s\left(\frac{r}{R}\right)$, with $a/2 = Q/A$. In this case, the 3D Navier-Stokes system (2.3) has two extra equations, related to u_x and u_y:

$$\begin{cases} \frac{\partial u_x}{\partial t} + \operatorname{div}(u_x \mathbf{u}) + \frac{1}{\rho}\frac{\partial P}{\partial x} - \nu \Delta u_x = 0 \\ \frac{\partial u_y}{\partial t} + \operatorname{div}(u_y \mathbf{u}) + \frac{1}{\rho}\frac{\partial P}{\partial y} - \nu \Delta u_y = 0 \\ \frac{\partial u_z}{\partial t} + \operatorname{div}(u_z \mathbf{u}) + \frac{1}{\rho}\frac{\partial P}{\partial z} - \nu \Delta u_z = 0 \\ \frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} + \frac{\partial u_z}{\partial z} = 0 \end{cases} \tag{2.11}$$

We integrate each equation in (2.11) over the volume V and then make $dz \to 0$. Applying this procedure to the two last equations in (2.11) results in the two equations of system (2.8). Indeed, noticing that $\int_\Gamma n_{r_1} \cos\theta \eta_x ds = \int_\Gamma n_{r_2} \sin\theta \eta_y ds = 0$, we have:

$$\int_\Gamma \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds = \int_\Gamma n_{r_1} \cos\theta \eta_x ds + \int_\Gamma n_{r_2} \sin\theta \eta_y ds + \int_\Gamma \frac{1}{R}\frac{\partial d}{\partial t} r \mathbf{e}_r \cdot \mathbf{n} ds \approx \frac{\partial A}{\partial t}$$

where $\mathbf{n} = (n_{r_1} \cos\theta, n_{r_2} \sin\theta, n_z)$ is the outward unitary normal. Thus

$$\int_V \operatorname{div} \mathbf{u} dv = \int_{\partial V} \mathbf{u} \cdot \mathbf{n} ds = \int_{S^+} u_z ds - \int_{S^-} u_z ds + \int_\Gamma \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds \approx \frac{\partial A}{\partial t} + \frac{\partial Q}{\partial z}$$

Since $\int_{z-\frac{dz}{2}}^{z+\frac{dz}{2}} \int_0^{2\pi} \int_0^R \frac{\partial d}{\partial t} \frac{r\cos\theta}{R} r dr d\theta dz = 0$, then

$$\int_V u_x dv = \int_{z-\frac{dz}{2}}^{z+\frac{dz}{2}} \int_0^{2\pi} \int_0^R \frac{\partial \boldsymbol{\eta}_x}{\partial t} r dr d\theta dz = \frac{\partial \eta_x}{\partial t}(z) A(z) dz + o(dz) \approx \frac{\partial \eta_x}{\partial t} A$$

Thus, defining $\Phi_x = \frac{\partial \eta_x}{\partial t} A$, then $\frac{d}{dt}\int_V u_x dv \approx \frac{\partial \Phi_x}{\partial t}$. Since $\frac{\partial \boldsymbol{\eta}}{\partial t} = 0$ on S, we have

$$\int_{\partial V} u_x \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds = \int_\Gamma u_x \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds$$

Hence, we obtain:

$$\int_V \frac{\partial u_x}{\partial t} dv = \frac{d}{dt}\int_V u_x ds - \int_{\partial V} u_x \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds \approx \frac{\partial \Phi_x}{\partial t} - \int_\Gamma u_x \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds$$

Using the divergence theorem and noticing that $\int_S \frac{1}{R}\frac{\partial d}{\partial t} r\cos\theta \, ds = 0$ and that

$$\int_S u_x u_z ds = \int_S \left(\frac{\partial \eta_x}{\partial t} + \frac{1}{R}\frac{\partial d}{\partial t} r\cos\theta\right)\left(1 - \frac{r^2}{R^2}\right) a(t,z) ds = \frac{\partial \eta_x}{\partial t} A \frac{Q}{A} = \Phi_x \frac{Q}{A}$$

we obtain

$$\int_V \mathrm{div}(u_x \mathbf{u}) dv = \int_{S^+} u_x u_z ds - \int_{S^-} u_x u_z ds + \int_\Gamma u_x \frac{\partial \boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds$$
$$\approx \frac{\partial}{\partial z}\left(\Phi_x \frac{Q}{A}\right) + \int_\Gamma u_x \frac{\boldsymbol{\eta}}{\partial t} \cdot \mathbf{n} ds$$

Pressure is still considered to be constant over each cross section, so that $\frac{\partial P}{\partial x} = 0$ and hence $\int_V \frac{\partial P}{\partial x} = 0$. Finally, for the diffusive term on x, we may neglect $\frac{\partial u_x}{\partial z}$, obtaining

$$\int_V \Delta u_x dv = \int_{\partial V} \nabla u_x \cdot \mathbf{n} ds = \int_{S^+} \frac{\partial u_x}{\partial z} ds - \int_{S^-} \frac{\partial u_x}{\partial z} ds + \int_\Gamma \nabla u_x \cdot \mathbf{n} ds = \int_\Gamma \nabla u_x \cdot \mathbf{n} ds$$

Noticing that $\nabla u_x \cdot n_z \propto \frac{\partial u_x}{\partial z} = 0$ we get

$$\int_\Gamma \nabla u_x \cdot \mathbf{n} ds = \int_\Gamma \nabla u_x \cdot n_r ds = \int_\Gamma \frac{\partial u_x}{\partial r} ds = \int_\Gamma \frac{1}{R} \frac{\partial d}{\partial t} \cos\theta ds = 0$$

Doing the limit as $dz \to 0$, the first equation of (2.11) becomes

$$\frac{\partial}{\partial t}(\Phi_x) + \frac{\partial}{\partial z}\left(\Phi_x \frac{Q}{A}\right) = 0.$$

A similar equation is obtained by integrating the second equation in (2.11), defining $\Phi_y = \frac{\partial \eta_y}{\partial t} A$. The resulting 1D system of equations is

$$\begin{cases} \frac{\partial A}{\partial t} + \frac{\partial Q}{\partial z} = 0 \\ \frac{\partial Q}{\partial t} + \frac{4}{3}\frac{\partial}{\partial z}\left(\frac{Q^2}{A}\right) + \frac{A}{\rho}\frac{\partial P}{\partial z} = -8\pi\nu\frac{Q}{A} \\ \frac{\partial \Phi_x}{\partial t} + \frac{\partial}{\partial z}\left(\Phi_x \frac{Q}{A}\right) = 0 \\ \frac{\partial \Phi_y}{\partial t} + \frac{\partial}{\partial z}\left(\Phi_y \frac{Q}{A}\right) = 0 \end{cases} \tag{2.12}$$

Notice that for a parabolic profile $\alpha = 4/3$. This system of PDEs is no longer strictly hyperbolic, being more difficult to deal with than system (2.8), both from the mathematical and numerical points of view. This 1D model can be further extended to the case where the axial velocity profile depends on x and y: $u_z = \left(1 - \frac{r^2}{R^2}\right)(a + bx + cy)$. In this case, two extra equations are needed, for the added quantities b and c, related with the axial velocity profile. These equations are obtained by integrating over the cross section the third equation of (2.11) multiplied by x and y, respectively. Using the same approach and arguments as before, the 1D model in

this case becomes:

$$\begin{cases} \frac{\partial A}{\partial t} + \frac{\partial Q}{\partial z} = 0 \\ \frac{\partial Q}{\partial t} + \frac{4}{3}\frac{\partial}{\partial z}\left(\frac{Q^2}{A}\right) + 6\pi\frac{\partial}{\partial z}\left(\frac{H^2}{A^2}\right) + 6\pi\frac{\partial}{\partial z}\left(\frac{G^2}{A^2}\right) + A\frac{\partial P}{\partial z} = -8\pi\nu\frac{Q}{A} \\ \frac{\partial H}{\partial t} + 2\frac{\partial}{\partial z}\left(\frac{QH}{A}\right) + \frac{1}{2}\frac{H}{A}\frac{\partial Q}{\partial z} + \frac{Q}{A}\Phi_x = -24\pi\nu\frac{H}{A} \\ \frac{\partial G}{\partial t} + 2\frac{\partial}{\partial z}\left(\frac{QG}{A}\right) + \frac{1}{2}\frac{G}{A}\frac{\partial Q}{\partial z} + \frac{Q}{A}\Phi_y = -24\pi\nu\frac{G}{A} \\ \frac{\partial \Phi_x}{\partial t} + \frac{\partial}{\partial z}\left(\Phi_x\frac{Q}{A}\right) - \frac{1}{2}\frac{\partial}{\partial z}\left(\frac{H}{A}\frac{\partial Q}{\partial z}\right) = 0 \\ \frac{\partial \Phi_y}{\partial t} + \frac{\partial}{\partial z}\left(\Phi_y\frac{Q}{A}\right) - \frac{1}{2}\frac{\partial}{\partial z}\left(\frac{G}{A}\frac{\partial Q}{\partial z}\right) = 0 \end{cases} \tag{2.13}$$

where $Q = \frac{a\pi R^2}{2} = \frac{aA}{2}$, as before, and $H = \frac{b\pi R^2 A}{12}$ and $G = \frac{c\pi R^2 A}{12}$. This system of PDE's is significantly more complex than the usual 1D model (2.8). Its mathematical and numerical analysis becomes extremely cumbersome, since many of the nice mathematical characteristics of system (2.8) are lost. Being simplified models, the usefulness o such complex 1D models is questionable. If we need more detail and the cost is such an increase in model complexity, than possibly the best is to use full detailed 3D models.

2.6 Conclusions

One dimensional models for blood flow in arteries are simplified and computationally low cost models, obtained by making simplifying assumptions and performing averaging procedures on the 3D FSI model. The simplest 1D model is described by an hyperbolic system of PDE's. Despite having a lower level of accuracy compared to the full 3D representation, it captures very effectively the pulsation of blood flow. More complex 1D models can be obtained by relaxation of some simplifying hypothesis, as for instance accounting for curvature. In these cases the 1D mathematical model becomes significantly more complex, losing some of the appealing mathematical properties of the simpler straight tube 1D model. Namely, it is no longer hyperbolic. Being simplified models, the extra complexity introduced in these cases is hardly justified. Except for very particular cases, in general when high accuracy on the blood flow solution is required, 3D FSI full mathematical models tend to be preferable.

References

1. J.R. Cebral, F. Mut, D. Sforza, R. Lohner, E. Scrivano, P. Lylyk, C.M. Putman. Clinical Application of Image-Based CFD for Cerebral Aneurysms. Int J Numer Method Biomed Eng. **27(7)**, 977–992 (2011). https://doi.org/10.1002/cnm.1373

2. S. Ramalho A. Moura A.M. Gambaruto A. Sequeira. Sensitivity to outflow boundary conditions and level of geometry description for a cerebral aneurysm. International Journal for Numerical Methods in Biomedical Engineering **28**, 697–713 (2012). https://doi.org/10.1002/cnm.2461
3. P. Moireau, N. Xiao, M. Astorino, C. A. Figueroa, D. Chapelle, C. A. Taylor, J.-F. Gerbeau External tissue support and fluid-structure simulation in blood flows. Biomech Model Mechanobiol **11**, 1–18 (2012). https://doi.org/10.1007/s10237-011-0289-z
4. Carlo Cavedon, Stephen Rudin eds. Cardiovascular and Neurovascular Imaging: Physics and Technology (2016) Taylor & Francis
5. B.P. Walcott, C. Reinshagen, C. J. Stapleton, O. Choudhri, V. Rayz, D. Saloner, M.T. Lawton. Predictive modeling and in vivo assessment of cerebral blood flow in the management of complex cerebral aneurysms J Cereb Blood Flow Metab. (2016) **36(6)**, 998–1003. https://doi.org/10.1177/0271678X16641125
6. P. Zunino, J. Tambaca, E. Cutri, S. Canic, L. Formaggia, F. Migliavacca. Integrated Stent Models Based on Dimension Reduction: Review and Future Perspectives. Annals of Biomedical Engineering. **44**, 604–617 (2016)
7. J. Alastruey, K.H. Parker, J. Peiró, S.M. Byrd, S.J. Sherwin. Modelling the circle of Willis to assess the effects of anatomical variations and occlusions on cerebral flows. J Biomech. **40(8)**, 1794–805 (2007)
8. L. Formaggia, F. Nobile, A. Quarteroni. A one dimensional model for blood flow: application to vascular prosthesis. In: I. Babuska, P.G. Ciarlet, T. Miyoshi eds, Mathematical Modeling and Numerical Simulation in Continuum Mechanics, I. 137–153 (2002) Springer-Verlag Berlin
9. L. Formaggia, A. Veneziani. Reduced and multiscale models for the human cardiovascular system. Lecture notes VKI Lecture Series 2003–07, Brussels (2003)
10. L. Euler. Principia pro motu sanguinis per arterias determinando. Opera posthima mathematica et physica anno 1844 detecta. **2**, 814–823 (1775)
11. A. Quarteroni, L. Formaggia. Mathematical Modelling and Numerical Simulation of the Cardiovascular System. Handbook of Numerical Analysis **12**. 2–127 (2004). https://doi.org/10.1016/S1570-8659(03)12001-7
12. M.E. Gurtin, An introduction to continuum mechanics. (2008) Academic Press, Elsevier Science
13. J.J. Wang and K.H. Parker. Wave propagation in a model of the arterial circulation. Journal of Biomechanics. **37**, 475–470 (2004)
14. M.S. Olufsen and C.S. Peskin and W.Y. Kim and E.M. Pedersen and A. Nadim and J. Larsen. Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions. Annals of Biomedical Engineering. **28**, 1281–1299 (2000)
15. E. Godlewski and P.-A. Raviart. Numerical approximation of hyperbolic systems of conservation laws. (1996) Springer-Verlag, New York
16. A. Quarteroni, A. Valli. Numerical approximations of partial differential equations. Computer Series in Computational Mathematics. (1997) Springer-Verlag.
17. J. Janela, A. Moura, A. Sequeira. Absorbing boundary conditions for a 3D non-Newtonian fluid–structure interaction model for blood flow in arteries. International Journal of Engineering Science. **48(11)**, 1332–1349 (2010). https://doi.org/10.1016/j.ijengsci.2010.08.004

Chapter 3
Uncertainty Quantification of Chemical Kinetic Reaction Rate Coefficients

É. Valkó and T. Turányi

3.1 Introduction

In chemical kinetics, the reaction rate coefficients (also called rate constants) characterize the speed of a chemical reaction. The temperature dependence of the rate coefficients can be defined by Arrhenius parameters A, n, and E. Chemical kinetics databases for many elementary gas-phase reactions provide the recommended values of the Arrhenius parameters, the temperature range of their validity and the uncertainty of rate coefficient k defined by uncertainty parameter f. However, in the kinetics databases the uncertainty parameter f is usually considered to be temperature independent. The goal of this project is to calculate temperature dependent uncertainty limits of rate coefficients of elementary reactions in such a way that these limits are consistent with the temperature dependence of the rate coefficient. The project consists of three tasks:

1. Visualization of the available kinetics data and calculation of the f_{original} uncertainty parameters of the investigated rate coefficient in selected temperature points.
2. Calculation of the f_{extreme} uncertainty parameters in selected temperature points.
3. Determination of a continuous, temperature dependent uncertainty function $f_{\text{prior}}(T)$, which is consistent with the rate coefficient-temperature function.

In accordance with the three tasks above, in Sects. 3.2, 3.3, and 3.4 all background information and definitions needed to solve the problems are provided. A possible solution is presented using the example of elementary reaction $OH+H_2=H_2O+H$.

É. Valkó (✉)
Institute of Mathematics, ELTE Eötvös Loránd University, Budapest, Hungary

T. Turányi
Institute of Chemistry, ELTE Eötvös Loránd University, Budapest, Hungary

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_3

This chapter is based on the publications of Nagy et al. [1–3]. All program codes used at the preparation of the examples are freely available from Web site [2, 4].

3.2 Calculation of the Uncertainty Parameters f_{original} in Selected Temperature Points

The temperature dependence of rate coefficient k of a chemical reaction is usually described by the modified Arrhenius equation

$$k(T) = A\{T\}^n \exp\left(-\frac{E}{RT}\right), \tag{3.1}$$

where R is the universal gas constant, T is temperature in Kelvin and parameters A, n, E are the so-called Arrhenius parameters. In accordance with the recommendations [5], curly brackets are used to denote the numerical value of the enclosed physical quantity at the predefined units, which are in this case cm, K, s, mol. Introducing transformed parameters $\kappa(T) := \ln\{k(T)\}$, $\alpha := \ln\{A\}$ and $\varepsilon := E/R$, the linearized form of the modified Arrhenius equation is

$$\kappa(T) = \alpha + n\ln\{T\} - \varepsilon T^{-1}. \tag{3.2}$$

To show the temperature dependence of κ, the logarithm of the rate coefficient is plotted as a function of the reciprocal of temperature. Such a $\kappa(T^{-1})$ curve is called an Arrhenius curve.

Figure 3.1 shows the Arrhenius curves of measured and theoretically determined rate coefficients for reaction OH+H_2=H_2O+H. The corresponding Arrhenius parameters and their temperature intervals are available in the Supplementary Material of article [2].

The rate coefficient of an elementary reaction can be determined by various experimental methods. If several measurements were carried out in different laboratories (maybe using different methods) at similar temperatures, then the uncertainty of the rate coefficient can be well assessed at a given temperature or in a narrow temperature interval. If the uncertainty of a rate coefficient is determined from literature data at different temperatures, then these uncertainties can be very different from each other even at nearby temperatures. However, since the measured rate coefficients are interrelated by a common Arrhenius expression, therefore the uncertainties determined at different temperatures are also related. Taking into account the temperature dependence of the rate coefficient, the uncertainty at a given temperature cannot be high if it is low at nearby temperatures.

The next pages discuss the determination of an Arrhenius-equation-consistent uncertainty function from the uncertainties of a rate coefficient valid at given temperatures (or in given temperature intervals) and the features of the corresponding uncertainty domain of the Arrhenius parameters.

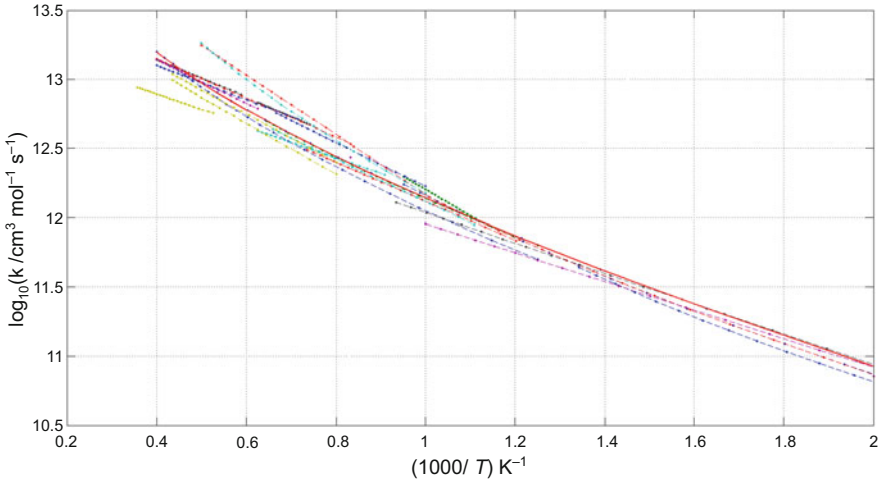

Fig. 3.1 The logarithm of measured and theoretically determined rate coefficients as a function of the reciprocal of temperature belonging to reaction $OH+H_2=H_2O+H$

Definition 3.1 The temperature dependent uncertainty function $f\,(T)$ is defined by Baulch et al. [6]:

$$f = \log_{10}\left(k^0(T)/k_{\min}(T)\right) = \log_{10}\left(k_{\max}(T)/k^0(T)\right), \tag{3.3}$$

where $k^0(T)$ is the recommended value of the rate coefficient at temperature T, $k_{\min}(T)$ and $k_{\max}(T)$ are the extreme, but still not excludable, physically realistic values at given temperature.

This definition of uncertainty is related to the limits and does not necessarily have a probabilistic inference. According to Eq. (3.3), the upper and lower extreme values differ from the recommended value by a multiplication factor, which means that, on a logarithmic scale, the extreme values are located symmetrically around the recommended value. The uncertainty bounds for rate coefficient k can be converted to the uncertainty bounds of $\kappa(T) := \ln\{k(T)\}$.

Definition 3.2 The uncertainty bound of the logarithm of the rate coefficient, $\kappa(T)$, is $\left[\kappa(T) - \frac{f(T)}{\ln 10}, \kappa(T) + \frac{f(T)}{\ln 10}\right]$.

Our aim is to define a temperature dependent uncertainty function $f(T)$, and thus defining a symmetrical bound for the recommended rate coefficient $\kappa^0(T)$. This bound will indicate the extreme, but physically realistic rate coefficient values at every temperature.

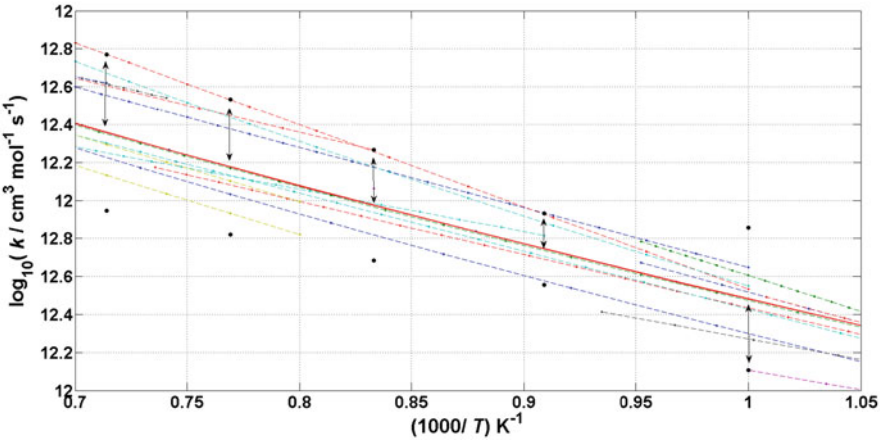

Fig. 3.2 The figure shows the Arrhenius curves of the rate coefficients that had been presented in Fig. 3.1, now in a narrower temperature interval (dashed lines). The mean value of the rate coefficient of reaction OH+H_2=H_2O+H was added (solid red line in the middle). The black circles and the sections with arrowheads ($\longleftrightarrow$) show the extent of the $f_{\text{original}}(T_i)$ values at $T = 1000$, 1100, 1200 and 1300 K

According to Eq. (3.3), the uncertainty of the investigated rate coefficient at a given temperature point, T_i, can be defined by

$$f_{\text{original}}(T_i) = \max\left(|\log_{10} k^0(T_i) - \log_{10} k_j(T_i)|\right) \quad j = 1, 2, \ldots, m_i;\ i = 1, 2, \ldots, n_T \tag{3.4}$$

where k_j is the j-th experimentally or theoretically determined value of the rate coefficient at temperature T_i, m_i is the number of all available values of the rate coefficient at temperature T_i and n_T is the number of temperature points.

Figure 3.2 shows the calculation of the $f_{\text{original}}(T_i)$ values for reaction OH+H_2=H_2O+H at $T = 1000$, 1100, 1200 and 1300 K. The investigated measurements and theoretical calculations for the values of the Arrhenius parameters and the corresponding temperature intervals are available in the Supplementary Material of [2].

Chemical kinetic databases contain suggested values of Arrhenius parameters A, n, E (or transformed Arrhenius parameters α, n, ε) and the temperature range where these values are relevant. Such information is available from the NIST Chemical Kinetics Database [7], the evaluations of Warnatz [8], Tsang et al. (see e.g. [9–11]), Baulch et al. (see e.g.[6, 12, 13]) and the review of Konnov [14].

Task 1 Collect all suggested values for the rate coefficient of reaction H+O_2=O+OH and create a user-friendly program code to visualize them in an Arrhenius plot. Using Arrhenius parameters A =2.07E+14 $\text{cm}^3\text{mol}^{-1}\text{s}^{-1}$, $n = -0.097$, $\frac{E}{R} = 7560$ K (the mean rate expression recommended by Baulch et al. [6]), calculate uncertainty parameter $f_{\text{original}}(T_i)$ at temperature points $T = 800, 900, \ldots, 2700$ K based on Eq. (3.4).

3.3 Calculation of the Uncertainty Values f_{extreme} at Selected Temperature Points

After the calculation of the $(T_i,\, f_{\text{original}}\,(T_i))$ point pairs, it is a natural idea fitting a polynomial to these points and using the fitted curve as a temperature dependent uncertainty function. However, the uncertainty bound defined by this function may contain physically not meaningful values of the rate coefficient. For this reason, a refined approach for the determination of the temperature-uncertainty point pairs is needed to obtain a physically relevant bound.

The procedure described here determines the uncertainty domain of Arrhenius parameters ($\mathbf{p} = (\alpha, n, \varepsilon)^{\mathrm{T}}$) from the uncertainty information for the rate coefficients. In several cases the temperature dependence of the rate coefficient can be described by two Arrhenius parameters (α, ε) or (α, n). In this case the third Arrhenius parameter is set to zero.

Assume that a central set of Arrhenius parameters $\mathbf{p}^0$ is available and the symmetric uncertainty of the rate coefficient is estimated at several temperatures by uncertainty parameters $f_{\text{original}}\,(T_i)\,, i = 1, \ldots, n_T$. It is possible to generate all Arrhenius curves $\kappa(T, \mathbf{p})$ that lie between the uncertainty limits, fulfilling the following $2n_T$ inequalities:

$$-f_{\text{original}}(T_i) \le \frac{\kappa(T_i; \mathbf{p}) - \kappa(T_i; \mathbf{p}^0)}{\ln 10} \le +f_{\text{original}}(T_i), \quad i = 1, 2, \ldots, n_T. \tag{3.5}$$

These curves are located symmetrically around the mean rate coefficient curve $\kappa\left(T; \mathbf{p}^0\right)$, since Arrhenius equation (3.2) is a linear function of parameters α, n, ε, and equation (3.3) defines symmetric linear constraints. A systematic procedure is proposed here for determining the extreme Arrhenius curves, which touch either the lower or the upper uncertainty limit at least at 2 or 3 temperatures for the 2- and the 3-parameter cases, respectively, and also go within the upper and lower uncertainty limits at all other temperatures. Formally, these criteria correspond to Arrhenius functions that fulfil at least 2 or 3 equality relations in Eqs. (3.5) and for the remaining $2n_T - 2$ or $2n_T - 3$ cases, respectively, either the equality or the inequality is fulfilled. The minimum and maximum values of these curves at a given temperature define the edges of the band of all possible Arrhenius curves.

In the case of the 3-parameter Arrhenius expression, term $n \ln T$ usually has a smaller contribution to the temperature dependence of the rate coefficient than $-\varepsilon/T$, since $\ln\{T\}$ changes more slowly than $1/T$ at combustion temperatures (800 K–2700 K). The effect of a change in the temperature exponent n on the rate coefficient at high temperatures can be well compensated by adjusting the pre-exponential factor α, leading to a very strong anti-correlation between α and n in most determinations. This implies that values of n, which significantly deviate (i.e. by ± 10) from the central n^0, can also fulfil all the inequality requirements in Eq. (3.5) if the initial uncertainty limits are not too tight. Both theoretical considerations [15] and the typical range of values of n in kinetic databases [7]

show that the temperature exponent n of elementary chemical reactions should take values of small negative or positive numbers. Therefore, we recommend confining the range of n values to a narrow (i.e. $\Delta n = 2$) symmetric interval around the central value n^0 when the band of possible Arrhenius curves is determined through finding extreme Arrhenius curves.

$$-\Delta n \leq n - n^0 \leq +\Delta n \tag{3.6}$$

The extreme Arrhenius curves are those which fulfil at least 2 or 3 equality relations in Eqs. (3.5) and (3.6) for the two-parameter and the three-parameter cases, respectively. To determine the extreme Arrhenius curves, uncertainty values need to be known at least at 2 temperatures, since in the three-parameter case a constraint is given for parameter n.

Definition 3.3 The minimum and maximum values of the extreme Arrhenius curves $(\kappa_{\min}(T))$ and $(\kappa_{\max}(T))$ define new uncertainty limits at any temperature, which are symmetrically located around the mean $\kappa\left(T; \mathbf{p}^0\right)$ curve. These new limits, obtained from a set of uncertainty values f and a user-defined Δn, uniquely define a new, continuous uncertainty function $f_{\text{extreme}}(T)$:

$$f_{\text{extreme}}(T) = \frac{\kappa\left(T; \mathbf{p}^0\right) - \kappa_{\min}(T)}{\ln 10} \equiv \frac{\kappa_{\max}(T) - \kappa\left(T; \mathbf{p}^0\right)}{\ln 10}. \tag{3.7}$$

By definition, this Arrhenius-equation-consistent uncertainty $f_{\text{extreme}}(T_i)$ is always less than or equal to the original uncertainty $f_{\text{original}}(T_i)$, at every temperature T_i $(i = 1, .., n_T)$.

Task 2 Create a program code to calculate $f_{\text{extreme}}(T_i)$ uncertainty values at temperature points $T = 800, 900, \ldots, 2700$ K based on the calculated $\left(T_i, f_{\text{original}}(T_i)\right)$ points for elementary reaction H+O_2=O+OH.

3.4 Calculation of Uncertainty Function $f_{\text{prior}}(T)$

The temperature dependent rate coefficient $k(T)$ (and its natural logarithm $\kappa(T)$) can be considered as a random variable deduced from measurements and calculations. Therefore, a probabilistic meaning may be attributed to f_{extreme}. According to this interpretation, if f_{extreme} corresponds to 3 standard deviations (3σ) [16–21] or 2 standard deviations (2σ) [22–24] of the rate coefficient on a decimal logarithmic scale, the uncertainty parameter f can be converted [17] to the standard deviation of the natural logarithm of the rate coefficient (σ_κ) at a given temperature T:

$$\sigma_\kappa(T) = \frac{\ln 10}{\mu} f(T), \tag{3.8}$$

where parameter μ is 3 or 2, respectively.

If the rate coefficients are considered as random variables, then Arrhenius parameters α, n and ε are also random variables, since these can be calculated from the random values of $\kappa(T)$ at three given temperatures using the linearized Arrhenius equation (Eq. (3.2)). The joint probability density function of the Arrhenius parameters is independent of temperature. This means that all central moments are also independent of temperature, including their expected values ($\bar{\alpha}$, $\bar{n}$, $\bar{\varepsilon}$), variances $\left(\sigma_\alpha^2,\ \sigma_n^2,\ \sigma_\varepsilon^2\right)$ and correlations ($r_{\alpha n},\ r_{\alpha\varepsilon},\ r_{\varepsilon n}$).

The following relation was deduced by Nagy [1] between the variance of $\kappa(T)$ and the elements of the covariance matrix of the Arrhenius parameters ($\mathbf{\Sigma_p}$):

$$\begin{aligned}\sigma_\kappa^2(T) &= \mathbf{\Theta}^T(\mathbf{\Sigma_p})\mathbf{\Theta} = \\ &= \sigma_\alpha^2 + \sigma_n^2 \ln^2 T + \sigma_\varepsilon^2 T^{-2} + 2r_{\alpha n}\sigma_\alpha\sigma_n \ln T + \\ &- 2r_{\alpha\varepsilon}\sigma_\alpha\sigma_\varepsilon T^{-1} - 2r_{n\varepsilon}\sigma_n\sigma_\varepsilon T^{-1}\ln T\end{aligned} \tag{3.9}$$

A method was proposed [1] for the determination of the covariance matrix of the Arrhenius parameters using Eqs. (3.8) and (3.9) from uncertainty parameter f of the rate coefficient at various temperatures. To determine the elements of the covariance matrix for the three-parameter Arrhenius expression, the uncertainty of the rate coefficient ($f_{\text{extreme}}(T_i)$) has to be known at least at six different temperatures. In the (α, ε) and (α, n) two-parameter cases, the uncertainty of the corresponding Arrhenius parameters can be handled in a similar way and the uncertainty of the rate coefficient has to be known at least at three temperatures [1]. Having calculated the $f_{\text{extreme}}(T_i)$ values at temperature points T_i, the $\sigma_\kappa^2(T_i)$ values are also available. The four or six elements of the covariance matrix have to be determined by minimising the difference between the values $\sigma_\kappa^2(T_i)$ and $\sigma_\alpha^2 + \sigma_n^2 \ln^2 T_i + \sigma_\varepsilon^2 T_i^{-2} + 2r_{\alpha n}\sigma_\alpha\sigma_n \ln T_i - 2r_{\alpha\varepsilon}\sigma_\alpha\sigma_\varepsilon T_i^{-1} - 2r_{n\varepsilon}\sigma_n\sigma_\varepsilon T_i^{-1}\ln T_i$ for all T_i, $(i = 1, \ldots, n_T)$.

Equations (3.8) and (3.9) provide a means for storing the $f_{\text{extreme}}(T)$ function in the form of the covariance matrix of Arrhenius parameters. The uncertainty parameter temperature function calculated from the covariance matrix can be used as a first guess (prior) and conservative estimation of the uncertainty of a rate coefficient and it can be used as a starting point of a more refined estimation [25] of the temperature dependent uncertainty of the rate coefficient.

Definition 3.4 The uncertainty function reconstructed from the covariance matrix is called prior uncertainty and denoted as $f_{\text{prior}}(T)$.

In Eq. (3.8), the parameter μ defines the proportionality between the uncertainty parameter f and the standard deviation σ_κ. When the uncertainty f_{prior} is calculated via σ_κ from the covariance matrix $\mathbf{\Sigma_p}$, the same parameter μ has to be used. This means that the value of μ is arbitrary in the storage of the f values in the covariance matrix, and the only important assumption here is that the uncertainty parameter f is proportional to the standard deviation of κ.

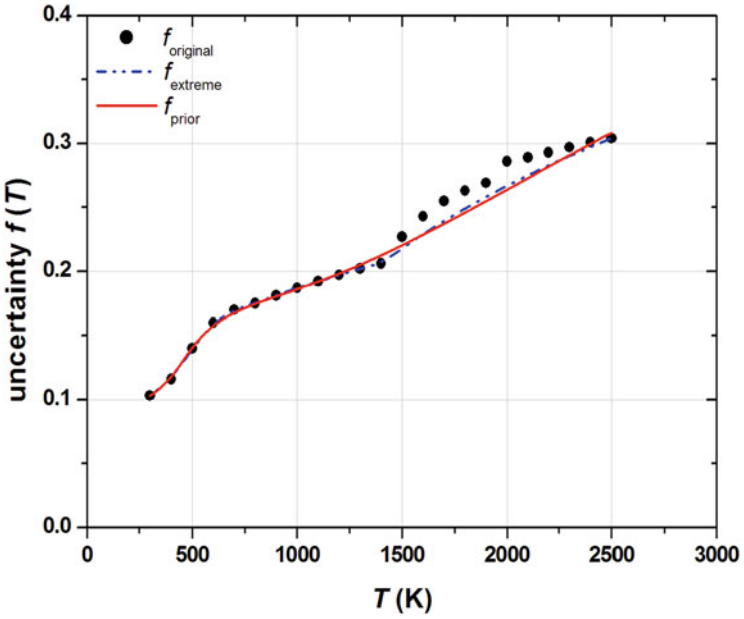

Fig. 3.3 The calculated $f_{\text{original}}(T_i)$ values (black circles), the extreme uncertainty function, $f_{\text{extreme}}(T)$ (blue dashed line) and the fitted $f_{\text{prior}}(T)$ function (red solid line) for reaction $OH+H_2=H_2O+H$ at temperature points T =300 K, 400 K, . . . , 2500 K

Figure 3.3 shows the calculated $(T_i, f_{\text{original}}(T_i))$ values and the calculated $f_{\text{extreme}}(T)$ function for reaction $OH+H_2=H_2O+H$. The red solid line in Fig. 3.3 represents the $f_{\text{prior}}(T)$ function, which was fitted to points $(T_i, f_{\text{extreme}}(T_i))$, where $T_i = 300\,\text{K}, 400\,\text{K}, \ldots, 2500\,\text{K}$.

Task 3 Using the results of Task 2, based on Eqs. (3.8) and (3.9), determine the values of the covariance matrix and define the $f_{\text{prior}}(T)$ function for reaction $H+O_2=O+OH$.

Acknowledgments Project no. ED_18-1-2019-0030 (Application-specific highly reliable IT solutions) has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme funding scheme.

References

1. Nagy, T., Turányi, T.: Uncertainty of Arrhenius parameters. International Journal of Chemical Kinetics 43, 359–378 (2011).
2. Nagy, T., Valkó, É., Sedyó, I., Zsély, I.G., Pilling, M.J., Turányi, T.: Uncertainty of the rate parameters of several important elementary reactions of the H_2 and syngas combustion systems. Combustion and Flame 162(5), 2059–2076 (2015).

3. Nagy, T., Turányi, T.: Determination of the uncertainty domain of the Arrhenius parameters needed for the investigation of combustion kinetic models. Reliability Engineering and System Safety 107, 29–34 (2012).
4. ReSpecTh information system. http://www.respecth.hu.
5. JCGM: International vocabulary of metrology—Basic and general concepts and associated terms (VIM). http://www.bipm.org/s (2008).
6. Baulch, D.L., Bowman, C.T., Cobos, C.J., Cox, R.A., Just, T., Kerr, J.A., Pilling, M.J., Stocker, D., Troe, J., Tsang, W., Walker, R.W., Warnatz, J.: Evaluated kinetic data for combustion modeling: Supplement II. Journal of Physical and Chemical Reference Data 34, 757–1397 (2005).
7. Manion, J.A., Huie, R.E., Levin, R.D., Burgess Jr., D.R., Orkin, V.L., Tsang, W., McGivern, W.S., Hudgens, J.W., Knyazev, V.D., Atkinson, D.B., Chai, E., Tereza, A.M., Lin, C.-Y., Allison, T.C., Mallard, W.G., Westley, F., Herron, J.T., Hampson, R.F., Frizzell, D.H.: NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.7, Data Version 2013.03, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899–8320. http://kinetics.nist.gov/ (2013).
8. Warnatz, J.: Rate coefficients in the C/H/O system. In: Gardiner, W.C. (ed.) Combustion chemistry. pp. 197–361. Springer, New York (1984)
9. Tsang, W., Hampson, R.F.: Chemical kinetic database for combustion chemistry 1. Methane and related compounds. Journal of Physical and Chemical Reference Data 15, 1087–1279 (1986).
10. Tsang, W.: Chemical kinetic data base for combustion chemistry Part V. Propene. Journal of Physical and Chemical Reference Data 20, 221–273 (1991).
11. Tsang, W., Herron, J.T. Journal of Physical and Chemical Reference Data 20, 609–663 (1991).
12. Baulch, D.L., Cobos, C.J., Cox, R.A., Esser, C., Frank, P., Just, T., Kerr, J.A., Pilling, M.J., Troe, J., Walker, R.W., Warnatz, J.: Evaluated kinetic data for combustion modeling. Journal of Physical and Chemical Reference Data 21, 411–734 (1992).
13. Baulch, D.L., Cobos, C.J., Cox, R.A., Frank, J.H., Hayman, G., Just, T.H., Kerr, J.A., Murrels, T., Pilling, M.J., Troe, J., Walker, B.F., Warnatz, J.: Summary table of evaluated kinetic data for combustion modeling—Supplement-1. Combustion and Flame 98, 59–79 (1994).
14. Konnov, A.A.: Remaining uncertainties in the kinetic mechanism of hydrogen combustion. Combustion and Flame 152, 507–528 (2008).
15. Pilling, M.J., Seakins, P.W.: Reaction Kinetics. Oxford University Press, Oxford (1995)
16. Brown, M.J., Smith, D.B., Taylor, S.C.: Influence of uncertainties in rate constants on computed burning velocities. Combustion and Flame 117, 652–656 (1999).
17. Turányi, T., Zalotai, L., Dóbé, S., Bérces, T.: Effect of the uncertainty of kinetic and thermodynamic data on methane flame simulation results Physical Chemistry Chemical Physics 4, 2568–2578 (2002).
18. Zsély, I.G., Zádor, J., Turányi, T.: Uncertainty analysis backed development of combustion mechanisms. Proceedings of the Combustion Institute 30, 1273–1281 (2005).
19. Zádor, J., Zsély, I.G., Turányi, T., Ratto, M., Tarantola, S., Saltelli, A.: Local and global uncertainty analyses of a methane flame model. The Journal of Physical Chemistry A 109, 9795–9807 (2005).
20. Zádor, J., Zsély, I.G., Turányi, T.: Local and global uncertainty analysis of complex chemical kinetic systems. Reliability Engineering and System Safety 91, 1232–1240 (2006).
21. Zsély, I.G., Zádor, J., Turányi, T.: Uncertainty analysis of NO production during methane combustion. International Journal of Chemical Kinetics 40, 754–768 (2008).
22. Sheen, D.A., You, X., Wang, H., Løvås, T.: Spectral uncertainty quantification, propagation and optimization of a detailed kinetic model for ethylene combustion. Proceedings of the Combustion Institute 32, 535–542 (2009).
23. Sheen, D., Wang, H.: Combustion kinetic modeling using multispecies time histories in shock-tube oxidation of heptane. Combustion and Flame 158, 645–656 (2011).

24. Sheen, D.A., Wang, H.: The method of uncertainty quantification and minimization using polynomial chaos expansions. Combustion and Flame 158, 2358–2374 (2011).
25. Turányi, T., Nagy, T., Zsély, I.G., Cserháti, M., Varga, T., Szabó, B.T., Sedyó, I., Kiss, P.T., Zempléni, A., J., C.H.: Determination of rate parameters based on both direct and indirect measurements. International Journal of Chemical Kinetics 44, 284–302 (2012).

Chapter 4
Nuclear Accidents: How Can Mathematicians Help to Save Lives?

Simone Göttlich

4.1 Introduction

The description of real-life problems by using mathematical modeling techniques enables the simulation and optimization of complex systems. In the present case study, this is the investigation of an evacuation process caused by a nuclear accident. To build up a real scenario and to work with data freely available, the problem was restricted to one of Germany's largest nuclear power plant located in Biblis (south-west of Germany) that has been closed in 2011 after the Fukushima accident.

The modeling includes a combined model to tackle the evacuation of people from a designated area and the spread of a radioactive cloud. From a mathematical point of view, this means to deal with flows on graphs and the approximation of travel times. Furthermore, knowledge on diffusion-advection equations is needed for the evolution of nuclear material over time and space. However, the main challenge is the combination of both approaches into an integrated framework, i.e., to set up an evacuation plan that is dependent on the current spread of radioactivity.

Within this contribution, a successful modelling approach is introduced which has been worked out by eight students during the *23-th ECMI Modelling Week 2010 in Milan/Italy*. The major subject of all students was related to applied mathematics. Due to the ECMI Modelling Week philosophy, the students came from different European universities so that the conversations were exclusively in English.

S. Göttlich (✉)
University of Mannheim, Department of Mathematics, Mannheim, Germany
e-mail: goettlich@uni-mannheim.de

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_4

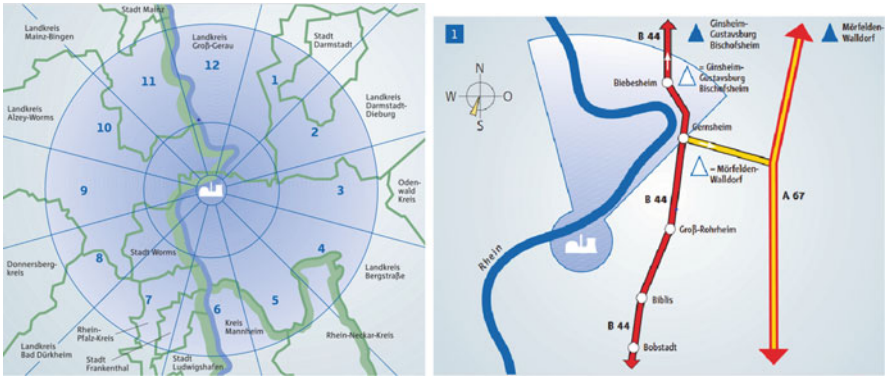

Fig. 4.1 Evacuation zones around Biblis (Germany) on the left and zoom into sector 1 on the right

4.1.1 Problem Description

A radiation accident is defined by the International Atomic Energy Agency[1] as "an event that has led to significant consequences to people, the environment or the facility. Examples include lethal effects to individuals, large radioactivity release to the environment, or reactor core melt." The worst-case scenario of a *major nuclear accident* is one in which a reactor core is damaged and large amounts of radiation are released, such as in the Chernobyl Disaster in 1986, or more recently, the Fukushima nuclear power plant accident in March 2011. So there is definitely a strong need for a fast and most of all safe evacuation, even nowadays.

For the description of a concrete scenario, we focus on the surrounding area of the Biblis power plant, see left picture in Fig. 4.1.

The picture is taken from the official emergency protection that is still available.[2] As we can see the immediate neighborhood is mainly divided into three zones: the central zone (radius of 1.5–2 km), the middle zone (radius of 10 km) and the outer zone (radius of 25 km). The zones are again arranged into 12 sectors to meet geographical restrictions, see right picture in Fig. 4.1.

A rough estimate on the number of inhabitants in the entire outer zone is given by 1.4 million people in 2010, see right picture in Fig. 4.2. Most of the people live in sectors 6–12, i.e. on the right side on the river Rhine, including cities such as Darmstadt, Mannheim or Ludwigshafen. However, this area can be only evacuated in the direction north or south since there is a mountain region in the east. The evacuation on the left side of the river is open to all directions (except east). The

[1] https://www.iaea.org.

[2] https://www.group.rwe/unser-portfolio-leistungen/betriebsstandorte-finden/kernkraftwerk-biblis.

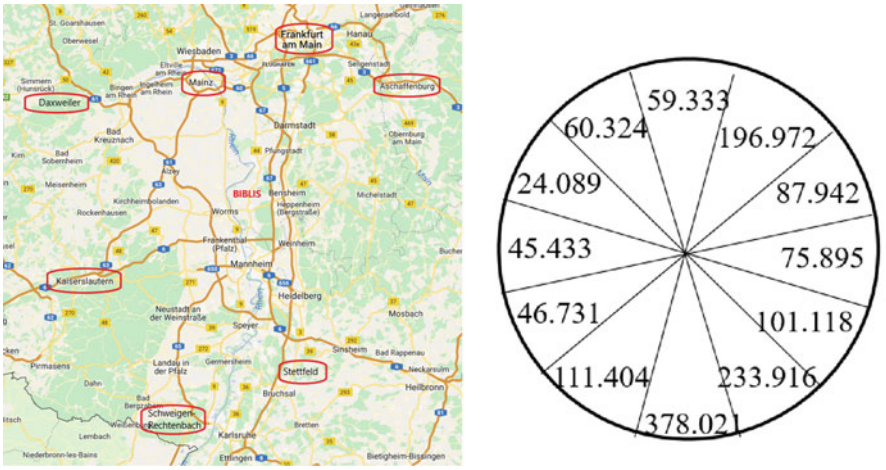

Fig. 4.2 Road network around Biblis with safety points in red (left) and estimates on the number of inhabitants per sector (right)

safety points for all sectors are seven cities outside the outer zone marked in red, see left picture[3] in Fig. 4.2.

A further idea of an evacuation plan is also to identify bottlenecks due to the given road network, see again Fig. 4.2. The critical area could be separated into two parts, east and west, where several highways are available. The main roads east of the river are A67, A5, B3, B44 and B47, where the prefix A denotes a German highway (average speed 130 km/h) and B a federal road (average speed 100 km/h). On the west of the river we have the roads A61, A63, A6 and B9. In the special case of Biblis, there are seven cities considered as safety points, namely Frankfurt, Mainz, Daxweiler, Kaiserslautern, Schweigen-Rechtenbach, Stettfeld and Aschaffenburg. We note that people should avoid traversing the road B9 since it is close to the Biblis power plant and might be massive congested.

Another important information is that the wind direction is usually (more than 70%) from west to east meaning that most of the radiation is spread towards the hilly region called Odenwald. Summarizing, an evacuation can be only carried out in north, south or west direction. This is an important aspect while rerouting the people to the safety points.

In the following, we present and comment on the key ideas of the group to solve the evacuation problem. We introduce mathematical models, solution methods and numerical simulations that might help to set up a reasonable evacuation plan. Several assumptions are needed to emphasize on features such as traffic congestion, people's behavior in case of panic and weather influences.

[3] https://www.falk.de.

4.2 Solution of the Problem Provided by the Group

The proposed solution approach can be mainly divided into three parts: the mathematical description of the evacuation process, the propagation of nuclear material and the combination of both.

As we have seen the evacuation model is influenced by predefined properties such as geographical restrictions, local infrastructure, road capacities and population densities. However, other modelling inputs as for instance individual human behavior or weather conditions are harder to predict.

In order to scale down the problem the following assumptions and simplifications have been suggested:

- The evacuation is considered by cars on highways or dederal roads, where each car drives at the maximum speed allowed.
- Accidents may occur and lead to a certain delay factor.
- The radioactive cloud is assumed to be homogeneous and the spread is only driven by the wind.

For the modelling of the people's behaviour, a kind of network model was applied to give each individual certain attributes and to include the spread of radiation later on. In contrast, the radioactive cloud was modelled using a diffusion-advection equation in two space dimensions that could be solved by standard numerical methods. After modelling these two parts separately, a combined model was introduced to find an evacuation plan which is as safe and smoothly as possible according to the radiation spread.

4.2.1 Modelling of the Evacuation Process

Biblis is connected to its surrounding towns and safety points by several highways, see Fig. 4.2. These towns, safety points and highways are modelled as a network or graph, respectively. As described by Bondy and Murty[2], a *graph* G is an ordered pair $(V(G), E(G))$ consisting of a set of vertices $V(G)$ and a set of edges $E(G)$. If $u, v \in V(G)$, then an edge connecting both vertices is denoted as $e = uv \in E(G)$. A *path* is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence. A path has a start vertex, an end vertex, but no repeated vertices and edges.

This means from an application point of view:

- V is the set of towns in 25 km radius Biblis area,
- E is the set of type A and B roads considered in the evacuation network,
- p is a set of cars, each transporting 5 people,
- $S(p) = v \in V$ is the safety point of p,
- $L(p, t)$ is the location of car p at time t.

Then, the aim is to find the shortest paths between cars and the nearest safety point, i.e. $\min ||L(p,t) - S(p)||$. The dynamics between different towns or on a certain path is driven by the people's behavior and the resulting consequences. This is explained in details in the next subsections on the initialization of the evacuation and the network dynamics.

4.2.2 Initialization of Evacuation Process

In the sense of an initialization of the problem, a *delay factor* is introduced to model the time that is needed to leave a town and enter the road network. In the given model, to each "individual" p (i.e. a car) there is an attributed delay factor that can be written as

$$D_p = D_{global} + D_{individual},$$

where D_{global} is a global delay factor describing the information spread over the media. From the viewpoint of an individual, it is the difference of time between the nuclear accident and the first moment when the individual might get the media information. The global delay factor has been estimated as

$$D_{global} = 30\,\text{min} = 360\ TimeUnits,$$

where $1\ TimeUnit = 5\,\text{s}$.

The second delay factor $D_{individual}$ is an individual-based factor and describes the time needed by each individual before to reach the highway and enter the given road network. Considering the whole set of individuals for a certain town, the goal of the $D_{individual}$-modeling is to obtain a distribution describing the amount of cars arriving per *TimeUnit* at the local highway. Therefore the two factors *size of the town*, i.e. the *number of inhabitants* and the individual-based *panic factor* contribute to this modelling and are now explained in more detail:

1. each car has an attributed *panic factor*. This factor is sampled from a distribution of panic factors $P(cars)$:

$$P(cars) \sim \begin{cases} N\,(0, 0.5) & \text{for a calm population,} \\ 1 - N\,(0, 0.5) & \text{for a panicked population.} \end{cases}$$

   Taking into consideration only the absolute values of the samples, the panic factors are $p_{ind} \in [0, 1]$, where

$$p_{ind} = \begin{cases} \text{low panic factor, direct movement} & \text{if } p_{ind} \leq 0.5, \\ \text{high panic factor, indirect movement} & \text{if } p_{ind} > 0.5. \end{cases}$$

2. All towns have a circular shape and a similar population density that is 500 people per km^2. Each town has a sufficient number of highway exits at the border of the town, so cars can therefore go straightforward along the radius of a town and reach a highway. The number of cars arriving in the network model per *TimeUnit* is the sum of cars reaching all highways per *TimeUnit*. The 2D-model can be therefore interpreted as a 1D-model, see Fig. 4.3.
3. People with direct movement ($p_{ind} \leq 0.5$) move the slower the higher the panic factor. The movement of people with indirect movement ($p_{ind} > 0.5$) becomes more random with rising panic factor.

Individuals move in-between the center of the city $X_0 = 0$ and the highways at the border of the town $X_{end} = r$. This r denotes the radius of the town and it can be calculated using the given inhabitants and population density in the city as

$$r = \sqrt{\frac{\text{inhabitants}}{500\pi}}.$$

The considered time grid is given by discrete time steps $t_0, \ldots, t_k = t_0 + k \cdot \Delta t$ with $\Delta t = 5$ s. At time t_0, the individuals are distributed as a $N(0, 0.4)$. Their position denotes the distance to the city center. The direct movement of an individual can be then defined as

$$X_{t_k} = X_{t_{k-1}} + \Delta X = \begin{cases} X_{t_{k-1}} + dir \cdot (1 - p_{ind}) & \text{low panic factor,} \\ X_{t_{k-1}} + dir \cdot (1 - p_{ind}) & \text{high panic factor,} \end{cases}$$

with

$$dir = \begin{cases} 2.5 & \text{low panic factor,} \\ U[-1, 2.5] & \text{high panic factor.} \end{cases}$$

This means that individuals with a low panic factor move directly with a panic-factor scaled velocity of maximal 30 km/h while individuals with a high panic factor may obtain negative velocities, i.e. they move towards the center and away from the highways.

4.2.3 Dynamics on Road Network

There are a few assumptions necessary to describe the movement of cars through the road network such that every car tries to reach its safety point as fast as possible, i.e.

- every driver knows its safety point,
- if a car reaches the safety point, it will be considered safe,

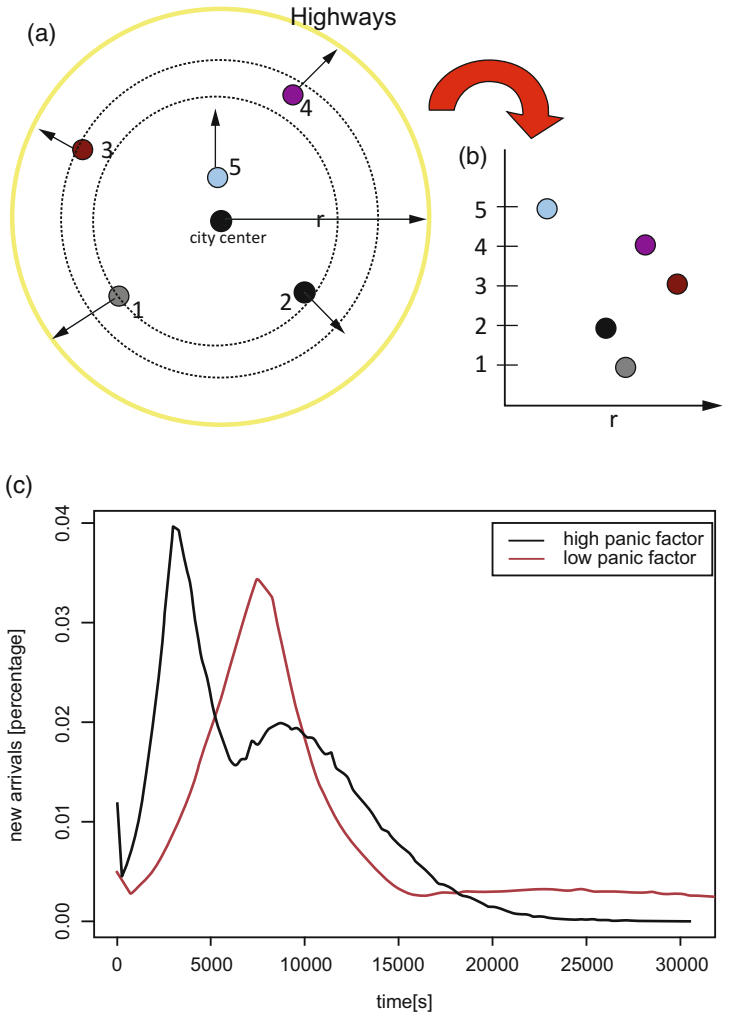

Fig. 4.3 Modeling the individual local delay factor for a single example town. (**a**) Schematic 2D-representation of an idealized town: in yellow the assumed continuous distribution of highways at the border of the town ($X = r$). Five individuals are placed randomly in a certain distance X to the city center X_0 and their moving direction is denoted. (**b**) Reduction to a 1D-model: all highways of the town count in a sum at the border of the town. For better visibility, individuals are labeled. (**c**) Different panic factors: percentage of new arrivals at the highway per time unit for the example town with 100,000 inhabitants. For a calm population, there are move arrivals after few time-steps. A panic real population needs more time to find the highways due to the loss of orientation

- every driver has basic knowledge of the roads in the area,
- every driver reconsiders his/her route only at the vertices of the graph,
- every driver has knowledge on the traffic situation in front.

Mathematically, the perceived distance to the safety point can be calculated as

$$\sum_{i,j \in P} \frac{D_{ij}}{\hat{v_{ij}}},$$

where P is a path from the current node to the safety point, i and j are vertices on the graph G, D_{ij} is the distance between vertex i and vertex j along the edge ij and $\hat{v_{ij}}$ is the estimated velocity on the edge.

A path-finding algorithm is needed to compute the routes between the multiple vertices. One commonly used algorithm for finding the shortest route through a graph is the A* search algorithm, as described by Nilsson et al. [6]. A* is a best-first search algorithm, which uses a heuristic cost function to find the shortest path to destination.

Furthermore, the traveling speed $\hat{v_{ij}}$ on a given edge from vertex i to j is determined by:

$$\rho = \frac{P_{ij}}{200 w_{ij} D_{ij}} \quad \text{and} \quad \rho_{crit} = \frac{0.125}{\left(1 + \alpha A_{ij}\right)},$$

where P_{ij} is the amount of people on the road, w_{ij} the amount of lanes, A_{ij} the amount of accidents happened on this road and α an experimental constant. Then, $vmax_{ij}$ is the maximum speed on the road and determined by

$$\hat{v_{ij}} = vmax_{ij}, \text{ when } \rho < \rho_{crit},$$
$$\hat{v_{ij}} = \left(1 - \rho_{ij}^2\right) vmax_{ij}, \text{ when } \rho \geq \rho_{crit}.$$

This represents a typical behavior of traveling speed, where on an empty road full speed is possible until ρ exceeds a threshold value ρ_{crit} and the speed drops down. Note that also the road capacity is reduced when accidents occur.

4.2.3.1 Modeling Car Accidents

Given the car density on each highway and the speed of the individuals, a simplified model for the number of accidents implies that the maximum speed of cars is thereby reduced.

Doing so, the minimal safety distance is given by a certain speed v of all individuals on a highway as

$$x_{min} = v/2.$$

However, the actual distance depends on the car density $\rho \in [0, 1]$ on each highway. Remember that the density counts the number of 5 m-long cars on a piece of 1 km of the highway. A density of 100% means therefore that 200 cars are on the highway without any distance in-between them. So the space $sp(\rho)$ in-between two 5 m-long cars given a density ρ can be computed as:

$$cars\rho = 200 \cdot \rho \text{and } sp(cars) = \frac{1000}{cars} - 5.$$

If the actual space is smaller than the minimum safety distance ($sp(\rho) < x_{min}$) then there are accidents occurring.

4.2.4 Modeling of the Radioactive Cloud

From a modelling point of view, a nuclear accident releases many types of different radioactive materials characterized by different weights, different particle sizes, different decay constants and different responses to weather conditions. To reduce complexity, the main assumption is to consider a homogenous cloud only and not to distinguish different material properties. As already pointed out, meteorological conditions play a major role for the spread of radiation once it has been released. The wind field is here the key ingredient to describe the drift or the direction of spread, respectively.

The unknown variable is then the concentration of the homogenous radioactive material in a unit volume, i.e. $C(t, \mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^2$. One possible way is to derive a differential equation that describes first the diffusion of the radioactive material is the continuity equation approach. Considering an arbitrary volume V, the equation reads

$$\begin{bmatrix}\text{rate of}\\ \text{change of}\\ \text{radioactive}\\ \text{material in } V\end{bmatrix} = \begin{bmatrix}\text{rate of}\\ \text{production of}\\ \text{radioactive}\\ \text{material in } V\end{bmatrix} - \begin{bmatrix}\text{rate of}\\ \text{decay of}\\ \text{radioactive}\\ \text{material in } V\end{bmatrix} - \begin{bmatrix}\text{rate of}\\ \text{leakage of}\\ \text{radioactive}\\ \text{material in } V\end{bmatrix}$$

and therefore,

$$\int_V \frac{\partial C}{\partial t} dV = \int_V S dV - \int_V \lambda C dV - \int_A \phi^T \cdot \mathbf{n} dA,$$

where $S = S(t, \mathbf{x})$ the source of radioactive material, λ the radioactive decay constant and $\phi^T = \phi^T(t, \mathbf{x})$ the total radioactive material flux. Note that ϕ^T is exactly the total radioactive material flux crossing the unit area of the unit volume, orthogonal to the flow direction.

Introducing the wind field in a second step, the problem becomes a diffusion-advection equation. Then, the total radioactive material flux ϕ^T also consists of convective fluxes ϕ^C and diffusive fluxes ϕ^F. So, considering an isotropic scattering radioactive material, Fick's law yields

$$\phi^F = -D\nabla C$$

where ϕ^F measures the amount of concentration that flows through a small area during a small time interval, while D is the constant diffusion coefficient.

Wind is modelled as a vector field describing the motion of the air. The length of each vector is the flow speed, in particular:

$$\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$$

which gives the velocity at a position $\mathbf{x}$ at time t.

Supposing that the radioactive material moves as fast as wind, $\mathbf{u}$ is the average velocity of the radioactive material. The convective flux is then

$$\phi^C = \mathbf{u}C.$$

Remembering that $\phi^T = \phi^C + \phi^F$, substitution and the use of the divergence theorem leads to:

$$\int_V \frac{\partial C}{\partial t} dV = \int_V S dV - \int_V \lambda C dV + \int_V \left[D\nabla^2 C - \nabla \cdot (\mathbf{u}C) \right] dV$$

Since the volume V is arbitrary, the following partial differential equation (PDE) of diffusion-advection type can be derived:

$$\frac{\partial C}{\partial t} = D\nabla^2 C - \nabla \cdot (\mathbf{u}C) - \lambda C + S$$

with initial condition $C(\mathbf{x}, 0) = f(\mathbf{x})$ and boundary conditions $C(0, y, t) = C(1, y, t) = C(x, 0, t) = C(x, 1, t) = 0$.

In view of combining the solution of the PDE problem with the road network model, a finite domain in 2D is needed. This a is reasonable choice due to the given infrastructure. Note that the main unknown variable is then just the concentration of the homogenous radioactive material in a unit surface instead of in a unit volume.

4.2.5 Combined Model

The combined model consists of the road network model and the diffusion-advection equation. Both approaches can be computed and simulated separately. In particular, for the network model this means:

- First, the delay for people leaving their current locations and entering the highways, i.e. the road network, is modelled as an influx to the network vertices. The rate of influx on each time step is determined by the distribution of people's delay times. This can be done in some preprocessing manner.
- Second, the congestion and accident models are closely connected. Both affect the traveling speeds on each road, which are used as input for the movement of people and the calculation of the escape route.
- Third, the coupling of the radioactive cloud model to the network model. The parameters for radioactive cloud model are set so that both of the models operate on the same time and space domain, i.e. one time step in the radioactive cloud model equals one time step in the network model and a coordinate point in the radioactive cloud model corresponds to the same point in the network model.

More precisely, the dosage D received in place $vecx$ on time step t is directly correlated with the concentration of radioactive material C:

$$Dose\,(x, y, t) = \varphi C\left((x, y)^T, t\right)$$

The constant φ is an empirically determined parameter for radiation effectiveness and depends on the weather—rainy or clear—and the type of fallout—small dust particles or large pieces of radioactive material. The total dosage and thus the effect of radiation, i.e. E_r on person i in a short time span, can be calculated as a sum of doses received (see [5]),

$$E_r\,(i, t) = \sum_{s=t_0}^{t} \varphi C\left(\left(X\,(i, s), Y\,(i, s)\right), s\right),$$

where $X(i, t)$ and $Y(i, t)$ are the coordinates of the person i on time step t.

4.3 Simulation Results

Running now the simulations on the Biblis test case, several important aspects about the evacuation planning can be identified. An illustration can be seen in Fig. 4.4.

A surprising results is that the total time taken for evacuation does not significantly change when using either the calm or panicky population distributions. The calm people exit the city in a short time frame, causing massive traffic jams

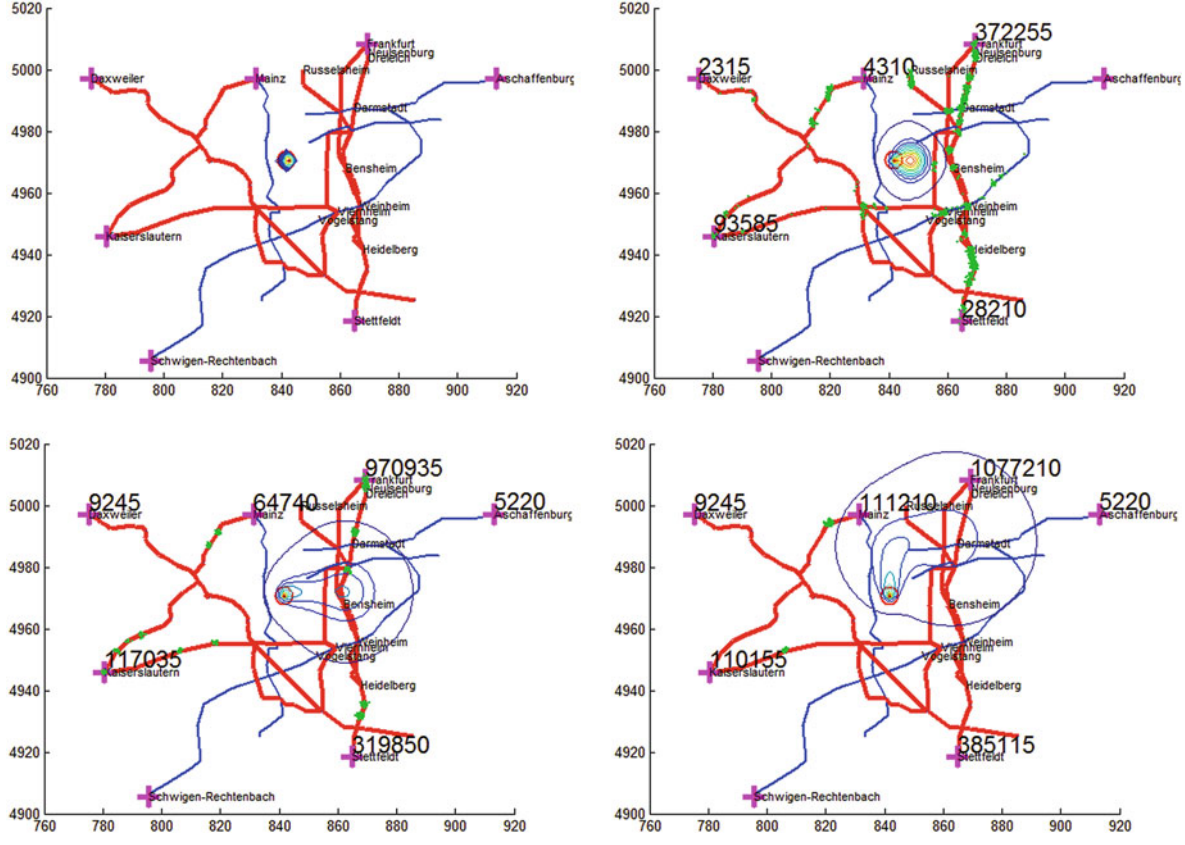

Fig. 4.4 Simulation results for the evacuation of the area around Biblis

immediately outside of the city. The exiting of panicked people in comparison is much slower, but this lessens the effect of congestion significantly, allowing the evacuation on highways be much more faster.

Other important observation is the fact that closing a safety point due to high radiation, the people are redirected. In such a situation, a new safety point is chosen that is closest in the euclidean metric. In some cases this meant that the people had to go back and forth between multiple safety points. A good way to inform people about and a good selection criteria for the new safety point would lead to another optimization problem.

Concerning the given infrastructure, a bottleneck/dangerous road is the federal road B9, which goes in the north-south direction from Mainz to Mannheim. This road is the shortest route across the map domain in the north-south direction and thus selected by a large fraction of the population. It passes very close to the Biblis nuclear power plant as the shortest distance is approximately 3 km between the plant and the road. This means that the road might get congested rather fast and is potentially highly radioactive.

4.4 Conclusion and Comment on Solution

The presented modelling problem requires competencies in different mathematical fields. On the one hand, ideas on network flows by Ahuja et al. [1] are needed to describe the traffic flows and, on the other hand, partial differential equations and their numerical solutions (see e.g. Grossmann et al. [3], Habermann [4]) to manage the spread of radiation. Therefore, the model approaches were developed simultaneously by different small groups and finally combined in the entire simulation.

The complexity of this modeling problem has been reduced in a good way such that first results were available after 1 week only. With the simulation tool at hand, more scenarios than shown here have been analyzed. The students did a great job to carry over various mathematical concepts for the evacuation planning. So the original challenge to provide numerical simulations for different evacuation scenarios has been solved successfully.

Acknowledgments The author wants to thank the group members **Aapo** Koivuniemi (Lappeenranta University of Technology, Finland), **Agney** Kilgi (University of Tartu, Estonia), **Cathleen** Heil (TU Dresden, Germany), **Giuseppe** Codazzi (University of Milan, Italy), **Muhammad** Soecipto Wibowo (TU Kaiserslautern, Germany), **Vanja** Tufvesson (Lund University, Sweden), **Victor** Bayona (Carlos III University of Madrid, Spain), **Yiannis** Hadjimichael (University of Oxford, UK) for the fruitful working atmosphere during the Modelling Week and providing their results.

References

1. R. Ahuja, T.L. Magnati, J.B. Orlin, Network Flows, 846. Prentice Hall (1993)
2. J.A. Bondy, U.S.R. Murty, Graph Theory, Springer (2008)
3. C. Grossmann, H.-G. Roos, M. Stynes, Numerical Treatment of Partial Differential Equations, 591. Springer (2007)
4. R. Haberman Elementary Applied Partial Differential Equations, Prentice Hall, Englewood Cliffs, New Jersey (1987)
5. C.H. Kearny, Nuclear War Survival Skills, updated and expanded in 1987 edition, Oregon Institute of Science and Medicine (1987)
6. P.E. Hart, N.J. Nilsson, B. Raphael, A Formal Basis for the Heuristic Determination of Minimum Cost Paths, IEEE Transactions of Systems Science and Cybernetics, 4, 100–107 (1968)

Chapter 5
Drug Delivery from Ophthalmic Lenses

José Augusto Ferreira

5.1 Introduction

Glaucoma is one of the most common diseases and is a consequence of disorders in the anterior segment of the eye. It is the result of anomalies in the aqueous humor dynamics that lead to increasing intraocular pressure (IOP). This pressure pushes the lens and consequently, the vitreous humor, inducing a pressure on the retina. It can lead to damaging of the optical nerve with subsequent vision loss. In extreme situations, it can even lead to blindness.

The mathematical modelling of drug delivery from a device and its transport until the target tissue requires the knowledge of the physiology of the eye, mainly (1) the anterior segment (2) the dynamics of the aqueous humor, responsible for such anomalous pressure. The increase in IOP is due to an increase of the resistance to the fluid outflow, an increase of the aqueous humor production or even both. It is necessary to identify the physiologic processes involved in aqueous humor production and in its drainage.

Therapeutical contact lenses is one of the drug delivery devices used to treat high IOP. Different drugs have been considered depending on the pathology that leads to IOP increasing:

1. *beta blockers* and *carbonic anhydrase inhibitors* reduce eye pressure by decreasing the production of intraocular fluid;
2. *prostaglandin analogs* induce a reduction of IOP, diminishing the resistance to aqueous humor outflow;
3. *alpha agonists* induce a decrease in the production of fluid and also increase the aqueous humor drainage.

J. A. Ferreira (✉)
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: ferreira@mat.uc.pt

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_5

This article aims to contribute to the mathematical modelling of drug delivery from therapeutical contact lenses to treat glaucoma. We start by presenting the anterior segment of the eye and the dynamics of aqueous humor. The set of anomalous situations that lead to increasing intraocular pressure is then described. An overview on some therapeutic strategies that can be used to treat open-angle glaucoma is presented. The main part of this work concerns a mathematical model that describes the drug release from therapeutic lenses and its evolution in the cornea and anterior chamber. To simplify, the model is built under some assumptions on the phenomena involved as well as on the geometry of the anterior chamber. We conclude remarking that this work was published in the ECMI Annual Report 2016 [9].

5.2 Anterior Segment of the Eye and Aqueous Humor Dynamics

Glaucoma is a group of diseases that lead to the damage of the optical nerve and it is usually associated with an increase of the IOP. This increase is due to pathological modifications of the physiology of the anterior segment of the eye, see Fig. 5.1. This part of the eye is composed by the cornea (the outer boundary), the anterior chamber, the iris, the lens and the ciliary body that define the anterior boundary of the anterior chamber. The cornea is composed by several layers:

1. the epithelium (the outer layer);
2. the stroma;
3. the endothelium (the inner layer). It is coated by a tear film known as precorneal film, see Fig. 5.1.

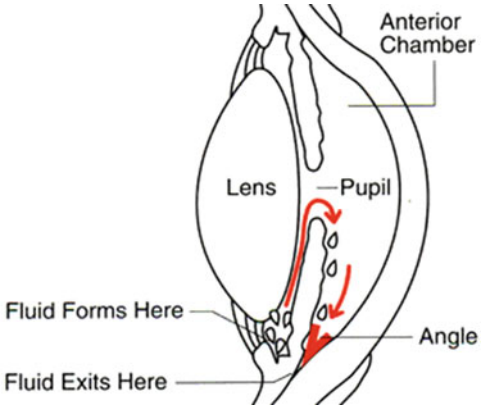

Fig. 5.1 Anatomy of the eye (http://www.theeyecenter.com/educational/005.htm)

The anterior segment of the eye is filled with the aqueous humor. This clear watery fluid is produced by the ciliary epithelium of the ciliary processes located in the ciliary body. It flows from the posterior chamber to the anterior chamber by the narrow space between the posterior iris and the anterior lens and enters in the anterior chamber by the pupil. The aqueous humor leaves the anterior chamber mainly through the trabecular meshwork. It reaches the episcleral venous system via Schlemm's canal. It is also drained from the anterior chamber by the uveoscleral route. The aqueous humor has a multi-purpose nature, see [3]:

1. provides nutrients to the avascular tissues of the anterior segment;
2. removes metabolic excretory products;
3. stabilizes the ocular structure;
4. contributes to the homeostasis of these tissues.

Two main factors contribute to aqueous humor flow:

1. the pressure difference between the trabecular meshwork, that induces a resistance to the outflow (porous structure), and the end of Schlemm's canal, which is similar to the blood pressure (8–10 mmHg);
2. the temperature difference near the cornea and lens.

Two convective flows are then induced that are driven by a pressure gradient and a temperature gradient. The balance between the aqueous humor production and its drainage maintains the IOP stable. The drainage through the uveoscleral route appears to be pressure independent, see [10].

5.3 Glaucoma

The *closed-angle glaucoma* can be induced by different causes that influence an iris dilation and its adheration to the lens.

The aqueous humor is produced in the ciliary body involving complex phenomena that include ultrafiltration, diffusion and active transport (active secretion). The ultrafiltration occurs in the capillaries of the ciliary processes and it is a passive movement of water and water soluble substances across cell membranes, diffusion of solute takes place in the tissue between the capillaries and the posterior chamber in response to concentration gradient. Active transport occurs in nonpigmented epithelial cells and it is the main responsible for the aqueous humor formation, see [10, 15]. An abnormal production of the aqueous humor can lead to increased IOP.

In the human eye, 75% of the resistance to the fluid outflow is due to the trabecular meshwork and 25% occurs beyond the Schlemm's canal. Trabecular meshwork's resistance to the drainage of aqueous humor is due to the hydration of the trabecular meshwork structure that can cause obstruction of its structure. Such obstruction is also associated with the formation of deposits within this tissue. Recently, the region of the trabecular meshwork that is responsible by the IOP regulation was identified: the *juxtacanalicular tissue*, which is adjacent

to Schlemm's canal. To keep the aqueous humor flow channels open in the juxtacanalicular tissue, the extracellular matrix of this tissue presents a continuous remodelling. An interference in this remodelling process compromises the aqueous humor drainage and increases the IOP, see [14].

The hypothesis that the biomechanical properties of Schlemm's canal contribute to the aqueous humor outflow was studied, for instance, in [2, 16]. It was observed that the pore formation is a mechanosensitive process: an increase of the biomechanical strain induces an increase of the porous density. Changing this biomechanical behaviour, it was observed that the porous formation decreases, leading to increased IOP.

5.4 Therapeutics Strategies to Open-Angle Glaucoma

To decrease the IOP it is necessary to attack the anterior segment of the eye fortress and introduce drugs in the anterior chamber. This fortress is defended by the *tear fluid barrier*, the tear film that coats the corneal epithelium, the permanent blink, the cornea (lower impermeable structure) and the blood-aqueous barrier. Eye drops are the most used ocular route to administer drugs. However the drug bioavailability in the anterior chamber is very low. The tear film turnover is one of the main contributors to this fact. The drug residence time in the corneal epithelium is equal to 5–6 min before being completely washed way. The permanent continuous blinking removes the mixture of drug solution with tear fluid from the corneal epithelium to the nasolacrimal ducts.

The low permeability of the corneal layers also contributes to the reduced amount of drug that reaches the anterior chamber. Less than 5% of the drug present in the eye drops reaches the ocular tissue. The use of the systemic rout to delivery drug into the anterior segment of the eye is also very inefficient. In fact, the blood-aqueous barrier restricts the entry of drugs from the blood stream into the posterior segment and consequently, to the anterior chamber. The poor eye drug absorption requires repeated applications during long periods to achieve drug concentrations in the anterior chamber within the therapeutic window.

Different drugs have been used to decrease the IOP and they depend on the specific pathology. For instance, if the increased IOP is due to an anomalous production of aqueous humor, β-blockers, α-agonists and carbonic anhydrase inhibitors lead to decreasing of aqueous humor inflow. Other approaches use prostaglandin analogs to enhance the uveoscleral outflow or muscarinic agonists to enhance the trabecular outflow, see [17].

Several approaches have been followed to avoid the limitations of classical topical administration, like, for instance, the use of viscosity enhancers, mucoadhesive and lens which aim at increasing the drug corneal residence time. Other strategies, like the use of penetration enhancers, prodrugs and colloidal systems aim to increase the corneal permeability, see [19]. The purpose of such strategies is to delivery drug

into the anterior chamber at a sustained and controlled rate complying the drug concentration in the target tissue to therapeutic window .

Since the nineties, several types of therapeutical contact lenses have been proposed by researchers to increase the drug residence time in the cornea. Without being exhaustive we mention

1. soaked contact lenses, see [1];
2. compound contact lenses with a hollow cavity, see [18];
3. entrapment of proteins, cells and drugs by polymerization of hydrogel monomers, see [20];
4. biodegradable contact lenses, see [4].

We remark that the corneal drug residence time for soaked lenses increases (it is around 30 min) by increasing the drug bioavailability. However, such increase is not significantly high because there are no barriers to the delivery and the loading is limited by the drug solubility. Compound lenses with hollow cavities as drug reservoirs present lower permeability to oxygen and carbone dioxide. In the polymerization process the drug can loose its therapeutic characteristics, see [21].

To delay the drug delivery process such that the corneal drug residence time and loading increases, several authors propose to encapsulate the drug in polymeric particles that are dispersed in the lens, see [5, 11–13]. In this case, the drug can also be dispersed in the polymeric structure leading to increasing drug loading. Such drug can be in three states in the polymer: free, bounded or encapsulated. The dispersed drug, when in contact with the tear fluid, is immediately released followed by the delivery of the bounded drug. The release on the encapsulated drug is delayed by the particles structure and the corneal drug residence time increases significantly. One of the main advantages of such devices is the possibility to build lenses that deliver the drug with a pre-defined profile.

5.5 Mathematical Modelling of Drug Delivery from Therapeutic Lenses

Building a mathematical model that describes the drug delivery process from a specific device and its transport to the target tissue is a complex work that requires different tasks. Let us consider the case of a therapeutic lens used to treat open-angle glaucoma, that is to deliver a specific drug in the anterior chamber to decrease IOP. Different tasks can be identified in the mathematical modelling of this drug delivery process.

The drug release and transport involves a set of complex phenomena presented before:

1. the drug release from the polymeric structure;
2. its clearance by the tear turnover;
3. the drug transport through the different layers of the cornea;

4. the drug transport and its drainage by the aqueous humor;
5. the dynamics of the fluid, which includes the aqueous humor production in the ciliary body, its transport in the posterior chamber and in the anterior chamber, its drainage through the trabecular meshwork and uveoscleral route, and its transport in the Schlemm's canal.

It should be remarked that a mathematical model describing all phenomena taking place will be very complex and its numerical simulation will be a very difficult task. Therefore it is necessary to identify the main phenomena involved and the spatial domains were they occur.

The drug delivery from a lens and its transport in the anterior chamber is naturally a three dimensional problem. However, to simplify the geometry of the domain we reduce the domain to a two dimensional one using the symmetry of the anterior segment of the eye and lens. We remark that in [6] the mathematical model is defined in bounded intervals for the lens and cornea and the anterior chamber was considered as a sac with passive role in the process. The mathematical models introduced in [7, 8] were defined in a two dimensional domain and the influence of the aqueous humour motion was taken into account. It is assumed that the fluid enters the anterior chamber through the space between the iris and the lens that we denote by $\Gamma_{ac,i}$, see Fig. 5.2, and it is drained through the trabecular meshwork. The tear turnover and the uveoscleral drainage were neglected, the aqueous humor production is not explicitly considered as well as its transport through the trabecular mesh until the Schlemm's canal. The last transport is described by a condition on the flux that leaves the anterior chamber through $\Gamma_{ac,tm}$. The aqueous humor production is described by a boundary source term specified at the fluid entrance $\Gamma_{ac,i}$. These assumptions allow us to consider the domain plotted in Fig. 5.2.

We point out that the properties of the polymer used to construct the lens and particles entrapping the drug should be provided. We assume that the drug is dispersed in the polymeric lens presenting three different states, free, bounded and entrapped, while the cornea is composed by a single layer.

Let Ω_ℓ, Ω_c and Ω_{ac} denote the lens, the cornea and the anterior chamber, respectively. By c_f, c_b and c_e we denote the free, bound and entrapped drug

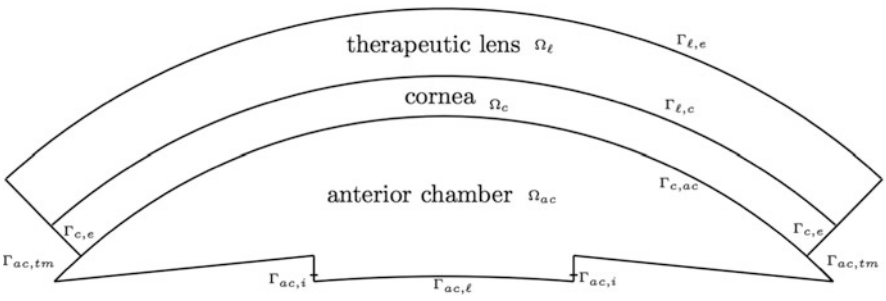

Fig. 5.2 Spatial domain

concentrations (g/m^3). In what follows we specify the phenomena and their mathematical laws in each domain:

1. Ω_ℓ—In the lens three different phenomena occur: the links between the polymeric chains and the drug molecules break and the bounded drug is converted in free drug that diffuses. Let $\lambda_{b,f}, \lambda_{e,f}$ be the transference coefficients $(1/s)$ between bounded and free drug and entrapped and free drug, respectively, and $\mathbf{D}_{f,\ell}$ denote the free drug diffusion tensor (m^2/s). Then the behaviour of the free and bound drugs is described by the diffusion equations

$$\begin{cases} \dfrac{\partial c_f}{\partial t} = \nabla \cdot (\mathbf{D}_{f,\ell} \nabla c_f) + \lambda_{b,f}(c_b - c_f) \\ \qquad\qquad + \lambda_{e,f}(c_e - c_f), \\ \dfrac{\partial c_b}{\partial t} = -\lambda_{b,f}(c_b - c_f), \\ \dfrac{\partial c_e}{\partial t} = -\lambda_{e,f}(c_e - c_f), \end{cases} \tag{5.1}$$

   in $\Omega_\ell \times (0, T]$, where $T > 0$ denotes a fixed time.
2. Ω_c—Only the free drug is released from the lens and enters in the cornea where it diffuses. If $\mathbf{D}_{f,c}$ represents the free drug diffusion tensor in the cornea then

$$\frac{\partial c_f}{\partial t} = \nabla.(\mathbf{D}_{f,c} \nabla c_f) - \lambda_c c_f \tag{5.2}$$

   in $\Omega_c \times (0, T]$. Equation (5.2) is established assuming that the clearance of the drug occurs here being λ_c the clearance rate $(1/s)$.
3. Ω_{ac}—In the anterior chamber the free drug diffuses and its transported by the aqueous humor to the trabecular meshwork. The evolution of c_f is described by the following convection-diffusion-reaction equation

$$\frac{\partial c_f}{\partial t} + \nabla(\mathbf{v} c_f) = \nabla.(\mathbf{D}_{f,ac} \nabla c_f) - \lambda_{ac} c_f \tag{5.3}$$

   in $\Omega_{ac} \times (0, T]$. In Eq. (5.3), $\mathbf{D}_{f,ac}$ and λ_{ac} $(1/s)$ represent the drug diffusion tensor and the drug clearance rate in the aqueous humor. As the aqueous humor is mainly composed by water, and its dynamics is mainly driven by the IOP, the velocity field $\mathbf{v}$ can be described by the incompressible Navier–Stokes equations

$$\begin{cases} \rho \dfrac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v} \cdot \nabla)\mathbf{v} - \nu \boldsymbol{\Delta} \mathbf{v} + \nabla p = \mathbf{0}, \\ \nabla \cdot \mathbf{v} = 0, \end{cases} \tag{5.4}$$

in $\Omega_{ac} \times (0, T]$. In system (5.4), p represents the intraocular pressure, ρ the density of the aqueous humor and ν its kinematic viscosity.

The velocity field $\mathbf{v}$ is time and space dependent if the drug molecules have a therapeutic effect in the trabecular meshwork. Otherwise the velocity does not change in time and then the system of equations (5.4) should be replaced by steady Navier–Stokes equations.

The boundary conditions are specified now. We start by defining the boundary conditions for drug concentration:

1. Let $\Gamma_{\ell,e}$ be the exterior boundary of Ω_ℓ, see Fig. 5.2. We assume that this surface is isolated, meaning that the drug mass flux is zero. Then

$$\mathbf{D}_{f,\ell} \nabla c_f \cdot \boldsymbol{\eta} = 0 \text{ on } \Gamma_{\ell,e} \times (0, T], \tag{5.5}$$

   where $\boldsymbol{\eta}$ denotes the outward unit normal to Ω_ℓ on $\Gamma_{\ell,e}$.
2. By $\Gamma_{c,e}$ we represent the exterior boundary of Ω_c, see Fig. 5.2. As no drug mass flux occur on $\Gamma_{c,e}$ we have

$$\mathbf{D}_{f,c} \nabla c_f \cdot \boldsymbol{\eta} = 0 \text{ on } \Gamma_{c,e} \times (0, T], \tag{5.6}$$

   where $\boldsymbol{\eta}$ denotes the outward unit normal to Ω_c on $\Gamma_{c,e}$.
3. On the fluid outflow boundary $\Gamma ac, tm$ (see Fig. 5.2) we assume that the drug mass flux depends on a function $A_c(c_f)$ that reflects the drug effect in the increasing of the porosity of the trabecular mesh. This function should increase as c_f increases reaching a maximum threshold. Therefore we assume that

$$\mathbf{J} \cdot \boldsymbol{\eta} = A_c(c_f) c_f \text{ on } \Gamma_{ac,tm} \times (0, T], \tag{5.7}$$

   where $\mathbf{J} = -\mathbf{D}_{f,ac} \nabla c_f + \mathbf{v} c_f$, and $\boldsymbol{\eta}$ denotes the outward unit normal to Ω_c on $\Gamma_{ac,tm}$.
4. In the boundary $\Gamma_{ac,\ell} \cup \Gamma_{ac,i}$ (see Fig. 5.2) we take

$$\mathbf{J} \cdot \boldsymbol{\eta} = 0 \text{ on } (\Gamma_{ac,e} \cup \Gamma_{ac,i}) \times (0, T], \tag{5.8}$$

   where $\boldsymbol{\eta}$ denotes the outward unit normal to this portion of the boundary.

The boundary conditions for the Navier–Stokes equations are specified in what follows:

1. In the inflow boundary $\Gamma_{ac,i}$ we assume that the normal component of the velocity is known

$$\mathbf{v} \cdot \boldsymbol{\eta} = v_{in} \text{ on } \Gamma_{ac,i} \times (0, T]. \tag{5.9}$$

2. There are several approaches to define the boundary condition when the pressure is known. One of them is to consider

$$(\nu \nabla \mathbf{v} - p\boldsymbol{I})\boldsymbol{\eta} = -p_0 \boldsymbol{\eta} \text{ on } \Gamma_{ac,tm} \times (0, T], \tag{5.10}$$

where p_0 denotes the pressure in Schlemm's canal which is taken equal to the blood pressure and $\boldsymbol{I}$ is the identity matrix.
3. On $\partial\Omega_{ac}\backslash(\Gamma_{ac,i} \cup \Gamma_{ac,tm})$ the normal component of the velocity is null

$$\mathbf{v} \cdot \boldsymbol{\eta} = 0 \tag{5.11}$$

on $(\Gamma_{c,ac} \cup \Gamma_{ac,e}) \times (0, T]$.

For interface boundaries we assume the next conditions for the free drug concentration.

1. Interface between the lens and cornea:

$$\begin{cases} \mathbf{D}_{f,\ell} \nabla c_{f,\ell} \cdot \boldsymbol{\eta} = \mathbf{D}_{f,c} \nabla c_{f,c} \cdot \boldsymbol{\eta} \\ -\mathbf{D}_{f,\ell} \nabla c_{f,\ell} \cdot \boldsymbol{\eta} = A_{\ell,c}(c_{f,\ell} - c_{f,c}) \end{cases} \tag{5.12}$$

on $\Gamma_{\ell,c} \times (0, T]$, where $\boldsymbol{\eta}$ denotes the outward unit normal to Ω_ℓ on $\Gamma_{\ell,c}$. Here $c_{f,\ell}$ and $c_{f,c}$ represent the drug concentrations in the lens and cornea, respectively, and $A_{\ell,c}$ (m/s) denotes the partition coefficient on $\Gamma_{\ell,c}$.
2. Interface between the cornea and anterior chamber:

$$\begin{cases} \mathbf{D}_{f,c} \nabla c_{f,c} \cdot \boldsymbol{\eta} = \mathbf{D}_{f,ac} \nabla c_{f,ac} \cdot \boldsymbol{\eta} \\ -\mathbf{D}_{f,c} \nabla c_{f,c} \cdot \boldsymbol{\eta} = A_{c,ac}(c_{f,c} - c_{f,ac}) \end{cases} \tag{5.13}$$

on $\Gamma_{c,ac} \times (0, T]$, where here $c_{f,ac}$ denotes the drug concentration in the anterior chamber, $\boldsymbol{\eta}$ the outward unit normal to Ω_c on $\Gamma_{\ell,c}$ and $A_{c,ac}$ (m/s) represents the partition coefficient on $\Gamma_{c,ac}$.

Finally, the initial conditions should be imposed to complete the system of partial differential equations (5.1)–(5.4) complemented with the boundary conditions (5.5)–(5.11) and interface conditions (5.12) and (5.13). We impose the following

$$c_f(0) = \begin{cases} c_{f,0} \text{ in } \Omega_\ell \\ 0 \text{ in } \Omega_c \cup \Omega_{ac} \end{cases}$$

$$c_b(0) = c_{b,0} \text{ in } \Omega_\ell,$$

$$c_e(0) = c_{e,0} \text{ in } \Omega_\ell,$$

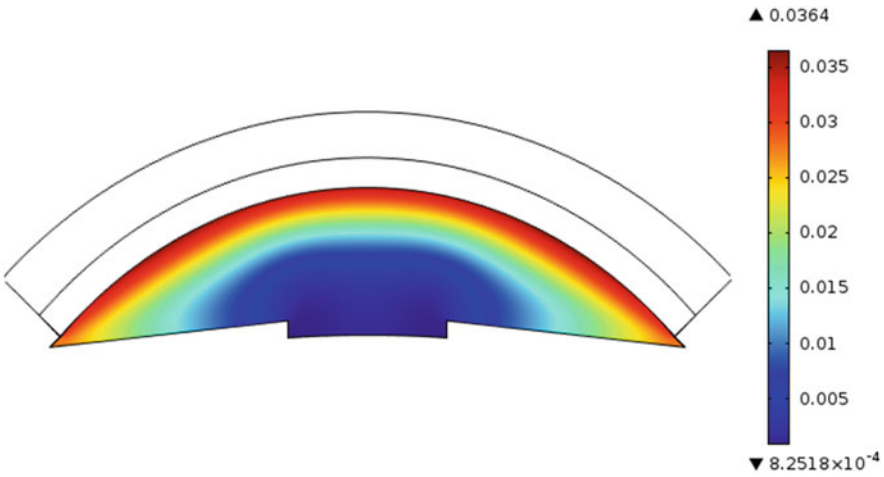

Fig. 5.3 Drug distribution in anterior chamber after 20 min for a lens

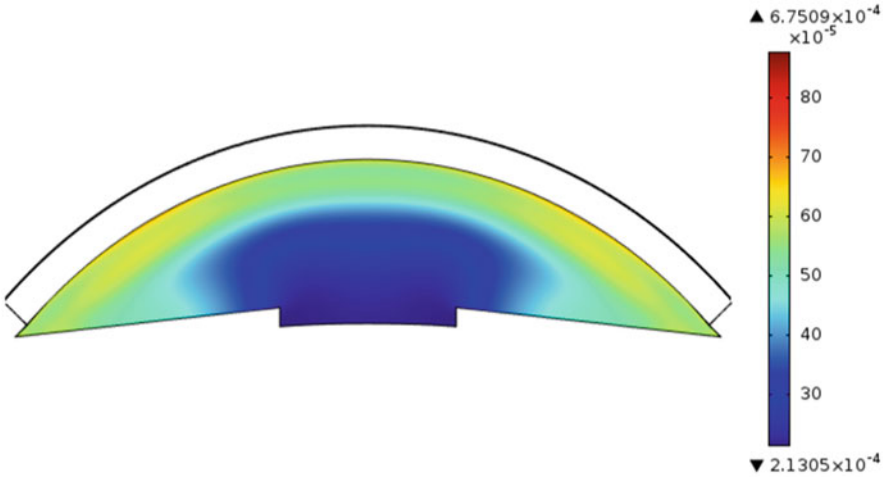

Fig. 5.4 Drug distribution in anterior chamber after 20 min for an eye drop

and

$$\mathbf{v}(0) = \mathbf{v}_0 \text{ in } \Omega_{ac},$$

where $c_{f,0}$, $c_{b,0}$, $c_{e,0}$ and $\mathbf{v}_0$ are known functions.

In Figs. 5.3 and 5.4 we present two typical plots for the drug distribution included before in [8]. Figure 5.3 illustrates the drug distribution in the anterior chamber when a lens is used where the drug is dispersed in the polymeric structure and entrapped in particles. The results obtained for the drop case are plotted in Fig. 5.4.

From the plots we can infer that the amount of drug that reaches the anterior chamber with lens is higher than for drops due to the initial loading. Moreover the drug release from a therapeutic lens is a slower process due to the polymeric barrier for the dispersed and entrapped drugs.

5.6 Conclusions

This work aims to contribute to the mathematical modelling of the drug delivery from a drug delivery device—the therapeutic lens—used to decrease IOP in a glaucoma scenario. The model is established under several simplifying assumptions in what concerns the geometry of the spatial domain and the phenomena involved.

Some new models can now be deduced with increasing complexity adding new phenomena and changing the geometry of the spatial domain to include new tissues or organs. For instance the tear turnover can be included requiring the inclusion of a tear film layer in the spatial domain. New equations should be added to the existing set of partial differential equations that describe the drug dynamics in this fluid. Different corneal layers—epithelium, stroma and endothelium—can also be added to the model where the drug presents different diffusion properties. Consequently, the diffusion equation in the cornea should be replaced by three diffusion equations with the correspondent compatibility conditions on the contact surfaces between layers. If the trabecular meshwork is considered (Ω_{tm}) with or without juxtacanalicular tissue, the drug transport is defined by an equation similar to (5.3) where the convective velocity $\mathbf{v}$ is given by Darcy equation (coupled with an incompressibility constraint)

$$\begin{cases} \mathbf{v} = -\dfrac{\mathbf{K}}{\mu\phi}\nabla p \\ \nabla \cdot \mathbf{v} = 0 \end{cases} \tag{5.14}$$

in $\Omega_{tm} \times (0, T]$. In (5.14) $\mathbf{K}$ denotes the permeability tensor and the porosity coefficient is represented by ϕ. It should be stressed that in this case the coupling between the Navier–Stokes equations (5.4) and (5.14) is a challenging topic namely due to the conditions required on the boundary of the trabecular meshwork that is contact with the anterior chamber.

There is a compromise between the complexity of the mathematical model and its utility to predict the IOP evolution in different scenarios. In fact, the number of parameters needed increases with the complexity and some of them are not known.

Acknowledgments This work was partially supported by the Centre for Mathematics of the University of Coimbra—UID/MAT/00324/2019, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

1. Bourlais, C., Acar, L., Sado, O., Needham, T., Leverge, R.: Ophthalmic drug delivery systems–recen tadvances. Prog. Ret. Eye Res. **17**, 33–58 (1998)
2. Braakman, S., Pedrigi, R., Read, A., Smith, J., Stamer, W., Ethier, C., Overby, D.: Biomechanical strain as a trigger for pore formation in schlemm's canal endothelial cells. Exp. Eye Res. **127**, 224–235 (2014)
3. Chader, G., Thassu, D.: Eye anatomy, physiology, and ocular barriers. Basic considerations for drug delivery. In : Ocular Drug Delivery Systems: Barriers and Application of Nanoparticulate Systems, Thassu. D., Chader, G., (eds), 17–40, CRC Press, London—New–York (2012)
4. Ciolino J., Hoare, T., Iwata, N., Behlau, I., Dohlman, C., Langer, R., Kohane, D.: A drug-eluting contact lens. Invest. Ophthal. Visual Sci. **50**,3346–3352 (2009)
5. Ferreira, J.A., Oliveira, P., Silva, P., Carreira, A., Gil, H., Murta, J.: Sustained drug release from contact lens. Comp. Model. Eng. Sci. **60**, 151–179 (2010)
6. Ferreira, J.A., Oliveira, P., Silva, P., Murta, J.: Drug delivery: from contact lens to the anterior chamber. Comp. Model. Eng. Sci. **71**, 1–14 (2011)
7. Ferreira, J.A., Oliveira, P., Silva, P.: Controlled drug delivery and medical applications. Chem. Biochem. Eng. Q. **26**, 331–342 (2012)
8. Ferreira, J.A., Oliveira, P., Silva, P., Murta, J.: Numerical simulation of aqueous humor flow: from healthy to pathologic situations. App. Math. Comp. *226*, 777–792 (2014)
9. Ferreira, J. A.: Drug delivery from ophthalmic lenses. In: Mathematics with Industry: Driving Innovation, ECMI Annual Report 2016, 34–42 (2016)
10. Goel, M., Picciani, R., Lee, R., Bhattacharya, S.: Aqueous humour dynamics: a review. Open Ophthalmol. J. **52**,52–59 (2010)
11. Gulsen, D., Chauhan, A., Ophthalmic drug delivery from contact lens. Invest. Ophthal. Visual Sci. **45**, 2342–2347 (2004)
12. Gulsen, D., Chauhan, A.: Dispersion of microemulsions drops in HEMA hydrogel: a potential ophthalmic drug delivery vehicle. Int. J. Pharm. **292**, 95–117 (2005)
13. Jung, H., Jaoude, M., Carbia, B., Plummer, C., Chauhan, A.: Glaucoma therapy by extended release of timolol from nanoparticle loaded silicone–hydrogel contact lenses. J. Control. Release **165**, 82–89 (2013)
14. Keller, K., Acot, T.: The juxtacanalicular region of ocular trabecular meshwork: a tissue with a unique extracellular matrix and specialized function. Journal of Ocular Biology **1**,1–7 (2013)
15. Kiel,J., Hollingsworth, M., Rao, R., Chen, M., Reitsamer, H.: Ciliary blood flow and aqueous humor production. Prog. Ret. Eye Res. **30**,1–17 (2011)
16. Overby, D., Zhou, E., Vargas-Pinto, R., Pedrigi, R., Fuchshofer, R., Braakman, S., Gupta, R., Sherwood, J., Vahabikashi, A., Dang, Q., Kim, J., Ethier, R., Stamer, W., Fredberg, J., Johnson, M.: Altered mechanobiology of schlemm's canal endothelial cells in glaucoma. Proc. Nat. Acad. Sci. **111**, 13876–13881 (2014)
17. McLan, N., Moroi, S.: Clinical implications of pharmacogenetics for glaucoma. Pharmacogenomics J. **30**, 197–201 (2003)
18. Nakada, K., Sugiyama, A.: Process for producing controlled drug-release contact lens, and controlled drug-release contact lens thereby produced. United States Patents, page 6027745 (1998)
19. Rupenthal, I.: Ocular drug delivery technologies: Exciting times ahead. ONdrugDelivery **54**, 7–11 (2015)
20. Santos, J., Alvarez-Lorenzo, C., Silva, M., Balsa, L., Couceiro, J., Torres-Labandeira, J., Concheiro, A.: Soft contact lenses functionalized with pendant cyclodextrins for controlled drug delivery. Biomaterials **30**, 1348–1355 (2009)
21. Xinming, L., Yingde, C., Lloyd, A., Mikhalovsky, S., Sandeman, S., Howel, C., Liewen, L.: Polymeric hydrogels for novel contact lens–based ophthalmic drug delivery systems: a review. Contact Lens & Anterior Eye **31**, 57–64 (2008)

Chapter 6
The Zombie Invasion

Jarosław Gruszka

6.1 Introduction

A possibility of making a dead person become alive again has been fascinating people for thousands of years. Zombies—also called *the undead*—can be considered a perfect example of that fascination. People have been imagining those creatures in a vast number of ways—almost every book or every film that zombies appeared in, approached the topic differently. There exists however some specific set of characteristics that zombies share in most pieces of works that mention their existence. They are usually described as very anomalous beings, extremely hostile to the humankind and willing to entirely destroy it or turn all 'ordinary' people into zombies. Hence, the zombie invasion could hypothetically be considered a threat for the entire human civilisation. Regardless of what the majority of people might think of it, there are still thousands who strongly believe that the ultimate end of the human civilisation will be a widespread rise of hordes of zombies.

Nevertheless, most of us would admit that the zombie apocalypse is not the most plausible scenario of the how our civilisation ends, especially considering that the today's world is facing multiple more tangible issues. However, from some perspectives, an event like this can be considered worth looking at with a scientific eye and one can name at least few reasons. First and foremost—modelling an event like the zombie invasion differs significantly from the 'classical' modelling process, i.e. creating a model of an existing and real-life phenomenon. In case of creating a mathematical description of anything we know from the world around us, the modeller needs to be very concious of the tools and concepts they employ. This is obviously driven by the fact that any scientific theory aiming to describe a real-

J. Gruszka (✉)
Hugo Steinhaus Center, Faculty of Pure and Applied Mathematics, Wroclaw University of Science and Technology, Wroclaw, Poland
e-mail: jaroslaw.gruszka@pwr.edu.pl

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_6

life phenomenon must necessarily be compared to the experimental data. And it is absolutely clear as well that scientific results are evaluated based on the agreement between scientist's theoretical model and the observations. Although this should be very natural for everyone—it is clear that we all want our theories to describe the surrounding world accurately—from some other perspective this urge for obtaining valuable and sterling results can be seen as a factor limiting creativeness. It is difficult to argue that the pervasive need for scientific usability makes us try to follow the paths which look most promising from the perspective of measurable results, which can be immediately applied. Arguably, it is also worth to look at the scientific progress form a kind of *reversed* perspective. Modelling a more abstract, less realistic ideas or processes (like, in this case, the zombie invasion for example) enables us to analyse the issue more freely, without unnecessary borders and limits, lets us to think outside of the box and somewhat play with the problem. Results obtained this way are not usually ready to be directly used right after developing them but they may change scientist's perspective and their point of view, they may feature new research methods and may become an inspiration for totally new studies which might turn out to unexpectedly pop out in some future research.

This was the actual reason for our research project of modelling the zombie invasion—not only does it sound more intriguing and more approachable in reception (although we admit that this can serve as an asset too), but it also creates an opportunity to do science differently, in a way oriented on methods rather than the results. So—as our main target we picked creating a model of an extremely fast-spreading epidemic totally from scratch, with minimal number of pre-settled assumptions, but general enough to possibly be 'calibrated' to some known contagious disease in the future. The only requirement, in some ways defining the direction of the development of the model was that the spreading scheme should be based on the population density maps and should embrace a premiss (taken purely from the common sense) that more densely populated areas should be more vulnerable to faster epidemic spreading. Agreeing only for those high-level assumptions marked a point where the actual project development could have been started.

6.2 Data Preparation

As mentioned in the Introduction, one of the very few assumptions that the project had was that the base for the entire model should be density population maps. It turns out that high-quality maps of this kind can easily be found on-line. One of the numerous web pages that can be used for this purpose was World Population Density [6]. There one can find the density population maps for the entire world, presented in a very clear manner. An interface of the webpage has been shown in Fig. 6.1.

Just like in case of most maps of this kind, the density population was represented by a colour scale. In order to use information that such maps convey, it is obviously

Fig. 6.1 Interface of the world population density webpage [6]

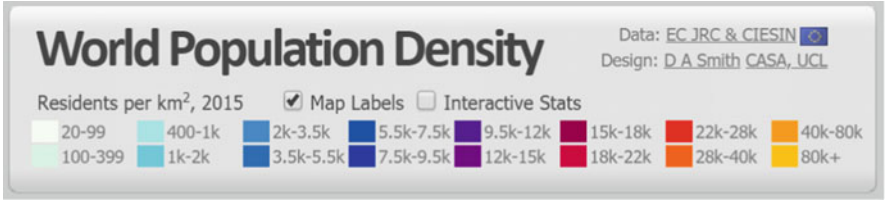

Fig. 6.2 Legend for the data population maps available under http://luminocity3d.org [6]

needed to quantify it, i.e. move it from a format of a picture to some specific numeric data type. The picture itself can be viewed as a structure composed of pixels—tiny, square pieces of the image. Each pixel can be uniquely identified by its position in a picture $(i, j) \in (1, n) \times (1, m)$, where n is height of the image and m is its width, expressed in the number of pixels. Having that said, any method that can be used to move the entire picture from any on-line location into our computer's workspace should start by scanning the image, pixel by pixel. This can be done effectively in many different programming languages and author's language of choice was Python 3 [7] with the usage of an image processing library called Pillow [2] and other side modules [4, 5]. The entire purpose of scanning all pixels of the image is actually to save the data of a colour value of each of them. The most common way to keep the information about the colour in a numeric way is by analysing what is called the RGB value of it. From the mathematical point of view, RGB is simply a vector of three numbers (r, g, b), $r, g, b \in [0, 255]$ each of them representing the saturation of the red, blue and green colour component respectively. Only having the value of the RGB for every pixel of the image one can actually start analysing it.

After getting the values of the colours of the map, the next step was to translate the RGB values of the colours into the actual values of the population density. For this purpose, the map legend should obviously be used. It has been presented in the Fig. 6.2.

The map's legend explicitly mentions 16 colours and precisely describes what range of population density (expressed in residents per square kilometre) each of them represents. However, one can expect that the map also features some smooth colour transitions between the colours that the legend mentions. Therefore the map needs to be *sharpened*—adjusted in a way so that it only contained the colours that the legend has. To this end, for each pixel, its colour needs to be substituted by one of the 16 map legend colours—by the one which is closest to the original one. The measure of *closeness* of colours can simply be defined as a sum of square differences between the RGB values of those two colours. Hence, the new, sharpened colour of a given pixel can be described as

$$(\hat{r}^{(i,j)}, \hat{g}^{(i,j)}, \hat{b}^{(i,j)}) = (r_m, g_m, b_m),$$
$$m = \underset{k\in\{1,2,\ldots,16\}}{\arg\min} ((r^{(i,j)} - r_k)^2 + (g^{(i,j)} - g_k)^2 + (b^{(i,j)} - b_k)^2)$$

where $(\hat{r}^{(i,j)}, \hat{g}^{(i,j)}, \hat{b}^{(i,j)})$ is the sharpened version of the original colour $(r^{(i,j)}, g^{(i,j)}, b^{(i,j)})$ of a given pixel and $(r_k, g_k, b_k), k \in \{1, 2, \ldots, 16\}$ is the RGB value of each of the 16 colours of the map.

Such a sharpened map can almost immediately be transformed into a matrix of the size $n \times m$, in which each entry corresponds to the number of citizens per square kilometre associated with the area represented by a pixel in that same position in the scanned map. The problem was that the legend only showed a range of the number of inhabitants related to a given colour, not the precise number. Since dealing with data intervals is much more difficult than with single numbers, for each interval (bound to each colour of the legend) we took a mean value of the interval's range. For example, a vivid red colour on the map represented density population between 22,000 and 28,000 people per square kilometre, so we assumed that each pixel of that colour on our sharpened map indicates a piece of land with the average density of 25,000 people per km^2. For the legend's last entry which was described as '80k+', we assumed the density of 90,000 per km^2. To simplify the data in the matrix even further, we decided to divide each value by the population of the most populous pixel read from the map, hence obtaining only relative values between 0 and 1 (obviously, this is with no loss of generality, as we can always multiply all entries by the divisor and have the original values back). Now we needed a way to visually represent the entries of the matrix that we prepared. Since the density is only a starting point for the model (ultimately we want to visualise spreading zombies), it seemed reasonable to save the vivid colours for later and stick to grey scale for now. We decided to assign white colour to the pixels representing zero population (like oceans) and black to the pixels with highest population density on a given map (usually—centres of the biggest cities). This can easily be done using the following equation

$$(\bar{r}^{(i,j)}, \bar{g}^{(i,j)}, \bar{b}^{(i,j)}) = (1 - \rho^{(i,j)}) \cdot (255, 255, 255), \tag{6.1}$$

where $(\bar{r}^{(i,j)}, \bar{g}^{(i,j)}, \bar{b}^{(i,j)})$ represents the grey scale colour value of the pixel in position (i, j) and $\rho^{(i,j)}$ is the entry in position (i, j) of the matrix containing data of population density normalised to only have values between 0 and 1 (obtained as described above).

This entire data processing flow has been presented in Fig. 6.3 and its ultimate visual effect—in Fig. 6.4.

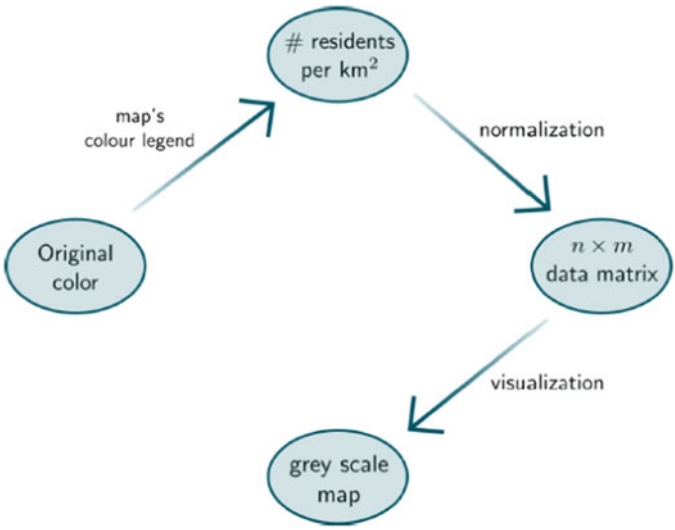

Fig. 6.3 Scheme illustrating the data pre-processing for the project

Fig. 6.4 Illustration of the visual effects of the data pre-processing

6.3 Model Description

Once we had the data prepared, the actual modelling part could have been started. Our starting point was the normalised density population matrix ρ. We now wanted to treat the elements of this matrix as indicators of how many people live in a given area and further—we wanted to split those people into two groups—healthy ones (denoted by H) and the zombies (which we will mark as Z). As we were interested in the evolution of those quantities in both space and time, they should both be space- and time-dependent, so that $H_t^{(i,j)}$ and $Z_t^{(i,j)}$ should be understood as indicators of the number of individuals occupying area represented by map position (i, j) at time t, which belong to healthy population or to the zombies, respectively. We now wanted to describe the mechanism of how the sizes of these two groups change in time, i.e. how people turn into zombies. We created 3 models of that process which vary in assumptions that were made and we shall now describe them in the order of ascending complexity.

6.3.1 Simple Deterministic Model

In the simplest version of the model, some initial population of zombies appears at a given point on the map, at time $t = 0$, it starts spreading to the neighbouring places with strength proportional to the number of zombies in the outbreak. To put it into a precise mathematical formulation, we came up with the following equations, encoding the time evolution of the zombie invasion

$$H_t^{(i,j)} = H_{t-1}^{(i,j)} - H_{t-1}^{(i,j)} \cdot c_t^{(i,j)}, \tag{6.2}$$

$$Z_t^{(i,j)} = Z_{t-1}^{(i,j)} + H_{t-1}^{(i,j)} \cdot c_t^{(i,j)}, \tag{6.3}$$

where by $c_t^{(i,j)}$ we denote what we call *the contamination*, i.e. the fraction of healthy people from a given area (i, j), which were turned into zombies at time t. More advanced readers may note the similarity of this description to the one of the classical SI model [3], however, since we were dealing not only with time-related but also with spacial aspects of the invasion spreading, we continued to develop our model ourselves, instead of trying to adjust the classical one to make use of it.

Taking a closer look at the structure of the equations themselves, we can note that they are recursive in parameter t, which means that the current state of both $H_t^{(i,j)}$ and $Z_t^{(i,j)}$ depends directly on $H_{t-1}^{(i,j)}$ and $Z_{t-1}^{(i,j)}$. This means that for the description

to be complete, we need to set up the initial conditions. Let us set them up in the following way

$$H_0^{(i,j)} = \begin{cases} \rho^{(i,j)} & \text{for } (i,j) \neq (u,v), \\ (1-\gamma)\rho^{(i,j)} & \text{for } (i,j) = (u,v) \end{cases} \tag{6.4}$$

$$Z_0^{(i,j)} = \begin{cases} 0 & \text{for } (i,j) \neq (u,v), \\ \gamma\rho^{(i,j)} & \text{for } (i,j) = (u,v) \end{cases}. \tag{6.5}$$

where $(u, v), 1 \leqslant u \leqslant n, 1 \leqslant v \leqslant m$ is a place where the outbreak occurs and $\gamma \in (0, 1]$ is the ratio of people who become zombies in that place at time $t = 0$. Now—coming back to the topic of the contamination. As mentioned at the beginning of this section, for each piece of area (represented by a single pixel) it should be dependent on the number of zombies that this particular area was in contact with at a given moment of time. Hence, we proposed the following way of calculating contamination

$$c_t^{(i,j)} = \max\left\{1, \alpha\frac{\sum_{(p,q)\in\mathcal{N}(i,j)} Z_{t-1}^{(p,q)}}{|\mathcal{N}(i,j)|}\right\}. \tag{6.6}$$

In formula (6.6) $\mathcal{N}(i, j)$ represents the neighbourhood of (i, j), i.e. the set of positions which we consider to affect (i, j) in terms of the possibility of the zombie attack. For example, the classical neighbourhood (often referred to as von Neumann neighbourhood) is the set of positions adjacent from top, bottom, left and right to the position in question. In such a set-up, we have for example $\mathcal{N}(2, 2) = \{(1, 2), (3, 2), (2, 1), (2, 3), (2, 2)\}$. Pixel's self-inclusion in its neighbourhood scheme ensures that if a given point already has some zombies, there will be more of them in the next time step, even if this spot does not yet have any contaminated neighbours. Since we are dealing with the data normalised to fit between 0 and 1, at any point of time the cumulative value of the zombies' component within a given neighbourhood constructed of k positions cannot exceed k. Hence, the value of the fraction $\frac{\sum_{(p,q)\in\mathcal{N}(i,j)} Z_{t-1}^{(p,q)}}{|\mathcal{N}(i,j)|}$ is between 0 and 1 so the value of this fraction itself could serve as a decent definition of contamination. However, to give ourselves a bit more modelling freedom we decided to introduce one more parameter, α, to artificially rise or decrease the contamination value. We called it *infectiousness* and explain it as a measure of invasion's strength in terms of spreading. Setting α to be a big number makes the invasion develop faster. It must be noted that multiplying by α, especially if its value is much bigger than 1, might make the contamination ratio rise above the value of 1, which is unacceptable (as the value of $H_t^{(i,j)}$ could drop below zero, as per Eq. (6.2)). Therefore, we added a safety feature in the form of a max function.

We ran the simple model described above for a map of the New York City and surrounding areas, we chose Central Park as the zombie's outbreak. We took

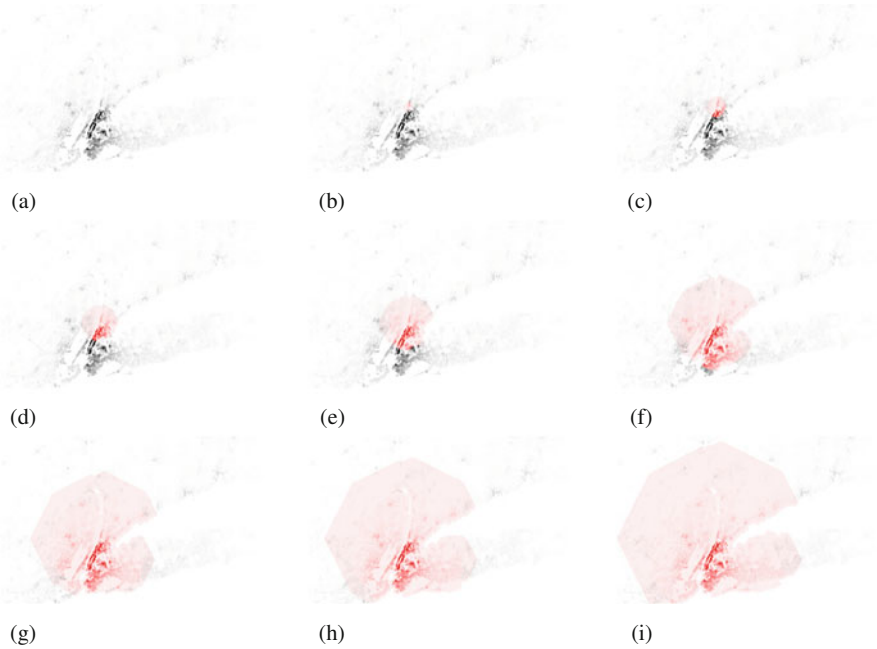

Fig. 6.5 Snapshots of the process of the zombie invasion, as per the rules of the model described in Sect. 6.3.1. (**a**) Zombie spread at $t = 0$. (**b**) Zombie spread at $t = 3$. (**c**) Zombie spread at $t = 6$. (**d**) Zombie spread at $t = 10$. (**e**) Zombie spread at $t = 15$. (**f**) Zombie spread at $t = 25$. (**g**) Zombie spread at $t = 35$. (**h**) Zombie spread at $t = 42$. (**i**) Zombie spread at $t = 50$

a snapshot of every stage of the time evolution as well as we plotted how the populations of healthy people and zombies were changing over time. These results were presented in Figs. 6.5 and 6.6 respectively.

6.3.2 Probabilistic Model (Without Recovery)

Up until now, the model was fully deterministic—every time a program was run for a given population density map as an input, the invasion flow was totally identical. However, we felt that a sudden and unexpected event—such as a zombie invasion—should probably be modelled in a way incorporating at least some random factors. Therefore, we created another model, which can be described in our evolution equations as follows

$$H_t^{(i,j)} = H_{t-1}^{(i,j)} - H_{t-1}^{(i,j)} \cdot c_t^{(i,j)} \cdot \mathrm{B}_t^{(i,j)}(f), \tag{6.7}$$

$$Z_t^{(i,j)} = Z_{t-1}^{(i,j)} + H_{t-1}^{(i,j)} \cdot c_t^{(i,j)} \cdot \mathrm{B}_t^{(i,j)}(f). \tag{6.8}$$

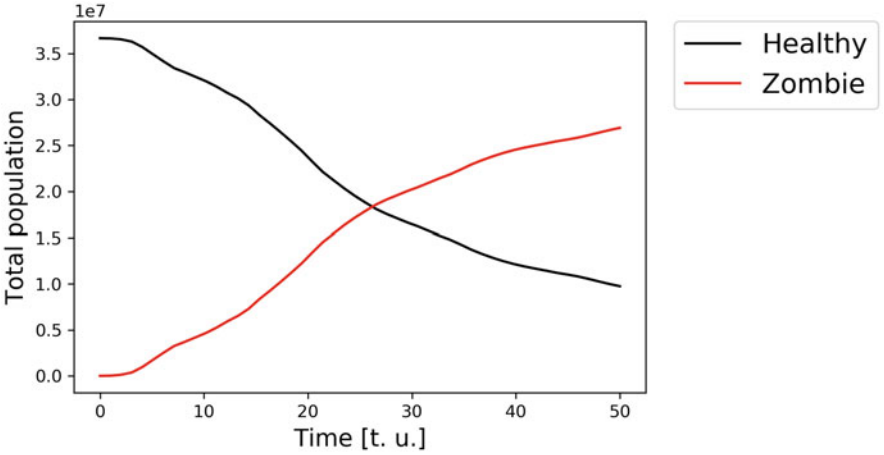

Fig. 6.6 Populations of healthy people and zombies as functions of time, within the model described in Sect. 6.3.1

We now included $\mathrm{B}_t^{(i,j)}(f)$ which is a random variable from a Bernoulli distribution with parameter f, $f \in [0, 1]$:

$$\mathrm{B}_t^{(i,j)}(f) = \begin{cases} 0 & \text{with probability } f, \\ 1 & \text{with probability } (1-f) \end{cases} \tag{6.9}$$

and (6.10)

$$\mathrm{B}_t^{(i,j)}(f) \text{ is independent of } \mathrm{B}_s^{(p,q)}(f) \text{ for any } i \neq p,\ j \neq q,\ t \neq s. \tag{6.11}$$

We called the f parameter *resistance*. Let us note that for $f = 0$, the new model reduces to Model 1, described by Eqs. (6.2) and (6.3)—in such model, whenever a given point on the map has a neighbour attacked by the zombies, it certainly becomes infected too. Note however that for $f = 1$ the system reduces again, this time to the form of

$$H_t^{(i,j)} = H_{t-1}^{(i,j)}, \tag{6.12}$$

$$Z_t^{(i,j)} = Z_{t-1}^{(i,j)}. \tag{6.13}$$

We therefore see no changes at all—we can say, healthy people are 100% effective in *resisting* zombies' spread. We can therefore think of f as of a slider which lets us decide on what is the actual probability that a given point on the map, in certain moment of the simulation, turns out to defend itself from the zombie attack. Thanks to this we successfully introduced an intuitive randomization to the flow of the model execution.

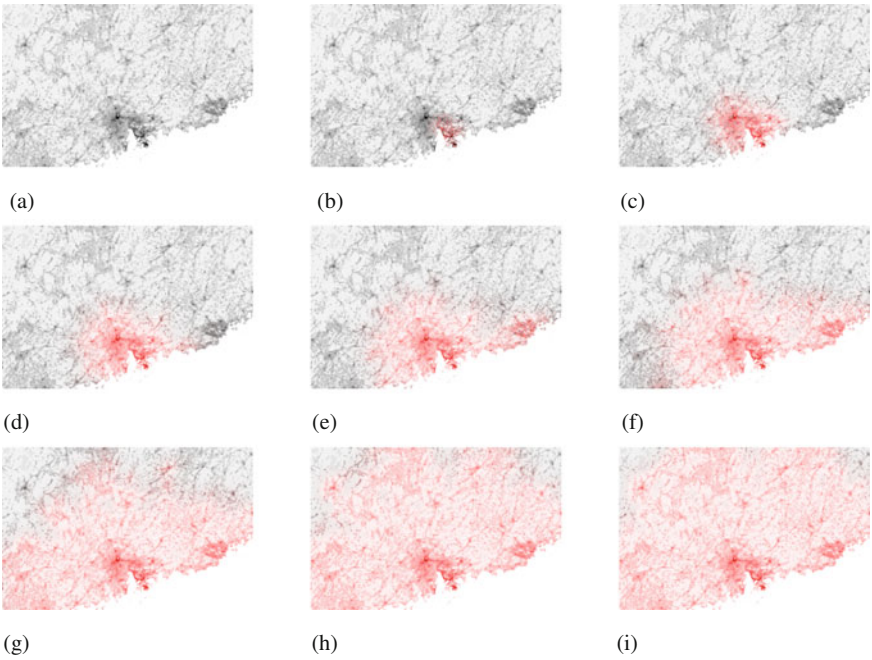

Fig. 6.7 Snapshots of the process of the zombie invasion, as per the rules of the model described in Sect. 6.3.2. (**a**) Zombie spread at $t = 0$. (**b**) Zombie spread at $t = 5$. (**c**) Zombie spread at $t = 15$. (**d**) Zombie spread at $t = 25$. (**e**) Zombie spread at $t = 35$. (**f**) Zombie spread at $t = 45$. (**g**) Zombie spread at $t = 65$. (**h**) Zombie spread at $t = 85$. (**i**) Zombie spread at $t = 100$

As done previously, also this time we checked performance of our model on the actual density population map. We used the map of southern coast of China, which is known to be very densely populated. The invasion started in the centre of Hong Kong and it quickly spread to the surrounding big cities—Shenzen, Dongguan, Guangzhou and Foshan.[1] The invasion slowed down when it came to the rural areas of the Guangdong province. The pace of invasion increased again when it reached another densely populated areas, near the city of Shantou and after overrunning this neighbourhood—the invasion went a bit slower again. Snapshots of the entire process and the plot of the populations can be seen in Figs. 6.7 and 6.8 respectively.

[1]We considered zombies not to be bound by the geopolitical difficulties of travelling between Hong Kong and continental China overland, which the 'ordinary' people are subject to.

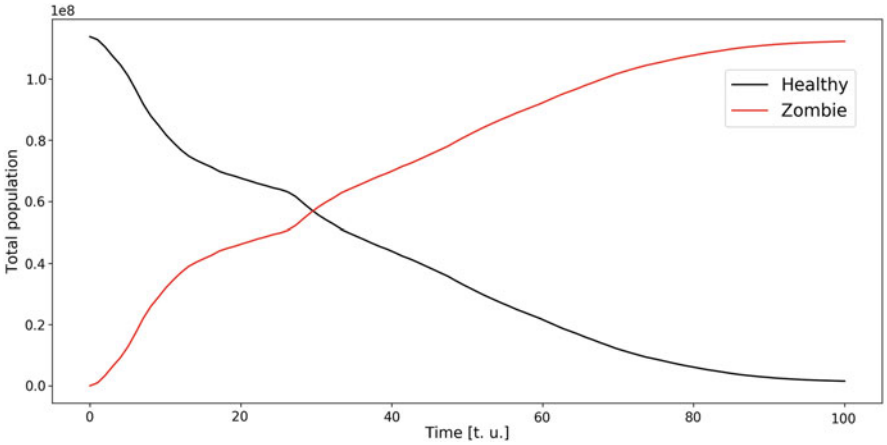

Fig. 6.8 Populations of healthy people and zombies as functions of time, within the model described in Sect. 6.3.2

6.3.3 Probabilistic Model (with Recovery)

For our last and final model we decided to allow for the possibility of the zombies to be recovered from their illness. The recovery process was based on the idea of research centres—big medical facilities which turn out to be able to develop a cure for what makes people zombies, but only after some time from the beginning of the invasion passes. This obviously requires us to add the third class of people that we consider in our model —- the recovered ones. The evolution equations took the following form

$$H_t^{(i,j)} = H_{t-1}^{(i,j)} - H_{t-1}^{(i,j)} \cdot c_t^{(i,j)} \cdot \mathrm{B}_t^{(i,j)}(f), \tag{6.14}$$

$$Z_t^{(i,j)} = \left(Z_{t-1}^{(i,j)} + H_{t-1}^{(i,j)} \cdot c_t^{(i,j)} \cdot \mathrm{B}_t^{(i,j)}(f)\right) \cdot \left(1 - \mathrm{B}_t^{(i,j)}(s) \cdot r_t^{(i,j)}\right), \tag{6.15}$$

$$R_t^{(i,j)} = R_{t-1}^{(i,j)} + \left(Z_{t-1}^{(i,j)} + H_{t-1}^{(i,j)} \cdot c_t^{(i,j)} \cdot \mathrm{B}_t^{(i,j)}(f)\right) \cdot \mathrm{B}_t^{(i,j)}(s) \cdot r_t^{(i,j)}. \tag{6.16}$$

Since we assume that some time must pass until first recovered people appear, we may safely assume that

$$R_0^{(i,j)} = 0 \quad \text{for all available } (i, j).$$

As we can see—two new parameters appeared in Eqs. (6.15) and (6.16), compared to the Eqs. (6.7) and (6.8), describing the previous model. Those parameters are $\mathrm{B}_t^{(i,j)}(s)$ and $r_t^{(i,j)}$. $\mathrm{B}_t^{(i,j)}(s)$ is again, a random variable from Bernoulli distribution with success parameter $s \in [0, 1]$ which we called the *development*

factor. It plays a similar role to the role of $\mathrm{B}_t^{(i,j)}(f)$ in the previous model—introduces randomness. We thought that the more developed a given country is, the more likely it will be that each and every cell will start curing itself if it only had a chance to do so (this will likely be a case for a big value of the development factor s) but if the level of the development of the society is low—there might be problems with transferring cure from one neighbourhood to the other (such situation would be modelled by a small value of s).

The actual possibility of a given cell to be cured is however described by $\mathrm{r}_t^{(i,j)}$—the *recovery* factor, i.e. the fraction of zombies who become recovered (note the similarity to the contamination parameter, first introduced in Eqs. (6.2) and (6.3)). Its precise definition is

$$r_t^{(i,j)} = \begin{cases} 1 & \text{if } (i,j) \in D \text{ and } t \geqslant t_d, \\ \beta \cdot \mathrm{I}_t^{(i,j)} & \text{otherwise} \end{cases}. \tag{6.17}$$

Here, D is a set of coordinates of, what we call, the research centres. These are the places on the map where we want the cure for being a zombie to be developed. As mentioned in the first paragraph of this section, we assume that the research centres need time to figure out the cure. This time is denoted by t_d. We can therefore see that points which we specify as the research centres will be cured precisely at time t_d. What about all the other points, outside of the ones defined to be research centres? Of them, the recovery indicator $\mathrm{I}_t^{(i,j)}$ takes care. It is defined as follows

$$\mathrm{I}_t^{(i,j)} = \max\left\{1, \sum_{(p,q)\in\mathcal{N}(i,j)} R_{t-1}^{(p,q)}\right\} \tag{6.18}$$

The idea of the indicator is quite simple—for the cell (i, j) at the moment t, the value of the indicator is 1 if there are any cells in the neighbourhood which contain any recovered people. This information is then used to establish the fraction of people that will be cured from being zombies—to do that we multiply $\mathrm{I}_t^{(i,j)}$ by an additional parameter $\beta \in (0, 1]$, which we call the *purification* constant. This constant helps us to control the speed in which zombies will be converted into recovered humans once they have access to cure through their neighbours.

As this was our final and most complex model, we naturally wanted to visualise it as well. Since we arrived at having three classes of individuals, we decided that one colour map is insufficient for this purpose, so we used two instead and we placed them side by side. One of the maps represents the number of zombies and the other—the number of recovered humans. For the final visualisation we chose the map of France. We placed the research centres in three major French cities—Paris, Toulouse and Bordeaux. The zombie outbreak was placed in a relatively minor town in the south-east of the country—Grenoble. As usually, we generated the full animation of the invasion spreading and the plot which presents the cumulative amounts in time—see Figs. 6.9 and 6.10 respectively. The snapshots in Fig. 6.9 are

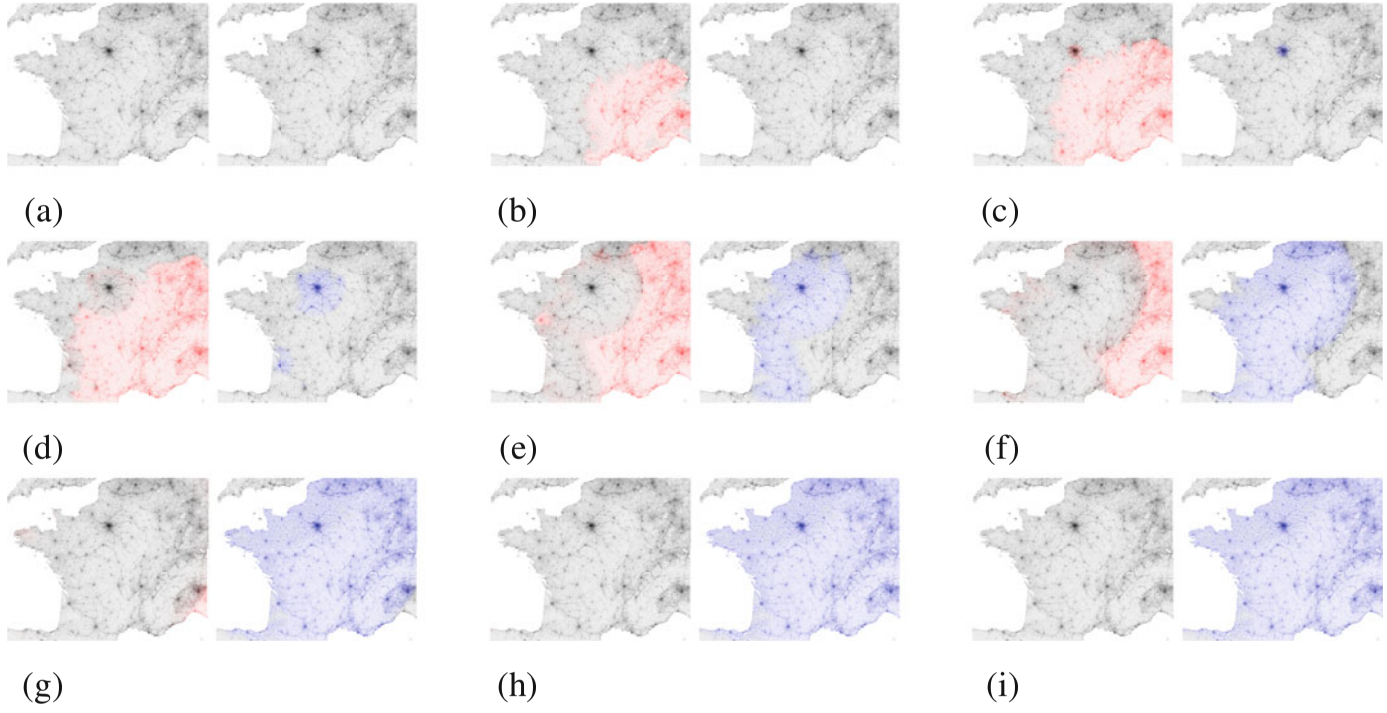

Fig. 6.9 Snapshots of the process of the zombie invasion, as per the rules of the model described in Sect. 6.3.3. (**a**) Zombie spread at $t = 0$. (**b**) Zombie spread at $t = 40$. (**c**) Zombie spread at $t = 50$. (**d**) Zombie spread at $t = 60$. (**e**) Zombie spread at $t = 70$. (**f**) Zombie spread at $t = 80$. (**g**) Zombie spread at $t = 100$. (**h**) Zombie spread at $t = 120$. (**i**) Zombie spread at $t = 150$

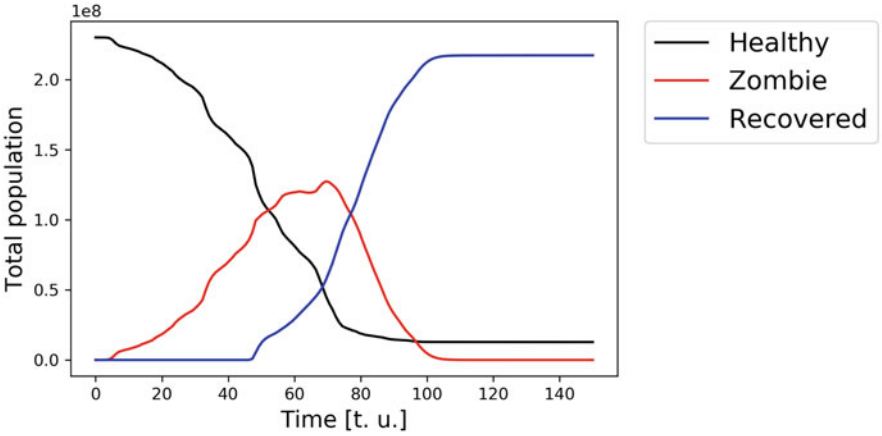

Fig. 6.10 Populations of healthy people and zombies as functions of time, within the model described in Sect. 6.3.3

particularly informative—we can clearly see from them that around $t = 50$ first recovered people appear around the three research centres and they start spreading quickly so that after $t = 100$ zombies are almost completely eradicated form the system. The plot in Fig. 6.10 shows us however that there are still some healthy people there, which means they have never become infected—and they will never be, as the entire epidemic has been defeated.

6.4 Conclusions

The project described above aimed to develop a model of a disorder rapidly spreading over some area for which the density population maps are available. As part of our work on the project we created as many as three such models which vary in the level of complexity. While creating the models we were not focusing on any concrete information about the actual epidemic that we were modelling, we also tried not to put any binding assumptions. Instead, we were turning nearly every variable that we used into a customisable parameter of the model. Thanks to that we were able to obtain the models which are very general and also possible to be calibrated to various kinds of real diseases, as long as they are able to spread relatively quickly. The next steps that can be taken to develop the project is to try to estimate the values of all the parameters so that the models themselves can be used for forecasting an modelling the spread of real-life contagious diseases.

The number of complex disease spreading models that are already available and well-studied is obviously huge. One of the biggest advantages of each of our models is their actual simplicity. Although the evolution equations that we used to describe the dynamics of populations of healthy and infected people might look a bit

overwhelming, they mostly consist of simple arithmetic operations and feature some random factors—both of these mathematical ideas can be easily understood by most people. Models based on complicated differential equations or stochastic processes might have a lot of advantages that our models do not have but they also might be more difficult to explain to a person which does not have a solid mathematical background. In our models however, changes can be made easily and the effects of those changes are usually quite easy to predict, they can also be verified quickly by the simulation that we have performed ourselves as well. Simple tools are usually more difficult to be broken, yet, if used appropriately, they may give us valuable insights into the problems that we are studying.

Acknowledgments Author's inspiration for considering this research problem was a post on the blog by Max Breggren [1].

The solution to the problem of modelling zombie invasion in the shape described above was obtained during the ECMI Modelling Week 2019 in Grenoble, France, under the supervision of the author of this work. It would not be possible to achieve those results without the group of remarkably hard-working students from all over Europe. The groups members were:

- Alejandro Tobio Pena,
- Alessandro Sfilio,
- Chiara Borsani,
- Miguel Rebocho,
- Sara Costa Faya,
- Simon Li Ying Yin.

The author would like to thank all of them for their cooperativeness and involvement in the project.

References

1. Max Berggren. Model of a zombie outbreak in Sweden, Norway and Finland (Denmark is fine). Website. last checked: 20.10.2019.
2. Alex Clark. Pillow. Website. last checked: 20.10.2019.
3. Herbert W. Hethcote. *Three Basic Epidemiological Models*, pages 119–144. Springer Berlin Heidelberg, Berlin, Heidelberg, 1989.
4. John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, May 2007.
5. Travis E. Oliphant. *Guide to NumPy*. CreateSpace Independent Publishing Platform, USA, 2nd edition, 2015.
6. Duncan Smith. World population density. Website. last checked: 20.10.2019.
7. Guido van Rossum. Python language reference, version 3.6. Website. last checked: 20.10.2019.

Chapter 7
Optimal Heating of an Indoor Swimming Pool

Monika Wolfmayr

7.1 Introduction

Modeling the heating of an object is an important task in many applicational problems. Moreover, a matter of particular interest is to find the optimal heating of an object such that it has a desired temperature distribution after some given time. In order to formulate such optimal control problems and to solve them, a cost functional subject to a time-dependent partial differential equation (PDE) is derived. One of the profound works paving the way for PDE-constrained optimization's relevance in research and application during the last couple of decades is Lion's work [9] from 1971. Some recent published monographs discussing PDE-constrained optimization as well as various efficient computational methods for solving them are, e.g., [1, 6], and [11], where the latter one is used as basis for the discussion on solving the optimal heating problem of this work.

The goal of this work is to derive a simple mathematical model for finding the optimal heating of the air in a glass dome represented by a half sphere, where a swimming pool is located in the bottom of the dome and the heat sources (or heaters) are situated on a part of the boundary of the glass dome. The process from the model to the final numerical simulations usually involves several steps. The main steps in this work are the setting up of the mathematical model for the physical problem, obtaining some analytical results of the problem, presenting a proper discretization for the continuous problem and finally computing the numerical solution of the problem. The parabolic optimal control problem is discretized by the finite element method in space, and in time, we use the implicit Euler method for performing the time stepping. The used solution algorithm for the discretized problem is the

M. Wolfmayr (✉)
Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
e-mail: monika.k.wolfmayr@jyu.fi

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_7

projected gradient method, which is for instance applied in [5] as well as in more detail discussed in [4, 6, 8, 10].

We want to emphasize that the model and the presented optimization methods for the heating process of this work are only one example for a possible modeling and solution. In fact, the stated model problem has potential for many modeling tasks for students and researchers. For instance, different material parameters for the dome as well as for air and water could be studied more carefully. The optimal modeling of the heat sources could be stated as a shape optimization problem or instead of optimizing the temperature of the air in the glass dome, one could optimize the water temperature. This would correspond to a final desired temperature distribution corresponding to the boundary of the glass dome, where the swimming pool is located, for the optimal control problem. Another task for the students could be to compute many simulations with, e.g., Matlab's `pdeModeler` to derive a better understanding of the problem in the pre-phase of studying the problem of this work. However, we only want to mention here a few other possibilities for modeling, studying and solving an optimal heating problem amongst many other tasks, and we are not focusing on them in the work presented here.

This article is organized as follows: First, the model of the heating process is formulated in Sect. 7.2. Next, Sect. 7.3 introduces the optimal control problem, which describes the optimal heating of the glass dome such that the desired temperature distribution is attained after a given time. In Sect. 7.4, proper function spaces are presented in order to discuss existence and uniqueness of the optimal control problem. We derive the reduced optimization problem in Sect. 7.5 before discussing its discretization and the numerical method for solving it, the projected gradient method, in more detail in Sect. 7.6. Numerical results are presented as well as conclusions are drawn in final Sect. 7.7.

7.2 Modeling

This section presents the modeling process. The physical problem is described in terms of mathematical language, which includes formulating an initial version of the problem, but then simplifying it in order to derive a version of the problem which is easier to solve. However, at the same time, the problem has to be kept accurate enough in order to compute an approximate solution being close enough to the original solution. That is exactly one of the major goals of mathematical modeling. In the following, we introduce the domain describing the glass dome, where an indoor swimming pool is located in the bottom of the dome, and the position of the heaters. The concrete equations describing the process of heating and the cost functional subject to them modeling the optimization task are discussed in next Sect. 7.3.

We have an indoor swimming pool which is located under a glass dome. For simplicity, it is assumed that we have an isolated system in the glass dome, so no heat can leak from the domain. The swimming pool covers the floor of the glass

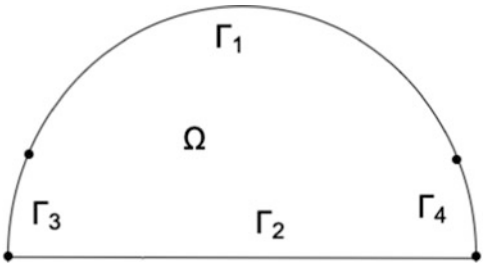

Fig. 7.1 The domain Ω reduced from 3d to 2d due to symmetry properties describing the glass dome and its heaters placed at the ground of the glass dome next to the floor, hence subdividing the boundary $\Gamma = \partial\Omega$ into four parts Γ_1, Γ_2, Γ_3 and Γ_4

dome and we assume that the heaters are placed next to the floor up on the glass all around the dome. The target of the minimization functional is to reach a desired temperature distribution at the end of a given time interval $(0, T)$, where $T > 0$ denotes the final time, with the least possible cost.

Due to the symmetry properties of the geometry as well as the uniform distribution of the water temperature, we reduce the three dimensional (3d) problem to a two dimensional (2d) one. The dimension reduction makes the numerical computations more simple. In the following, the 2d domain is denoted by Ω and its boundary by $\Gamma = \partial\Omega$. We assume that $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain. We subdivide its boundary Γ into four parts: the glass part Γ_1, the floor which is the swimming pool Γ_2 and the heaters Γ_3 and Γ_4. The domain Ω and its boundaries are illustrated in Fig. 7.1.

7.3 Optimal Control Problem

In this section, the optimal control problem is formulated, where an optimal control function u has to be obtained corresponding to the heating of the heat sources on the boundary $\Gamma_R := \Gamma_3 \cup \Gamma_4$ such that the state y reaches a desired temperature distribution y_d after a given time T. This problem can be formulated in terms of a PDE-constrained optimization problem, which means minimizing a cost functional subject to a PDE and with u being the control function.

Let $Q_T := \Omega \times (0, T)$ denote the space-time cylinder with the lateral surface $\Sigma := \Gamma \times (0, T)$, where $T > 0$ denotes the final time. The optimal control problem is given as follows:

$$\min_{(y,u)} J(y, u) = \frac{1}{2} \int_\Omega (y(\boldsymbol{x}, T) - y_d(\boldsymbol{x}))^2 \, d\boldsymbol{x} + \frac{\lambda}{2} \int_0^T \int_{\Gamma_R} u(x, t)^2 \, ds \, dt \tag{7.1}$$

such that

$$y_t - \triangle y = 0 \qquad \text{in } Q_T := \Omega \times (0, T), \tag{7.2}$$

$$\frac{\partial y}{\partial n} = 0 \qquad \text{on } \Sigma_1 := \Gamma_1 \times (0, T), \tag{7.3}$$

$$y = g \qquad \text{on } \Sigma_2 := \Gamma_2 \times (0, T), \tag{7.4}$$

$$\frac{\partial y}{\partial n} + \alpha y = \beta u \qquad \text{on } \Sigma_R := \Gamma_R \times (0, T), \tag{7.5}$$

$$y(0) = y_0 \qquad \text{in } \Omega, \tag{7.6}$$

where g is the given constant water temperature, $\lambda \geq 0$ is the cost coefficient or control parameter, and α and β are constants describing the heat transfer, which are modeling parameters and have to be chosen carefully. The control u denotes the radiator heating, which has to be chosen within a certain temperature range. Hence, we choose the control functions from the following set of admissible controls:

$$u \in U_{\text{ad}} = \{v \in L^2(\Sigma_R) : u_a(\boldsymbol{x}, t) \leq u(\boldsymbol{x}, t) \leq u_b(\boldsymbol{x}, t) \text{ a.e. on } \Sigma_R\}, \tag{7.7}$$

which means that u has to fulfill so called box constraints. Equations (7.3)–(7.5) are called Neumann, Dirichlet and Robin boundary conditions, respectively. Equation (7.3) characterizes a no-flux condition in normal direction. Equation (7.4) describes a constant temperature distribution. Regarding Eq. (7.5), $\alpha = \beta$ would be a reasonable choice from the physical point of view because this would mean that the temperature increase at this part of the boundary is proportional to the difference between the temperature there and outside. However, a decoupling of the parameters makes sense too, see [11], and does not change anything for the actual discussion and computations, since $\alpha = \beta$ could be chosen at any point.

The goal is to find the optimal set of state and control (y, u) such that the cost is minimal.

7.4 Existence and Uniqueness

In this section, we discuss some basic results on the existence and uniqueness of the parabolic initial-boundary value problem (7.2)–(7.6), whereas we exclude the details. They can be found in [11]. We first introduce proper function spaces leading to a setting, where existence and uniqueness of the solution can be proved.

Definition 7.1 The normed space $W_2^{1,0}(Q_T)$ is defined as follows

$$W_2^{1,0}(Q_T) = \{y \in L^2(Q_T) : D_i y \in L^2(Q_T) \, \forall i = 1, \ldots, d\} \tag{7.8}$$

with the norm

$$\|y\|_{W_2^{1,0}(Q_T)} = \left(\int_0^T \int_\Omega (|y(\boldsymbol{x}, t)|^2 + |\nabla y(\boldsymbol{x}, t)|^2) \, d\boldsymbol{x} \, dt \right)^{1/2}, \tag{7.9}$$

where $D_i y$ denotes the spatial derivative of y in i-direction and d is the spatial dimension.

For the model problem of this work, the dimension is $d = 2$. In the following, let $\{V, \|\cdot\|_V\}$ be a real Banach space. More precisely, we will consider $V = H^1(\Omega)$ in this work.

Definition 7.2 The space $L^p(0, T; V)$, $1 \leq p < \infty$, denotes the linear space of all equivalence classes of measurable vector valued functions $y : [0, T] \to V$ such that

$$\int_0^T \|y(t)\|_V^p \, dt < \infty. \tag{7.10}$$

The space $L^p(0, T; V)$ is a Banach space with respect to the norm

$$\|y\|_{L^p(0,T;V)} := \left(\int_0^T \|y(t)\|_V^p \, dt \right)^{1/p}. \tag{7.11}$$

Definition 7.3 The space $W(0, T) = \{y \in L^2(0, T; V) : y' \in L^2(0, T; V^*)\}$ is equipped with the norm

$$\|y\|_{W(0,T)} = \left(\int_0^T (|y(t)|_V^2 + |y'(t)|_{V^*}^2) \, dt \right)^{1/2}. \tag{7.12}$$

It is a Hilbert space with the scalar product

$$(u, w)_{W(0,T)} = \int_0^T (u(t), w(t))_V \, dt + \int_0^T (u'(t), w'(t))_{V^*} \, dt. \tag{7.13}$$

The relation $V \subset H = H^* \subset V^*$ is called a Gelfand or evolution triple and describes a chain of dense and continuous embeddings.

The problem (7.2)–(7.6) has a unique weak solution $y \in W_2^{1,0}(Q_T)$ for a given $u \in U_{\text{ad}}$. Moreover, the solution depends continuously on the data, which means that there exists a constant $c > 0$ being independent of u, g and y_0 such that

$$\max_{t \in [0,T]} \|y(\cdot, t)\|_{L^2(\Omega)} + \|y\|_{W_2^{1,0}(Q_T)} \leq c(\|u\|_{L^2(\Sigma_R)} + \|g\|_{L^2(\Sigma_2)} + \|y_0\|_{L^2(\Omega)}) \tag{7.14}$$

for all $u \in L^2(\Sigma_R)$, $g \in L^2(\Sigma_2)$ and $y_0 \in L^2(\Omega)$. Hence, problem (7.2)–(7.6) is well-posed in $W_2^{1,0}(Q_T)$. Furthermore, since $y \in W_2^{1,0}(Q_T)$ and it is a weak solution of problem (7.2)–(7.6), y also belongs to $W(0, T)$. The following estimate holds:

$$\|y\|_{W(0,T)} \leq \tilde{c}(\|u\|_{L^2(\Sigma_R)} + \|g\|_{L^2(\Sigma_2)} + \|y_0\|_{L^2(\Omega)}) \tag{7.15}$$

for some constant $\tilde{c} > 0$ being independent of u, g and y_0. Hence, problem (7.2)–(7.6) is also well-posed in the space $W(0,T)$.

Note that the Neumann boundary conditions on Γ_1 are included in both estimates (7.14) and (7.15) related to the well-posedness of the problem (as discussed in [11]). However, the Neumann boundary conditions (7.3) are equal to zero.

Under the assumptions that $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain with boundary Γ, $\lambda \geq 0$ is a fixed constant, $y_d \in L^2(Q_T)$, $\alpha, \beta \in L^\infty(\Sigma_R)$, and $u_a, u_b \in L^2(\Sigma_R)$ with $u_a \leq u_b$ a.e. on Σ_R, together with the existence and uniqueness result on the parabolic initial-boundary value problem (7.2)–(7.6) in $W(0,T)$, the optimal control problem (7.1)–(7.7) has at least one optimal control $\bar{u} \in U_{\text{ad}}$. In case of $\lambda > 0$ the optimal control $\bar{u}$ is uniquely determined.

7.5 Reduced Optimization Problem

In order to solve the optimal control problem (7.1)–(7.7), we derive the so called reduced optimization problem first.

Since the problem is well-posed as discussed in the previous section, we can formally eliminate the state equation (7.2)–(7.6) and the minimization problem reads as follows

$$\min_u \bar{J}(u) = \frac{1}{2}\int_\Omega (y_u(\boldsymbol{x},T) - y_d(\boldsymbol{x}))^2\, d\boldsymbol{x} + \frac{\lambda}{2}\int_0^T \int_{\Gamma_R} u(x,t)^2\, ds\, dt \tag{7.16}$$

such that (7.7) is satisfied. The problem (7.16) is called reduced optimization problem. Formally, the function y_u denotes that the state function is depending on u. However, for simplicity we can set again $y_u = y$. In order to solve the problem (7.16), we apply the projected gradient method. The gradient of $\bar{J}$ has to be calculated by deriving the adjoint problem which is given by

$$-p_t - \triangle p = 0 \qquad \text{in } Q_T, \tag{7.17}$$

$$\frac{\partial p}{\partial n} = 0 \qquad \text{on } \Sigma_1, \tag{7.18}$$

$$p = 0 \qquad \text{on } \Sigma_2, \tag{7.19}$$

$$\frac{\partial p}{\partial n} + \alpha p = 0 \qquad \text{on } \Sigma_R, \tag{7.20}$$

$$p(T) = y(T) - y_d \qquad \text{in } \Omega. \tag{7.21}$$

The gradient of $\bar{J}$ is given by

$$\nabla \bar{J}(u(\boldsymbol{x},t)) = \beta \chi_{\Gamma_R} p(\boldsymbol{x}, T-t) + \lambda u(\boldsymbol{x},t) \tag{7.22}$$

with χ_{Γ_R} denoting the characteristic function on Γ_R. The projected gradient method can be now applied for computing the solution of the PDE-constrained optimization problem (7.1)–(7.7). We denote by

$$\mathscr{P}_{[u_a,u_b]}(u) = \max\{u_a, \min\{u_b, u\}\} \tag{7.23}$$

the projection onto the set of admissible controls U_{ad}.

Now putting everything together, the optimality system for (7.1)–(7.7) and a given $\lambda > 0$ reads as follows

$$\begin{aligned}
y_t - \triangle y = 0 \qquad & - p_t - \triangle p = 0 && \text{in } Q_T,\\
\frac{\partial y}{\partial n} = 0 \qquad & \frac{\partial p}{\partial n} = 0 && \text{on } \Sigma_1,\\
y = g \qquad & p = 0 && \text{on } \Sigma_2,\\
\frac{\partial y}{\partial n} + \alpha y = \beta u \qquad & \frac{\partial p}{\partial n} + \alpha p = 0 && \text{on } \Sigma_R,\\
y(0) = y_0 \qquad & p(T) = y(T) - y_d && \text{in } \Omega,\\
u = \mathscr{P}_{[u_a,u_b]}(-\frac{1}{\lambda}\beta p).
\end{aligned} \tag{7.24}$$

In case that $\lambda = 0$, the projection formula changes to

$$\begin{aligned}
u(\boldsymbol{x}, t) &= u_a(\boldsymbol{x}, t), \qquad \text{if } \beta(\boldsymbol{x}, t)p(\boldsymbol{x}, t) > 0,\\
u(\boldsymbol{x}, t) &= u_b(\boldsymbol{x}, t), \qquad \text{if } \beta(\boldsymbol{x}, t)p(\boldsymbol{x}, t) < 0.
\end{aligned} \tag{7.25}$$

Remark 7.1 In case that there are no control constraints imposed, the projection formula simplifies to $u = -\lambda^{-1}\beta p$.

7.6 Discretization and Numerical Method

In order to numerically solve the optimal control problem (7.1)–(7.7), which is equivalent to solving (7.24), we discretize the heat equation in space by the finite element method and in time. We use the implicit Euler method for performing the time stepping.

We approximate the functions y, u and p by finite element functions y_h, u_h and p_h from the conforming finite element space $V_h = \text{span}\{\varphi_1, \ldots, \varphi_n\}$ with the basis functions $\{\varphi_i(\boldsymbol{x}) : i = 1, 2, \ldots, n_h\}$, where h denotes the discretization parameter

with $n = n_h = \dim V_h = O(h^{-2})$. We use standard, continuous, piecewise linear finite elements and a regular triangulation $\mathscr{T}_h$ to construct the finite element space V_h. For more information, we refer the reader to [3] as well as to the newer publications [2, 7]. Discretizing problem (7.24) by computing its weak formulations and then inserting the finite element approximations for discretizing in space leads to the following discrete formulation:

$$M_h \underline{y}_{h,t} + K_h \underline{y}_h + \alpha M_h^{\Gamma_R} \underline{y}_h = \beta M_h^{\Gamma_R} \underline{u}_h, \qquad \underline{y}_h(0) = y_0, \tag{7.26}$$

$$-M_h \underline{p}_{h,t} + K_h \underline{p}_h + \alpha M_h^{\Gamma_R} \underline{p}_h = 0, \qquad \underline{p}_h(T) = \underline{y}_h(T) - y_d, \tag{7.27}$$

together with the projection formula

$$\underline{u}_h = \mathscr{P}_{[u_a,u_b]}(-\frac{1}{\lambda}\beta \underline{p}_h) \tag{7.28}$$

for $\lambda > 0$. The problem (7.26)–(7.28) has to be solved with respect to the nodal parameter vectors

$$\underline{y}_h = (y_{h,i})_{i=1,\dots,n}, \quad \underline{u}_h = (u_{h,i})_{i=1,\dots,n}, \quad \underline{p}_h = (p_{h,i})_{i=1,\dots,n} \in \mathbb{R}^n$$

of the finite element approximations $y_h(\boldsymbol{x}) = \sum_{i=1}^n y_{h,i}\varphi_i(\boldsymbol{x})$, $u_h(\boldsymbol{x}) = \sum_{i=1}^n u_{h,i}\varphi_i(\boldsymbol{x})$ and $p_h(\boldsymbol{x}) = \sum_{i=1}^n p_{h,i}\varphi_i(\boldsymbol{x})$. The values for $\underline{y}_h$ are set to g in the nodal values on the boundary Γ_2 and the problems are solved only for the degrees of freedom. The matrices M_h, $M_h^{\Gamma_R}$ and K_h denote the mass matrix, the mass matrix corresponding only to the Robin boundary Γ_R and the stiffness matrix, respectively. The entries of the mass and stiffness matrices are defined by the integrals

$$M_h^{ij} = \int_\Omega \varphi_i \varphi_j \, d\boldsymbol{x}, \qquad K_h^{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, d\boldsymbol{x}.$$

For the time stepping we use the implicit Euler method. After implementing the finite element discretization and the Euler scheme, we apply the following steps of the projected gradient method:

1. For $k = 0$, choose an initial guess u_h^0 satisfying the box constraints $u_a \le u_h^0 \le u_b$.
2. Solve the discrete forward problem (7.26) corresponding to (7.2)–(7.6) in order to compute y_h^k.
3. Solve the discrete backward problem (7.27) corresponding to (7.17)–(7.21) in order to obtain p_h^k.
4. Evaluate the descent direction of the discrete gradient

$$d^k = -\nabla \bar{J}(u_h^k) = -(\beta \chi_{\Gamma_R} p_h^k + \lambda u_h^k). \tag{7.29}$$

5. Set $u_h^{k+1} = \mathscr{P}_{[u_a,u_b]}(u_h^k + \gamma^k d^k)$ and go to step 2 unless stopping criteria are fulfilled.

Remark 7.2 For a first implementation, the step length $\gamma = \gamma^k$ can be chosen constant for all k. However, a better performance is achieved by applying a line search strategy as for instance the Armijo or Wolfe conditions to obtain the best possible γ^k in every iteration step k. We refer the reader to the methods discussed for instance in [5]. However, these strategies are not subject to the present work.

7.7 Numerical Results and Conclusions

In this section, we present numerical results for solving the type of model optimization problem discussed in this article and draw some conclusions in the end. The numerical experiments were computed in Matlab. The meshes were precomputed with Matlab's `pdeModeler`. The finite element approximation and time stepping as well as the projected gradient algorithm were implemented according to the discussions in the previous two sections.

In Fig. 7.2, the nodes corresponding to the interior nodes ('g.'), the Neumann boundary Γ_1 ('kd'), the Dirichlet boundary Γ_2 ('bs') and the Robin boundary Γ_R ('r*') are illustrated.

In the numerical experiments, we choose the following given data: the water temperature $g = 20$, the parameters $\alpha = \beta = 10^2$, the final time $T = 1$, the box constraints $u_a = 20$ and $u_b = 60$, the desired final temperature $y_d = 30$ and the initial value $y_0 = 0$ satisfying the boundary conditions. For the step lengths γ^k of the projected gradient algorithm, we choose the golden ratio $\gamma^k = \gamma = 1.618$ constant for all iteration steps k. The stopping criteria include that the norm of the errors $e_{k+1} := \|u_h^{k+1} - u_h^k\|/\|u_h^k\| < \epsilon_1$ or $|e_{k+1} - e_k| < \epsilon_2$ with $\epsilon_1 = 10^{-1}$ and $\epsilon_2 = 10^{-2}$ have to be fulfilled as well as setting a maximum number of iteration steps $k_{\max} = 20$ with $k < k_{\max}$.

In the first numerical experiment, we choose a fixed value for the cost coefficient $\lambda = 10^{-2}$ and compute the solution for different mesh sizes $n \in \{76, 275, 1045, 4073, 16081\}$. Table 7.1 presents the number of iterations needed until the stopping criteria were satisfied, for different mesh sizes and time steps. The number of time steps was chosen corresponding to the mesh size in order to guarantee that the CFL (Courant-Friedrichs-Lewy) condition is fulfilled.

In the set of Figs. 7.3, 7.4, 7.5, 7.6, 7.7, and 7.8, the approximate solutions y_h defined in Matlab as y are presented for the final time $t = T = 1$ computed on the different meshes including one figure, Fig. 7.6, presenting the adjoint state p for the mesh with 1045 nodes. We present only one figure for the adjoint state, since for other mesh sizes the plots looked similar.

In the second set of numerical experiments, we compute solutions for different cost coefficients λ on two different grids: one with mesh size $n = 275$ and 250 time steps, and another one with mesh size $n = 1045$ and 1000 time steps. The numerical

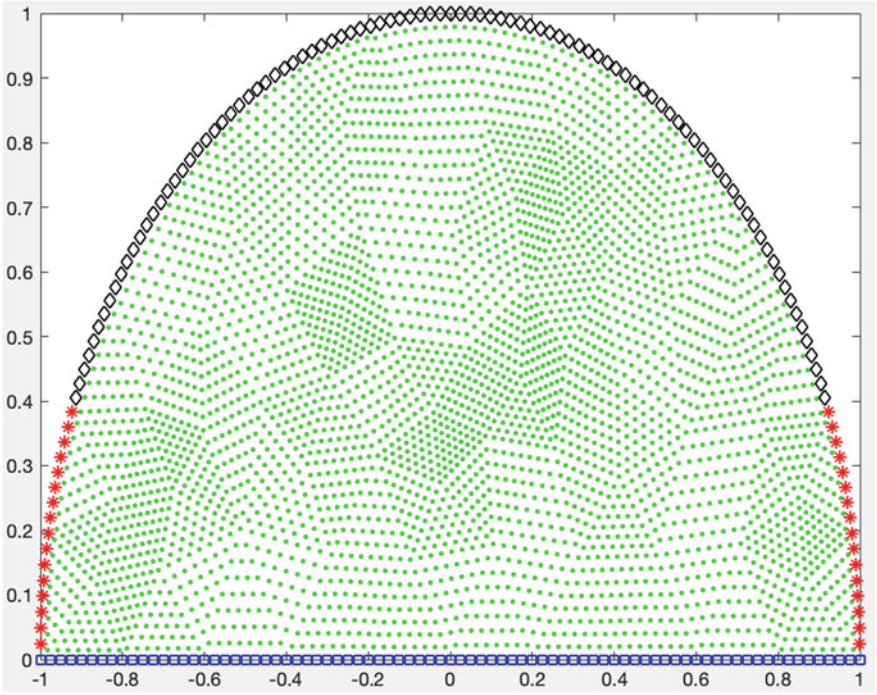

Fig. 7.2 The nodes marked corresponding to different boundaries and the interior of the domain Ω on a mesh with 4073 nodes

Table 7.1 Number of iterations needed to satisfy the stopping criteria for different mesh sizes and numbers of time steps for a fixed cost coefficient $\lambda = 10^{-2}$

Mesh size	Time steps	Iteration steps
76	125	7
275	250	5
1045	1000	19
4073	4000	19
16,081	16,000	4

results including the number of iteration steps needed are presented in Table 7.2. It can be observed that the numbers of iteration steps for the first case (275/250) are all very similar for different values of λ, even for the lower ones. In the second case (1045/1000), the numbers of iteration steps are getting higher the lower the values of λ. However, the results are satisfactory for these cases too. As example the approximate solution y for the final time $t = T = 1$ computed for $\lambda = 10^{-4}$ is presented in Fig. 7.9. The approximate solution for $\lambda = 10^{-2}$ has already been presented in Fig. 7.5.

The results of Tables 7.1 and 7.2 were included as example how one can perform different tables for different parameter values or combinations. Students or

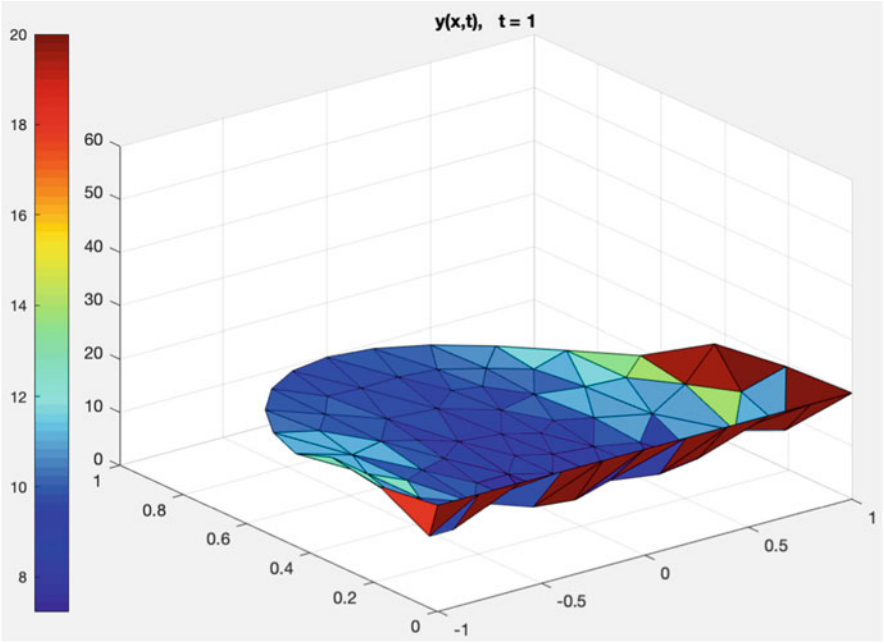

Fig. 7.3 The approximate solution y for final time $t = T = 1$ on a mesh with 76 nodes

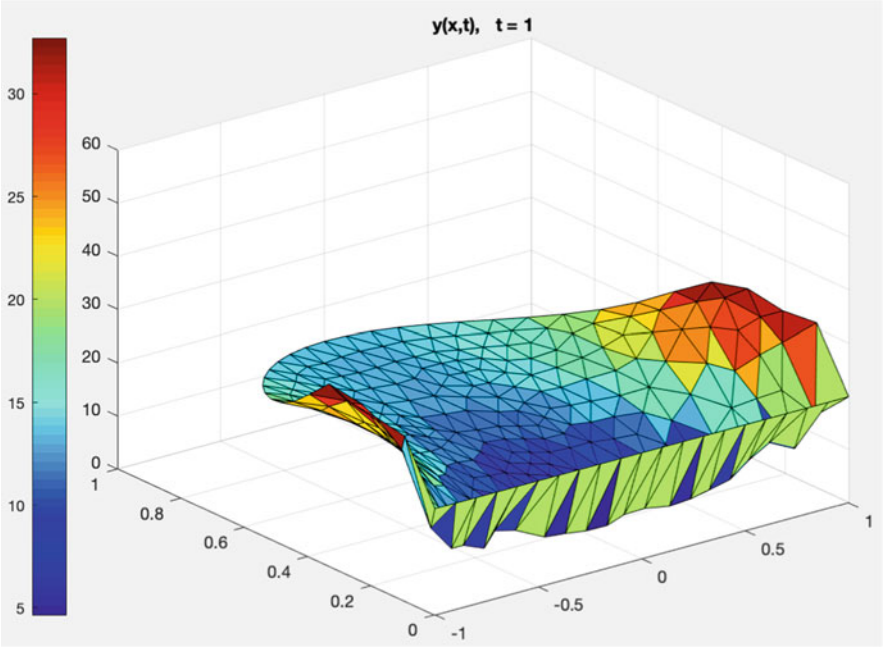

Fig. 7.4 The approximate solution y for final time $t = T = 1$ on a mesh with 275 nodes

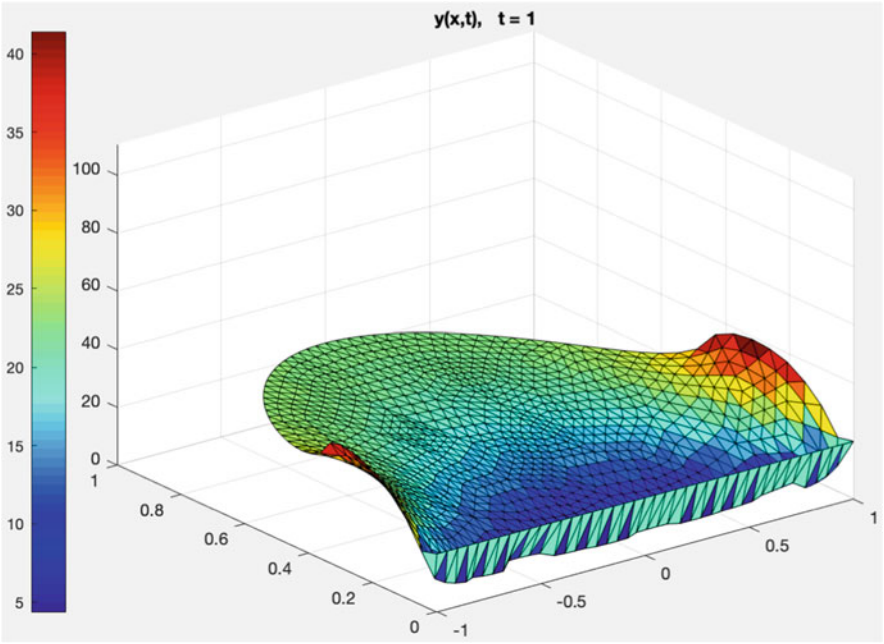

Fig. 7.5 The approximate solution y for final time $t = T = 1$ on a mesh with 1045 nodes

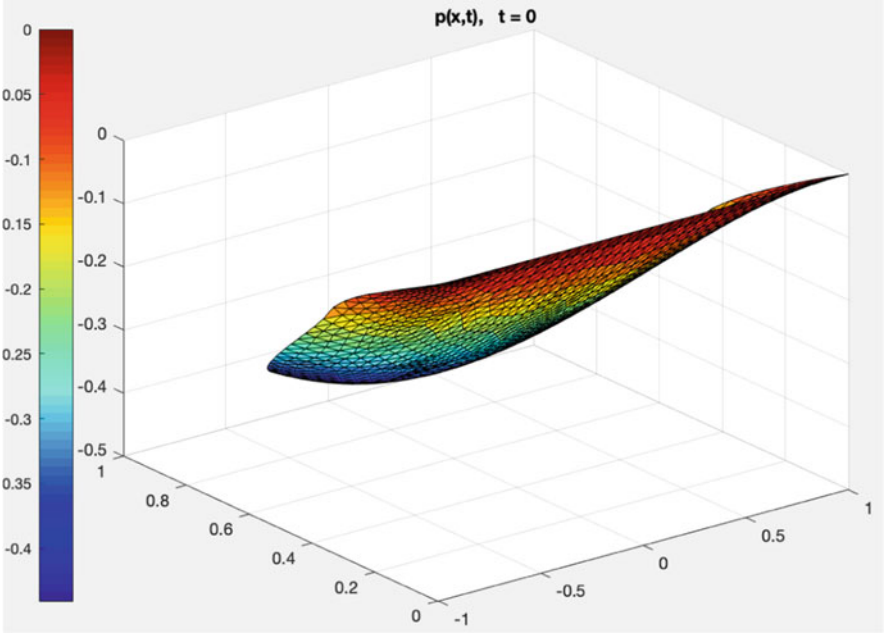

Fig. 7.6 The approximate adjoint state p for time $t = 0$ on a mesh with 1045 nodes

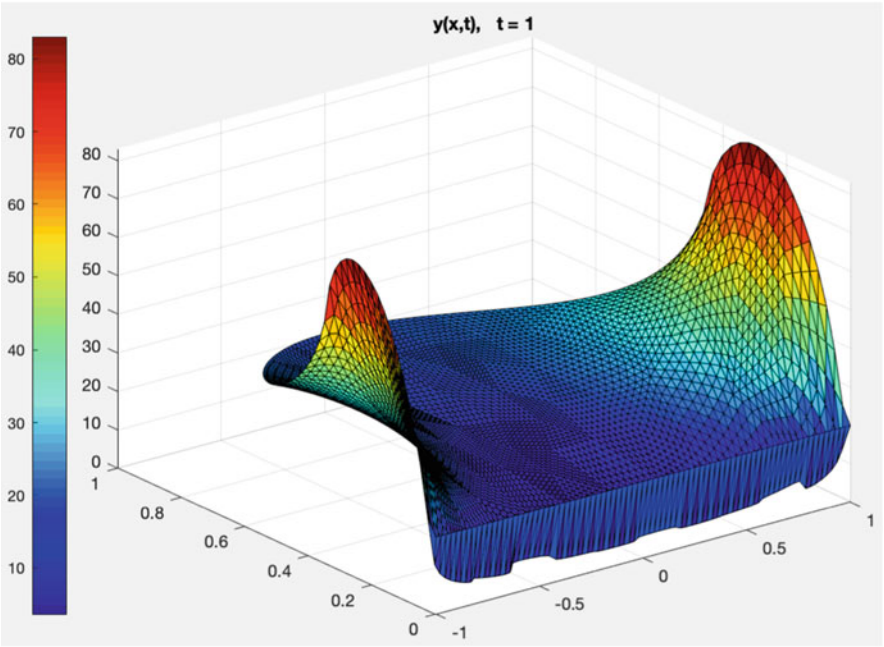

Fig. 7.7 The approximate solution y for final time $t = T = 1$ on a mesh with 4073 nodes

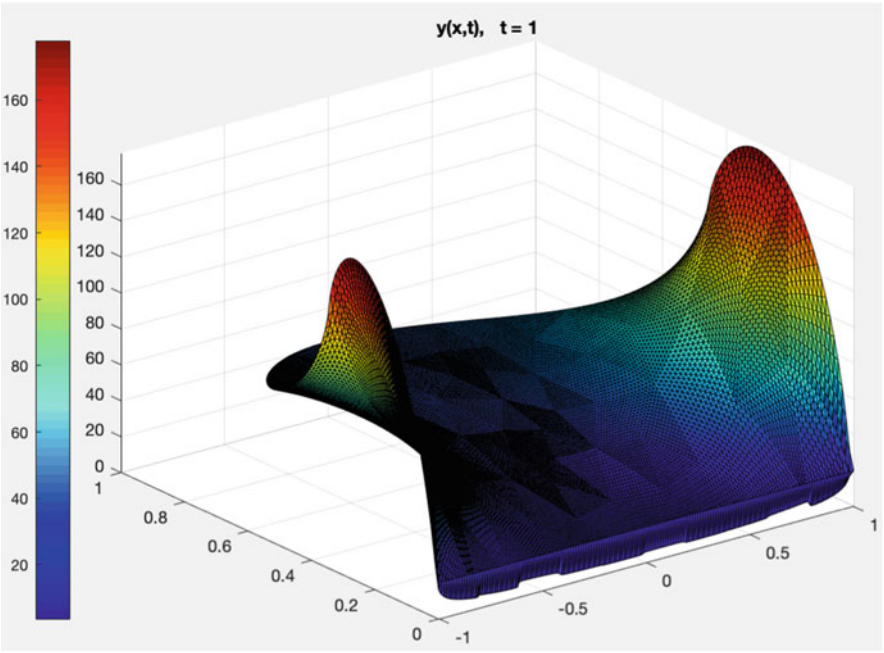

Fig. 7.8 The approximate solution y for final time $t = T = 1$ on a mesh with 16,081 nodes

Table 7.2 Number of iterations needed to satisfy the stopping criteria for different cost coefficients $\lambda \in \{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$ on grids with mesh sizes $n = 275$ and $n = 1045$ with 250 and 1000 time steps, respectively

λ	Iteration steps (275/250)	Iteration steps (1045/1000)
10^{-4}	5	19
10^{-2}	5	19
1	7	19
10^2	4	4
10^4	4	4

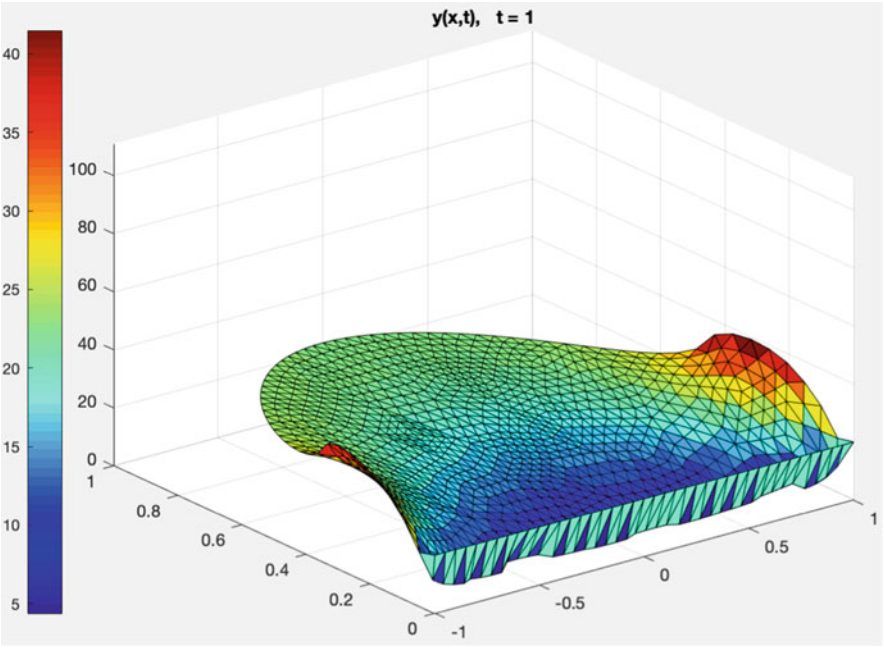

Fig. 7.9 The approximate solution y for final time $t = T = 1$ on a mesh with 1045 nodes for the value $\lambda = 10^{-4}$

researchers could compute exactly these different kinds of numerical experiments in order to study the practical performance of the optimization problem.

After presenting the numerical results, we have to mention again that the step length $\gamma = \gamma^k$ of the projected gradient method has been chosen constant for all k in all computations. However, better results should be achieved by applying a suitable line search strategy, see Remark 7.2. With this we want to conclude that the optimal control problem discussed in this work is one model formulation for solving the optimization of heating a domain such as the air of a swimming pool area surrounded by a glass dome. Modeling and solving the optimal heating of a swimming pool area has potential for many different formulations related to

mathematical modeling, discussing different solution methods and performing many numerical tests including "playing around" with different values for the parameters, constants and given functions, and finally choosing proper ones for the model problem. All of these tasks can be performed by students depending also on their previous knowledge, interests and ideas.

Acknowledgments I would like to thank my students T. Bazlyankov, T. Briffard, G. Krzyzanowski, P.-O. Maisonneuve and C. Neßler for their work at the 26th ECMI Modelling Week which provided part of the starting point for this work. I gratefully acknowledge the financial support by the Academy of Finland under the grant 295897.

References

1. Borzì A. and Schulz V.: Computational Optimization of Systems Governed by Partial Differential Equations. SIAM book series on Computational Science and Engineering, SIAM, Philadelphia (2012)
2. Braess D.: Finite elements: Theory, Fast Solvers, and Applications in Solid Mechanics. Cambridge University Press, second edition (2005)
3. Ciarlet P. G.: The Finite Element Method for Elliptic Problems. Studies in Mathematics and its Applications 4, North-Holland, Amsterdam (1978), Republished by SIAM in 2002.
4. Gruver W. A. and Sachs E. W.: Algorithmic Methods in Optimal Control. Pitman, London (1980)
5. Herzog R. and Kunisch K.: Algorithms for PDE-constrained optimization. GAMM–Mitteilungen 33.2, 163–176 (2010)
6. Hinze M., Pinnau R., Ulbrich M., and Ulbrich S.: Optimization with PDE constraints. Mathematical Modelling: Theory and Applications 23, Springer, Berlin (2009)
7. Jung M. and Langer U.: Methode der finiten Elemente für Ingenieure: Eine Einführung in die numerischen Grundlagen und Computersimulation. Springer, Wiesbaden, second edition (2013)
8. Kelley C. T.: Iterative Methods for Optimization. SIAM, Philadelphia (1999)
9. Lions J. L.: Optimal Control on Systems Governed by Partial Differential Equations. Springer, Berlin-Heidelberg-New York (1971)
10. Nocedal J. and Wright S. J.: Numerical Optimization. Springer, New York (1999)
11. Troeltzsch F.: Optimal Control of Partial Differential Equations. Theory, Methods and Applications. Graduate Studies in Mathematics 112, AMS, Providence, RI (2010)

Chapter 8
Some Basic Epidemic Models

Danijela Rajter-Ćirić

8.1 Introduction

Spreading of infectious diseases has always been a threat to human health and people have been trying to fight against it (which is especially important nowadays when contagious diseases are spreading faster and further than ever). So far great achievements have been made. In order to prevent the spread of a particular disease, one should first try to understand and explain the mechanism how it spreads in the population. However, no experiments are possible due to ethical (and many other) reasons. Therefore, mathematical models present a very useful tool. Although they are only theoretical and usually simplify the real situation, they still describe the behaviour of population members well enough so that they can be successfully used for describing the dynamics of a disease, predicting epidemics, measuring the effects of some prevention measures, etc.

At the beginning of the twentieth century Dr. Ross, later awarded the Nobel Prize for Medicine for his significant contribution to research, used a differential equation model to describe malaria transmission between humans and mosquitoes. Later, William Kermack and Anderson McKendrick formulated a model to study the Black Death outbreak in London and the plague outbreak in Mumbai. They published their results in 1927 in the paper "A Contribution to the Mathematical Theory in Epidemic". They have used one of the simplest forms of, so-called, SIR model which has been studied, improved and generalized afterwards by many authors.

Today there are many different mathematical epidemic models and mathematical approach to the epidemic modelling is widespread. In this paper we introduce

D. Rajter-Ćirić (✉)
Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia
e-mail: rajter@dmi.uns.ac.rs

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_8

the readers to some of these models. All mathematical models can roughly be divided into two groups: deterministic and stochastic models. In the paper, we first describe a few deterministic models for spreading of a contagious diseases in a large population, which are based on the, so-called, mass action law (population members make contacts to other members independently of each other and each individual has an equal chance of contacting any other individual). Further on, we present a few ideas of how a stochastic approach can be used in epidemic modelling. A stochastic approach is reasonable and it is justified by the fact that a population does not behave in a precisely determined way as it is assumed in deterministic models.

It is important to emphasize that in this paper models are presented on a very basic level, without complicated mathematical proves and getting deeper into the theory. The paper should simply serve to introduce readers into a beautiful field of applied mathematics called epidemic modelling and to present how nicely mathematics can be applied to such a serious research area. There are no original results in the paper. The models presented here have been considered in many student books and papers.

8.2 Some Deterministic Models

Deterministic models assume that the population behaves exactly as assumed in the model and there are no randomness in the behavior of the population. The population is large and divided in groups by the epidemic state of individuals. The number of groups depends on the disease and hence on the model, as the reader will see. Here we present only some of many models. In all of presented models ordinary differential equation approach has been proposed. Although in most cases, corresponding systems of ordinary differential equations describing the model are not solvable (if not analytically, these ordinary differential equation systems can be solved numerically), they still play a significant role in the mathematical analysis of the disease spread and in the prediction of epidemics. For more about deterministic epidemic models we refer the reader, for instance, to [2, 3].

8.2.1 SIR Model

An SIR model is a very simple epidemic model that one can use to calculate the number of individuals infected with an epidemic (a contagious) disease in a large population over time. One of the simplest SIR models is the Kermack-McKendrick model.

We consider the population of size N, where N is a constant, and assume that the population consists of three types of individuals based on the state of the individual concerning the disease. In this model we assume that there are only three possible states: a subject is sensitive, infected or immune to a virus. Therefore, the

population is divided into three different groups. The first group consists of those individuals who have not developed immunity against the virus. That is the group of susceptibles (population members that are not infected but could become infected). The second group consists of infectives (subjects who are infected with the virus which means that they have the disease and can transmit it to the members of the group of susceptibles). Finally, members of the third group are individuals who have recovered from the disease and gained lasting immunity or who have died from the disease. In both cases, those individuals are said to be removed. (Some authors call the third group Recovered instead of Removed since they consider the individuals who have died from the disease being the same, from the epidemic point of view, as those who have recovered, since both recovered and dead are, in some sense, immune to the virus).

So, basically, there is a very simple "rule" in SIR model: After becoming infected a susceptible subject immediately enters the infected group. Afterwards (after recovering or dying from the disease) the subject enters the group of removed.

The numbers of group members for these three groups are denoted by the letters S, I and R, respectively, which is the reason why this is called SIR model. All these numbers are actually functions of time t:

$$S(t) \text{ – denotes the number of susceptibles at time } t$$

$$I(t) \text{ – denotes the number of infectives at time } t$$

$$R(t) \text{ – denotes the number of removed at time } t.$$

In the simplest SIR model that we first consider, a short time scale has been assumed so that births and deaths (other than deaths from this disease) can be neglected. One can consider the case when births and deaths are taken into account which yields to a slightly more complicated model as we will see later.

The usual assumptions for SIR model are:

1. Individuals infect each other directly rather than through disease vectors.
2. Contacts between individuals are random.
3. Immediately after a contact with the infected person, susceptible person shows symptoms and can infect someone else.
4. An arbitrary population member makes βN contacts (within the population) in a unit of time, where β denotes disease transmission rate.
5. The number of population members that move from the group of infectives to the group of removed in the unit of time is $\alpha I(t)$, where α denotes, so-called, recovery rate.

Now we want to answer the question: How $S(t)$, $I(t)$ and $R(t)$ vary with time?

As an answer, the SIR model proposes a system of ordinary differential equations representing the transition from one group to another. More precisely, the numbers of susceptibles, infectives and removed change according to the system:

$$S'(t) = -\beta S(t) I(t) \tag{8.1}$$

$$I'(t) = \beta S(t) I(t) - \alpha I(t) \tag{8.2}$$

$$R'(t) = \alpha I(t). \tag{8.3}$$

Assuming that every population member belongs to one of the three groups one has that, at every time t,

$$S(t) + I(t) + R(t) = N.$$

Therefore Eq. (8.3) can be omitted.

Note that, based on one of the model assumptions, every infected individual makes βN different contacts i.e., contacts with βN members of population, but the chance to make the contact with a susceptible person is $\frac{S}{N}$. Thus

$$\beta N \, \frac{S}{N} \, I = \beta S I$$

is the number of individuals who move from the susceptible group to the group of infected in unit of time. Therefore, $S(t)$ decreases, while $I(t)$ increases for that number. This explains the differential equations in the model above.

The dynamics of the infectious group depends on the ratio $R_0 = \frac{\beta}{\alpha}$. The number

$$R_0 = \frac{\beta N}{\alpha}$$

is the, so-called, basic reproduction number and it represents the expected number of new infections from a single infection in a population where all subjects are susceptible. (For more about this number we refer the reader, for instance, to [4].)

The basic reproduction number is very important since it is a good epidemic indicator. If $R_0 > 1$, many susceptible individuals will be infected, i.e., the epidemic will start. If $R_0 = 1$ the disease becomes endemic. If one wants to prevent the epidemic, it is necessary to keep R_0 less than 1. For instance, the vaccination is a possible way for keeping the basic reproduction number lower than 1. Assume that p is the proportion of population members who have been successfully vaccinated before the appearance of the first infected individual. For preventing the epidemic the following condition has to be satisfied:

$$R_0 = \frac{\beta}{\alpha} (1 - p) N < 1,$$

i.e.,

$$p > 1 - \frac{1}{R_0} = 1 - \frac{\alpha}{\beta N}.$$

In the model above the function $F = \beta I$ models the transition rate from the group of susceptible individuals to the group of infectious individuals. Therefore, it is called the force of infection. For many contagious diseases it is more realistic to consider a force of infection that does not depend on the absolute number of infectious subjects, but on their fraction $F = \beta \frac{I}{N}$. Some authors have even proposed nonlinear forces of infection to model more realistically the processes of contagious diseases.

Let us just briefly mention the more general case when birth and death rates influence the model. Suppose that λ denotes the birth rate and that μ denotes the death rate in the population. We still assume that the size of the population is constant. In that case, the SIR model is described by the following system:

$$S'(t) = \lambda - \mu S(t) - \beta I(t) S(t) \tag{8.4}$$

$$I'(t) = \beta I(t) S(t) - \alpha I(t) - \mu I(t) \tag{8.5}$$

$$R'(t) = \alpha I(t) - \mu R(t). \tag{8.6}$$

Also, one can go a step further and study the (more realistic) SIR model that includes the vital dynamics (birth and death rates) in the population of size which is not a constant anymore, but varies with time. Here we will not consider that case.

8.2.2 SEIR Model

Now we present a modification of the SIR model that is very realistic for many infectious diseases. Instead of the assumption that after every contact with an infected subject a susceptible subject gets immediately infected and can infect others, here we assume that there is an incubation period during which the individuals have been infected but are not yet infectious themselves. That period is significant for many contagious diseases. Therefore, in this model another group of population members is formed - the group of exposed members (individuals who are in the incubation period), which means that now the population is divided in four groups: susceptible, exposed, infected and removed. Newly infected members do not immediately move from the susceptible group to the infected group, but first they go into the exposed group. Same as in SIR model, after being in the group of infected, subjects move to the group of removed.

Denote by $E(t)$ the number of exposed individuals at time t.

Assuming that the average of the incubation period is κ^{-1} and that births and deaths (other than deaths due to the disease) have no influence to the model, the SEIR model is represented by the following system of ordinary differential equations:

$$S'(t) = -\beta S(t) I(t) \tag{8.7}$$

$$E'(t) = \beta S(t)I(t) - \kappa E(t) \tag{8.8}$$

$$I'(t) = \kappa E(t) - \alpha I(t). \tag{8.9}$$

The same as in the SIR model that we considered above, it is enough to consider only three out of four differential equations, since $S(t) + E(t) + I(t) + R(t) = N$ is a constant.

If infectivity of an exposed person can be reduced by some factor δ then one obtains more general SEIR model represented by the system:

$$S'(t) = -\beta S(t)I(t) - \delta\beta S(t)E(t) \tag{8.10}$$

$$E'(t) = \beta S(t)I(t) + \delta\beta S(t)E(t) - \kappa E(t) \tag{8.11}$$

$$I'(t) = \kappa E(t) - \alpha I(t). \tag{8.12}$$

Note that Eq. (8.10) describes the following: The number of susceptible subjects decreases by contacts with an infected subject or with an exposed subject but not every contact with an exposed person leads to infection transmission (the number of contacts that lead to infection is reduced by factor δ).

The basic reproduction number is now given by:

$$R_0 = \frac{\beta N}{\alpha} + \frac{\delta\beta N}{\kappa}.$$

It shows how many subjects can be infected by one exposed subject entering the group S and it can be explained in the following way: An exposed person makes $\frac{\beta N}{\kappa}$ different contacts in the group during the incubation period of the length $\frac{1}{\kappa}$, but not every contact leads to infection, as we mentioned above. Therefore there are $\frac{\delta\beta N}{\kappa}$ newly infected subjects. After the incubation period, the exposed person from above becomes infected and can make $\frac{\beta N}{\alpha}$ contacts but now every contact leads to the infection transmission.

In the end, let us briefly remark that, similarly as in SIR model, one can consider SEIR model assuming the presence of birth and death rates. It is very common to consider the case with birth and death rates that are equal. If μ denotes the birth/death rate, one has the model:

$$S'(t) = \mu N - \mu S - \beta \frac{I}{N} S$$

$$E'(t) = \beta \frac{I}{N} S - (\mu + \kappa)E$$

$$I'(t) = \kappa E - (\alpha + \mu)I$$

$$R'(t) = \alpha I - \mu R.$$

Here we wrote all four equations although we could have omitted the last one since $S(t) + E(t) + I(t) + R(t) = N$ is a constant due to the assumption that birth and death rates are equal. However, in general N is a variable, not a constant.

8.2.3 SLIAR Model

This model includes a, so-called, latent period during which the person is infected, but there are still no symptoms of the disease and the person cannot transmit the virus to other members of the population. This is the case with influenza, and some authors call this model an influenza model.

For this model first it is necessary to form a population group of individuals that are in the latent period. When an individual gets out from the latent period, symptoms of the disease may or may not develop. If symptoms develop, then the individual moves into the infected group, and if that does not happen, then the person is in the, so-called, asymptomatic period when he or she does not have the symptoms of the disease but can transmit the infection to the others with a reduced factor ε. So, the model requires one more group to be formed—the group that consists of population members who are in the asymptomatic period.

Denote by $L(t)$ the number of population members who are in the latent period and by $A(t)$ the number of population members who are in the asymptotic period.

We also assume that the proportion of p out of the total number of those who are in the latent period goes into the infected group, which implies that the proportion of 1-p goes into the group of those who are in the asymptomatic period.

The model is called SLIAR model by the first letters of the names of five groups: Susceptible, Latent, Infected, Asymptotic and Removed.

The system of ordinary differential equation that describes the SLIAR model is:

$$\begin{aligned} S'(t) &= -\beta S(t)\,[I(t) + \varepsilon A(t)] \\ L'(t) &= \beta S(t)\,[I(t) + \varepsilon A(t)] - \kappa L(t) \\ I'(t) &= p\kappa L(t) - \alpha I(t) \\ A'(t) &= (1-p)\kappa L(t) - \eta A(t). \end{aligned}$$

From the first equation one sees that the number of susceptibles decreases after the contact with an infected subject or with a subject which is in asymptotic period (but not every contact with a subject in an asymptotic period yields to the infection, that is why one has ε multiplying A in the equation). As in previous cases, the equation that shows how $R(t)$ varies has been omitted due to the fact that number $S(t) + L(t) + I(t) + A(t) + R(t) = N$ is a constant.

The basic reproduction number in the SLIAR model is:

$$R_0 = p\frac{\beta N}{\alpha} + (1-p)\frac{\varepsilon\beta N}{\eta}$$

and it shows how many susceptible population members get infected by one subject who is in the latent period.

8.2.4 SIS Model

This model describes the disease that is endemic. It is a model of a disease in which the infected do not acquire immunity after recovery. This means that there are only two groups here: the group of the vulnerable and the group of the infected ones. After recovery the infected subjects return back to the sensitive group. Such a model can be applied in modeling the spread of diseases caused by a bacterium, because then immunity is not acquired against a new infection caused by the same bacterium. A model that does not include birth and death rates will be considered first. We again assume that the population size is constant. Then the system of differential equations corresponding to this model is as follows

$$\begin{aligned} S'(t) &= -\beta S(t)I(t) + \alpha I(t) \\ I'(t) &= \beta S(t)I(t) - \alpha I(t). \end{aligned}$$

Since $N = S(t) + I(t)$, for every t, the previous system reduces to the equation

$$I'(t) = \beta\left[N - I(t)\right]I(t) - \alpha I(t)$$

The equation above can be written in the form:

$$I' = \beta I\left(M - I\right),$$

where $M = N - \dfrac{\alpha}{\beta}$. The last equation can easily be solved:

$$I(t) = \frac{M}{\exp\{-M(\beta t + c)\} + 1}, \quad \text{for } M > I$$

$$I(t) = \frac{-M}{\exp\{-M(\beta t + c)\} - 1}, \quad \text{for } M < I$$

However, in any case one can see that the following holds:

- If $M > 0$ then $\lim_{t\to\infty} I(t) = M$.
- If $M \leq 0$ then $\lim_{t\to\infty} I(t) = 0$.

The basic reproduction number in this model is the same as in SIR model: $R_0 = \frac{\beta N}{\alpha}$. Therefore $M = N\left(1 - \frac{1}{R_0}\right)$ and one concludes the following:

- If $R_0 > 1$ then $\lim_{t\to\infty} I(t) = M$ i.e., the disease remains in the population
- If $R_0 \leq 1$ then $\lim_{t\to\infty} I(t) = 0$ i.e., the disease vanishes.

Finally, let us just mention that one can consider the SIS model with birth and death rates both being equal to μ and obtain a generalization of the previous model:

$$I'(t) = \beta\,[N - I(t)]\,I(t) - (\mu + \alpha)I(t). \tag{8.13}$$

8.3 Some Stochastic Models

The assumption that a population behaves exactly as assumed in the model is not very realistic. There are always some randomness that affect the population behavior. Therefore, a stochastic approach to the epidemic modelling problems is reasonable. There are different stochastic approaches depending on many factors and hence there are many different stochastic models. One of the simplest approaches is the one that uses discrete-time Markov chain models. Thus here we present a model of that type first. Some other stochastic models involve stochastic differential equations. Here we just briefly mention one of such models. Finally, there are many stochastic processes that can be used in epidemic modelling and here we present how a Poisson process can be used.

8.3.1 SIS Model in the Form of Discrete-Time Markov Chain

In this section we describe the SIS model with birth and death effects in a form of discrete-time Markov chain (see [1] for details). We assume that birth and death rates are equal and denoted by μ. The population size is constant and denoted by N. Therefore, as we concluded in the SIS deterministic case, it is enough to consider only one variable and this will (again) be the number of infected subjects, $I(t)$.

So, we consider a stochastic process $\{I(t),\ t \in T\}$, where $T = \{0, \Delta t, 2\Delta t, \dots\}$, as a discrete-time Markov chain. From the epidemic point of view it is reasonable to assume that the number of infectives at a time moment depends only on the number of infectives in the previous moment, so it is reasonable to assume that $I(t)$ satisfies the Markov property.

As we saw in the deterministic case (see (8.13)) the following holds:

$$I'(t) = \beta\,[N - I(t)]\,I(t) - (\mu + \alpha)\,I(t).$$

Since $I(t)$ is number of infected subjects at time t it is obvious that the set of possible states in this case is $S = \{0, 1, \ldots, N\}$. The probability that process I is in the state $i \in S$ at time t, is denoted by $p(t)$, i.e. $p(t) = P\{I(t) = i\}$. The, so-called, probability vector is given by

$$p(t) = [p_0(t), \ldots, p_N(t)]^T,$$

and $p_0(t) + p_1(t) + \cdots + p_N(t) = 1$.

The next step is to determine the transition probabilities from one state to another for a short period of time Δt:

$$p_{ij}(t + \Delta t) = P\left\{I(t + \Delta t) = j | I(t) = i\right\}.$$

Based on the deterministic case, we assume that the Markov chain $I(t)$ is homogeneous i.e., that transition probabilities do not depend on time. Thus, we can write $p_{ij}(\Delta t)$ instead of $p_{ij}(t + \Delta t)$.

In order to make the model as simple as possible, we also assume that Δt is small enough so that during that time period the number of infected subjects can change for one at most, i.e., there are three possible state changes:

$$i \to i + 1, \quad i \to i - 1 \quad \text{or } i \to i.$$

Now the transition probabilities are given by:

$$p_{ij}(\Delta t) = \begin{cases} \beta\ i\ (N - i)\ \Delta t, & j = i + 1 \\ (\mu + \alpha)\ i\ \Delta t, & j = i - 1 \\ 1 - [\beta i(N - i) + (\mu + \alpha)i]\,\Delta t, & j = i \\ 0, & \text{otherwise} \end{cases} \tag{8.14}$$

If we set $b_i := \beta i(N - i)$ and $d_i := (\mu + \alpha)i$ we obtain:

$$p_{ij}(\Delta t) = \begin{cases} b_i \Delta t, & j = i + 1 \\ d_i \Delta t, & j = i - 1 \\ 1 - [b_i + d_i]\,\Delta t, & j = i \\ 0, & \text{otherwise} \end{cases} \tag{8.15}$$

Note that Δt has to be small enough to provide that $p_{ij} \in [0, 1]$. Therefore, the following must hold:

$$\max_{i\in\{1,\ldots,N\}} \{(b_i + d_i)\Delta t\} \leq 1.$$

Using the transition probabilities from (8.15) one can determine the probability that there are i infected subjects at time $t + \Delta t$:

$$p(t+\Delta t) = p_{i-1}(t)b_{i-1}\Delta t + p_{i+1}(t)d_{i+1}\Delta t + p_i(t)\left(1 - [b_i + d_i]\,\Delta t\right), \quad i = 1, \ldots, N.$$

Finally, although we will not prove it here, let us mention that for expected number of infected subjects the following holds:

$$E\left(I(t + \Delta t)\right) = E(I(t)) + [\beta N - (\mu + \alpha)]\,E(I(t))\Delta t - \beta E(I^2(t))\Delta t.$$

Using the fact that $E(I^2(t)) \geq E^2(I(t))$ and letting $\Delta t \to 0$ one obtains that

$$\frac{dE(I(t))}{dt} \leq \beta\,[N - E(I(t))]\,E(I(t)) - (\mu + \alpha)E(I(t)). \tag{8.16}$$

8.3.2 A Note on Stochastic Differential Equation for SIS Model

Here we introduce and just briefly describe the stochastic differential equation for SIS model. For details we refer the reader to [5].

As already mentioned, for the SIS model Eq. (8.13) holds. This equation can be written as

$$dI(t) = \beta\,(N - I(t))\,I(t)dt - (\mu + \alpha)\,I(t)dt. \tag{8.17}$$

Consider the first summand in the sum above: $\beta\,(N - I(t))\,I(t)dt = \beta S(t)I(t)dt$. It represents the number of the newly infected individuals in the time interval od the length dt. If we make a reasonable assumption that β is actually the random variable and that instead of βdt in (8.17) one can write $\beta dt + \sigma dW(t)$, where $W(t)$ is standard Brownian motion (Wiener process), then we obtain a stochastic differential equation:

$$dI(t) = I(t)\left([\beta(N - I(t)) - (\mu + \alpha)]\,dt + \sigma\,(N - I(t))\,dW(t)\right). \tag{8.18}$$

One can prove the following: If $R_0^S = R_0 - \dfrac{\sigma^2 N^2}{2(\mu+\alpha)} = \dfrac{\beta N}{\mu+\alpha} - \dfrac{\sigma^2 N^2}{2(\mu+\alpha)} < 1$ and $\sigma^2 \leq \frac{\beta}{N}$ then, for every initial data $I(0) \in (0, N)$, the solution $I(t)$ to stochastic

differential equation (8.18) exponentially tends to zero, almost surely. In another words, the disease vanishes with probability 1.

8.3.3 A Poisson Process Model for Tracking the Number of HIV Infections

We present a very simple Poisson process model for tracking the number of HIV infections, as done in [6].

One of many difficulties with HIV infection is the fact that the incubation period is relatively long. So, there may be many individuals who are infected with the virus but still not showing the symptoms. The following model is a very simple approximation model that helps obtaining a rough estimate of the number of such individuals.

In this model we assume that

- HIV infections appear in accordance with a Poisson process with unknown rate λ,
- the time from the moment when an individual becomes infected until symptoms of the disease appear is a random variable that has a known distribution G,
- the incubation periods of different infected individuals are independent.

Let $N_1(t)$ denotes the number of individuals who have shown symptoms of the disease by time t and $N_2(t)$ denotes the number of individuals who are HIV positive but still don't show any symptoms of the disease by time t.

Since a subject who gets infected at time s will have symptoms by time t with probability $G(t-s)$ and will not with probability $1-G(t-s) = \bar{G}(t-s)$, it follows that $N_1(t)$ and $N_2(t)$ are independent Poisson random variables with means

$$E(N_1(t)) = \lambda \int_0^t G(t-s)\,ds = \lambda \int_0^t G(y)\,dy,$$
$$E(N_2(t)) = \lambda \int_0^t \bar{G}(t-s)\,ds = \lambda \int_0^t \bar{G}(y)\,dy.$$

Since λ is unknown, we must estimate it. Suppose that we have reliable records and that we know how many individuals are ill by time t. Denote that number by n_1. Then we can estimate that

$$n_1 \approx E(N_1(t)) = \lambda \int_0^t \bar{G}(y)\,dy.$$

So, we can estimate λ by $\tilde{\lambda}$ given by

$$\tilde{\lambda} = \frac{n_1}{\int_0^t G(y)\,dy}.$$

Using this estimation of λ, we can estimate the number of infected individuals with no symptoms at all at time t by

$$\tilde{N}_2(t) = \tilde{\lambda} \int_0^t \bar{G}(y)\,dy = n_1 \frac{\int_0^t \bar{G}(y)\,dy}{\int_0^t G(y)\,dy}.$$

If, for example, G is exponential with mean μ, then $\bar{G}(y) = e^{-\frac{y}{\mu}}$, and

$$\tilde{N}_2(t) = \frac{n_1\mu(1 - e^{-\frac{t}{\mu}})}{t - \mu(1 - e^{-\frac{t}{\mu}})}.$$

In [6] Ross gives the following concrete example based on the previous assumptions and calculations: If we suppose that $t = 16$ years, $\mu = 10$ years, and $n_1 = 220{,}000$, then the estimation of the number of infected but symptomless individuals at time 16 is

$$\tilde{N}_2(16) = \frac{220 \cdot 10(1 - e^{-1.6})}{16 - 10(1 - e^{-1.6})} = 218.96.$$

So, if the incubation period is exponential with mean 10 years and if the total number of individuals who have AIDS symptoms during the first 16 years of the epidemic is 220,000, then we can expect that approximately 219,000 individuals are HIV positive but with no symptoms at time 16.

So, the model above can be used for getting a rough estimation of number of HIV infections. However, the assumption that the infection rate λ is a constant is not very realistic. It would be much better to use an infection rate that changes over time.

References

1. Allen, L.J.S.: An Introduction to Stochastic Epidemic Models. In: Brauer F., van den Driessche P., Wu J. (eds) Mathematical Epidemiology. Lecture Notes in Mathematics, Vol 1945. Springer, Berlin, Heidelberg (2008)
2. Brauer, F., Castillo-Chavez, C.: Epidemic Models. In: Mathematical Models in Population Biology and Epidemiology. Text in Applied Mathematics, Vol 40. Springer, New York, NY (2012)
3. Brauer, F., Castillo-Chavez, C.: Models for Endemic Diseases. In: Mathematical Models in Population Biology and Epidemiology. Text in Applied Mathematics, Vol 40. Springer, New York, NY (2012)

4. Brauer, F., Driessche, P.V.D., Wu, J.: Mathematical Epidemiology. Springer Science and Business Media, Berlin (2008)
5. Gray, A., Greenhalgh, D., Hu, L., Mao, X., Pan, J.: A Stochastic Differential Equation SIS Epidemic Model. University of Strathclyde, Glasgow, Donghua University Shanghai (2006)
6. Ross, S.: Introduction to Probability Models (Tenth Edition). Academic Press as an Imprint of Elsevier (2010)

Chapter 9
Mathematical Model for the Game Management Plan

Milana Pavić-Čolić

9.1 Introduction

We consider an enclosed hunting area, where the wild animals are grown, such as red deer, wild boar, mouflon. In those areas, humans can have a great influence on breeding. Usually, there is a management team that supervises animal raising, and thus an enclosed hunting area can be understood as a farm.

The overall goals of the management are to protect, sustain, and manage hunted wildlife, provide hunting opportunity, protect and enhance wildlife habitat, and minimize adverse impacts to residents, other wildlife, and the environment [1].

In this work we are particularly interested in the management activities of game population control. Usually this activity is focused on producing high quality trophies: the management controls a number of individuals over some period (or population dynamics) in order to raise the ones with trophy potential. The trophy is usually part a of the animal that the hunter keeps as a souvenir to represent the success of the hunt. It can be horn of a red deer or mouflon, teeth of a wild boar.

Trophy production is usually achieved by following two simple rules: (1) governing population of any initial structure towards the prescribed optimal one and (2) keeping high the number of trophy candidates. The population dynamics that obeys these rules is guided by the so-called Game Management Plan. It is a 10 years time frame document that in particular contains a plan of harvest for each hunting season, which is a main mechanism of population dynamics control.

In our study we focus on red deer population and the hunting area "Vorovo" in Serbia, whose management kindly provided the necessary data. In this case,

M. Pavić-Čolić (✉)
Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia
e-mail: milana.pavic@dmi.uns.ac.rs

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_9

trophy candidates are male deers 8–10 years old. All the predictions that are made in the Game Management Plan for this hunting area are based on experience and do not rely on any mathematical model, so far. Our aim is to put the problem on mathematical ground and to build a model that will improve the current management strategy keeping its simplicity. This is the first important result in direction of the game population control in enclosed hunting area such as "Vorovo".

From the mathematical point of view, a very general framework can be derived, that can be applied to any enclosed hunting area that breeds any type of animals. Therefore, the answer that mathematics can provide would be useful all over the world.

This project was originally presented under the title "Red deer import" at the ECMI Modelling Week 2016, that took place in Sofia, Bulgaria. We studied the mathematical model for the Game Management Plan, and this paper partially contains the observations and results we obtained at the time. Besides that, we tried to understand the inbreeding problem, that is known to be the main problem in enclosed hunting areas. After some time living in those areas, the animals start to mate within family members, which provokes malformations and diseases in the new generations, and has irreparable consequences on trophy production. In order to avoid inbreeding, a hunting area imports new animals. The import process requires good strategy. First, the right time should be determined, and then the question of a number and an optimal structure of imported population regarding both sex and age arise. These problems are still open.

The plan of this chapter is as follows. In Sect. 9.2 we will introduce "Vorovo" hunting area and present the current management strategy, as well as some of its shortcomings. Section 9.3 is devoted to the new model for the Game Management Plan which aims to improve the current model, keeping its simplicity. We illustrate our model on the example of "Vorovo" hunting area in Sect. 9.4.

9.2 "Vorovo" Hunting Area

In this work, we are concerned with hunting areas of red deer population. In particular, we study the hunting area "Vorovo", that is supervised by the Fruška Gora National Park in Serbia. This project aims at improving its strategy for production of the red deer trophies.

9.2.1 The Current Management Strategy

Let us explain the current management strategy for this specific enclosed hunting area. First, an optimal structure of red deer population is determined. This structure is fixed for the hunting area over the years. It is determined on the basis of

Table 9.1 The optimal structure of red deer population in "Vorovo" hunting area

		Age of individuals										
		0	1	2	3	4	5	6	7	8	9	10
Number of individuals	Male	15	10	8	8	8	8	7	6	5	0	0
	Female	15	10	8	8	8	8	7	6	5	0	0

environmental conditions, that include food and water availability, vegetation, climate, peace requirements, pedological composition of the soil (Table 9.1).

The first goal of the management is to reach this optimal structure for any initial population state. Achievement of this goal is driven by a 10 years time-frame document, the so called Game Management Plan.

The Game Management Plan contains for each hunting year (hunting year is from March 1 year until the March next year), within the interval of 10 years, the following data:

1. optimal structure of population,
2. number of individuals on March 31,
3. number of individuals on April 1 (this date can be understood as "dears birthday"),
4. expected number of newborns,
5. number of individuals before hunting,
6. expected loss,
7. expected harvest,
8. number of individuals after hunting,
9. comparison of the number of individuals after hunting with the optimal structure.

We should mention that all the numbers that appear in the Game Management Plan are actually predictions. The only real data, besides the optimal structure, is a number of individuals on March 31 the first hunting year. In order to be sure to follow the Game Management Plan, hunting area needs to organize additional activities each hunting year. Two main such activities are game counting on March (at the beginning of hunting season) and making an Annual Plan. Annual Plan is a document whose aim is to plan the harvest so that the number of individuals at the end of the hunting season matches with the corresponding one from the Game Management Plan. Therefore, the Annual Plan is a sort of a comparison between real data and the predicted one in the Game Management Plan.

All predictions presented in the Game Management Plan are based on experience and do not rely on any mathematical model so far. In particular, harvest is planned in such a way that it ensures two tasks at the same time: (1) reach the prescribed optimal number of individuals at some moment, and (2) produce hunting trophies. Candidates for hunting trophies are 8–10 years old male deers. It is remarkable that the Game Management Plan preserves equilibrium state. More precisely, once the optimal structure is reached, the population dynamics do not change anymore

(for example, the population structure before hunting is the same for each of the following years). Such equilibrium state is represented in Table 9.2.

As we have already mentioned, besides this 10 years time frame document, hunting area organizes a lot of activities during the hunting year.

At the beginning, these activities are related to the observation of the real situation. First, in March (when the vegetation is still low) game counting is organized, when the real data is collected. Then the management makes a first version of the Annual Plan in April, which mostly aims at planning the harvest so that at the end of the hunting year the number of species coincides with the number from the Game Management Plan. In April, calving of females starts. Until the mid June, game warden monitors the hunting area, and in particular pays attention on losses. In mid June, with counted losses, hunting area makes a final version of the Annual Plan with a definite harvest prediction.

Then, the hunting is organized. First, in June and July hunting area organizes hunting for selective culling of newborns. Then, the main event is trophy hunting in September.

For the Modelling Week, "Vorovo" provided necessary data, more precisely the two Game Management Plans, for the intervals 2009–2019 and 2015–2025, as well as Annual Plans for hunting years 2010/2011, 2011/2012, 2012/2013, 2013/2014 and 2014/2015.

9.2.1.1 Some Shortcomings of the Current Management Strategy

When analyzing the data obtained from the park "Vorovo", it can be noticed few shortcomings of the current management strategy. First, it is supposed to become stable in a short time, or in other words, the equilibrium state is supposed to be reached shortly after the period starts. According to the real data we get from Annual Plans, this seems unrealistic, since even the Annual Plan as a sort of control of the Game Management Plan is not successful in reaching desired structure at the end of hunting year.

Besides that, the population dynamics of newborns was not accurately modelled. According to the current model, the number of newborns was predicted as 70% of the number of females 2 years old and older. From the data we obtained, it was clear that the number of newborns was actually much higher.

Our main contribution is at improving this current management strategy. In the rest of the paper we explain our new proposed model and compare it with the current one. We intent to make a more accurate model for prediction of number of newborns, and to build a new model for the Game Management Plan that is less ambitious in how fast the equilibrium should be reached (and therefore seems more realistic) and that offers a good chance for trophy hunting.

Table 9.2 The equilibrium state

Population dynamics	Sex	Age of individuals											Sum	
		0	1	2	3	4	5	6	7	8	9	10		
Optimal number	Male	15	10	8	8	8	8	7	6	5	0	0	75	150
	Female	15	10	8	8	8	8	7	6	5	0	0	75	
Number on March 31	Male	15	10	8	8	8	8	7	6	5	0	0	75	150
	Female	15	10	8	8	8	8	7	6	5	0	0	75	
Number on April 1	Male	*xx*	15	10	8	8	8	8	7	6	5	0	75	150
	Female	*xx*	15	10	8	8	8	8	7	6	5	0	75	
Expected growth	Male	21	*xx*	*xx*	xx	xx	xx	xx	xx	xx	xx	xx	21	42
	Female	21	*xx*	*xx*	xx	xx	xx	xx	xx	xx	xx	xx	21	
Number before hunting	Male	21	15	10	8	8	8	8	7	6	5	0	96	192
	Female	21	15	10	8	8	8	8	7	6	5	0	96	
Expected loss	Male	2	1	0	0	0	0	1	0	0	1	0	5	10
	Female	2	1	0	0	0	0	1	0	0	1	0	5	
Expected harvest	Male	4	4	2	0	0	0	0	1	1	4	0	16	32
	Female	4	4	2	0	0	0	0	1	1	4	0	16	
Number after hunting	Male	15	10	8	8	8	8	7	6	5	0	0	75	150
	Female	15	10	8	8	8	8	7	6	5	0	0	75	
Relation to optimal number	Male	0	0	0	0	0	0	0	0	0	0	0	0	0
	Female	0	0	0	0	0	0	0	0	0	0	0	0	

9.3 A New Proposed Model for the Management Strategy

The first problem that attracted our attention was how to improve the current management strategy, or to obtain a new model for Game Management Plan, such that it keeps the fundamental goals:

1. it predicts well the number of newborns,
2. it reaches the optimal structure within 10 years,
3. it keeps high the number of males older than 8 years, which are candidates for hunting trophies.

In the following two sections we describe our proposed model.

9.3.1 A New Model for the Number of Newborns

According to the current strategy, the number of newborns in one hunting year is planned as 70% of the number of females 2 years old and older on the date March 31 that hunting year. If one wants to check this model, it is needed to deal with the real data. This information is not available for the newborns. What is known is the number of "0" years old deers the following hunting season. In the meantime, we had harvest and some loss. Assuming that the harvest is realized with 100%, then one strategy for estimating the real number of newborns in one hunting year is the following: count number of "0" years old deers the next year and add the expected number of hunted and lost newborns the current year.

If we try to compare the current model with the real data obtained in such a way, we obtain a lot of disagreements, that are illustrated in Table 9.3, the third and the first row, respectively. We can see that, for example, in 2011/2012 the expected number of newborns is 66 according to the current model. But on March 31 the following 2012/2013 it was found 81 individuals, and 29 individuals was planned to be hunted in 2011/2012 . Therefore, if 29 individuals was hunted, then actually the

Table 9.3 Modelling the number of newborns: estimated number of newborns according to different models over hunting seasons

	Hunting season				
	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015
Real data with 100% realized harvest	102	110	90	68	81
Real data with 70% realized harvest	91	101	82	60	71
Current model	78	66	73	68	66
New model	86	76	90	82	70

number of newborns in 2011/2012 was 110. Thus, there is a mistake of 44 newborns. The same error appears in all other years for which we have data.

Having discovered an inaccuracy of the current model, let us try to fix it. Still, we want to keep two main properties of the current model: (1) we want to have the number of newborns modeled as a certain percentage of certain population on the current year i.e. on March 31, and (2) once the equilibrium is reached, we need to stay in that state.

The big error in expected and real (under the assumption of 100% realization of harvest) number of newborns led us to assume that the harvest is realized with 70% of success. These new estimated real data are shown in the second row of Table 9.3.

In order to keep the main ideas of the current model, we looked into relation of number of newborns (real, with weighted harvest by 0.7) with respect to many cases: number of females of 2 years old and older, the sum of females older than 2 years and males older than 6 years and the whole population. We obtained that the relation of the number of newborns with respect to the whole population on date March 31, had the least standard deviation with respect to the average that we approximate with $21/150$. Moreover, as we will see, this number preserves the equilibrium state.

With this conclusion, we build our model. Let Σ_j denote the sum of the whole population on March 31 of the hunting year $j/j+1$, $j = 0, \ldots, 9$. Then our new proposed model predicts the number of newborns, males and females, in the hunting year $j/j+1$, denoted with $N_{j/j+1}$, as follows

$$N_{j/j+1} = 2\frac{21}{150}\Sigma_j. \tag{9.1}$$

The results are shown in the last row of Table 9.3.

In Fig. 9.1 we plot the real data (assuming harvest is realized with 70% of success) and predictions according to the current model and our new proposed model.

Note that in the equilibrium state for some hunting season $j/j+1$ the whole population sum on March 31 is $\Sigma_j = 150$, and therefore the number of newborns is then

$$N^{eq}_{j/j+1} = 42, \tag{9.2}$$

which coincides with expected growth in equilibrium state shown in Table 9.2.

9.3.2 A New Model for the Game Management Plan

The current schema for the Game Management Plans do not follow any mathematical model, so far. It is based solely on experience. Given the real data on March 31 the first year of the Game Management Plan, the number of newborns is predicted, as well as losses and harvest within the whole population. So we obtain the state of

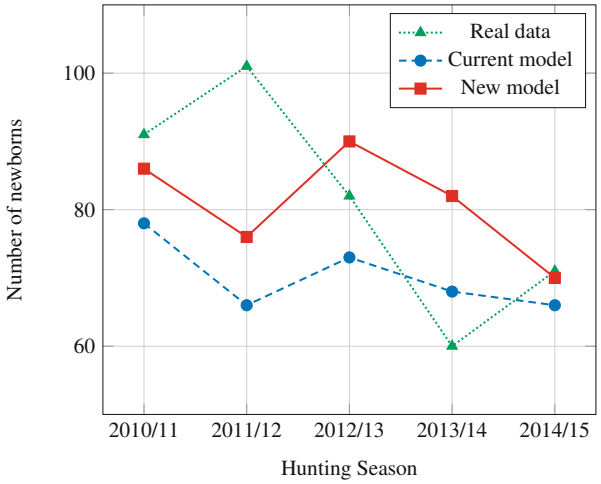

Fig. 9.1 Number of newborns, the real data versus the current and our new proposed model

population after the hunting current hunting year, which is precisely the population state on the date March 31 the next year. Then again newborns, losses and harvest are predicted and we obtain the population structure after the hunting the following hunting year. This procedure is repeated for the whole period of 10 years. We observed that (1) the Annual Plans as a sort of control of the Game Management Plan show that the real state differs significantly from the predicted one, and (2) the equilibrium state is reached very fast, in 2–3 years, which seems unrealistic (according to the Annual Plans, the equilibrium state was actually never reached!).

We propose a new model for the Game Management Plan that drives population dynamics for the next 10 years. Our model is more relaxed regarding expected harvest, in the sense that it predicts less hunting activities (and therefore seems more realistic than the current one). Consequently, the population structure according to our model does not go to the equilibrium state very fast, but it achieves the original goal of the Gama Management Plan, that is, it reaches equilibrium state within 10 years.

Our model is based on what we call *survival probabilities*. Let denote with $p_{i;j/j+1}$ the survival probability of a male deer population of age i before hunting on the hunting year $j/j+1$, where i ranges from 0 to 10. Moreover, let $M_{i;j/j+1}$ be the number of male deers of age i before hunting on April 1 the hunting year $j/j+1$. With the survival probability we calculate the number of deers of age $i+1$ before hunting on the following hunting year $j+1/j+2$, as follows

$$M_{i+1;j+1/j+2} = p_{i;j/j+1} M_{i;j/j+1}. \tag{9.3}$$

For the female deers consideration is similar. We do not make difference in survival probabilities between males and females, but only in their numbers. We denote with

$F_{i;j/j+1}$ the number of female deers of age i before hunting on the hunting year $j/j+1$. Their number the following year is modelled with

$$F_{i+1;j+1/j+2} = p_{i;j/j+1} F_{i;j/j+1}. \tag{9.4}$$

The initial state is the number of individuals 1 year old and older on April 1 the first hunting season 0/1, represented by the matrix

$$\begin{bmatrix} M_{1;0/1} & M_{2;0/1} & M_{3;0/1} & M_{4;0/1} & M_{5;0/1} & M_{6;0/1} & M_{7;0/1} & M_{8;0/1} & M_{9;0/1} & M_{10;0/1} \\ F_{1;0/1} & F_{2;0/1} & F_{3;0/1} & F_{4;0/1} & F_{5;0/1} & F_{6;0/1} & F_{7;0/1} & F_{8;0/1} & F_{9;0/1} & F_{10;0/1} \end{bmatrix}. \tag{9.5}$$

We resume the idea of our population dynamics model: take the initial state matrix (for the season 0/1), calculate the number of newborns according to the model we presented in Sect. 9.3.1, model the survival probabilities $p_{i;0/1}$, $i = 0, \ldots, 10$, and then get the population (of 1 year old deers and older) prediction for April 1 the next season 1/2. We repeat the procedure: calculate the number of newborns (for the season 1/2), from the modelling of survival probabilities $p_{i;1/2}$, $i = 0, \ldots, 10$, obtain the population (of 1 year old deers and older) prediction for April 1 the following season 2/3. The aim is to repeat this procedure iteratively for 10 years. Therefore, the main aspect of our model is to get the survival probabilities.

In what follows, we are going to model the population survival probabilities. We make a difference between the survival probabilities for newborns ($i = 0$) and all the other deers ($i = 1, \ldots, 10$). Let us model first the survival probability for the population of deers older than 1 year in Sect. 9.3.2.1. Then in Sect. 9.3.2.2 we present our model of survival probabilities for newborns. Finally, we are going to apply our model on a concrete example in Sect. 9.4. For this, we will use the population state from the Annual Plan 2015/2016 as an initial state

$$\begin{bmatrix} 23 & 15 & 18 & 7 & 8 & 7 & 9 & 5 & 5 & 7 \\ 24 & 17 & 14 & 15 & 8 & 10 & 6 & 12 & 9 & 0 \end{bmatrix}. \tag{9.6}$$

This is precisely the first data in the Game Management Plan 2015–2025 that we got from "Vorovo", which allow us to compare our new proposed model and the current one.

9.3.2.1 The Survival Probabilities for 1 Year Old Deers and Older

When modelling the survival probabilities for the deers 1 year old and older, we take the simplest model. First, we assume all survival probabilities are constant over the years,

$$p_{i;j/j+1} = p_i, \quad \forall j = 1, \ldots, 9,$$

Table 9.4 Determining the survival probabilities

Population dynamics	Sex	Age of individuals										
		0	1	2	3	4	5	6	7	8	9	10
Number before hunting 1 year	Male	21	15	10	8	8	8	8	7	6	5	
	Female	21	15	10	8	8	8	8	7	6	5	
Number before hunting the following year	Male		15	10	8	8	8	8	7	6	5	0
	Female		15	10	8	8	8	8	7	6	5	0

Table 9.5 Equilibrium survival probabilities

p_0^{eq}	p_1	p_2	$p_3 = p_4 = p_5$	p_6	p_7	p_8	$p_9 = p_{10}$
$\frac{15}{21}$	$\frac{10}{15}$	$\frac{8}{10}$	1	$\frac{7}{8}$	$\frac{6}{7}$	$\frac{5}{6}$	0

where $i = 1, \ldots, 10$, and therefore our probabilities now depend only on deer age. In order to determine these probabilities, we recall that any model needs to preserve the equilibrium state once it is reached. Thus, we are motivated to look into the equilibrium state from Table 9.2 and extract information of the population state before hunting in two successive hunting years (Table 9.4).

We can see that the deer of age one (male or female) will "survive" and have 2 years the following hunting season with the probability $p_1 = \frac{10}{15}$. Similarly, 2 years old deer will have 3 years the following hunting season with the probability $p_2 = \frac{8}{10}$. In the same fashion we obtain all the other probabilities, represented in Table 9.5. We call them *equilibrium survival probabilities* since they are determined while referring to the equilibrium state.

For deers 1 year old and older for the survival probabilities we take the equilibrium ones. Only for newborns we make a different model in the following section.

9.3.2.2 The Survival Probability for the Newborns

In this section we model the probabilities $p_{0;j/j+1}$, for $j = 0, \ldots, 9$.

For the first hunting season $0/1$, we start with the given initial state (9.5). We calculate the sum of the whole population that year, denoted with Σ_0,

$$\Sigma_0 = \sum_{i=1}^{10} \left(M_{i;0/1} + F_{i;0/1} \right).$$

Then we predict the number of newborns according to the model (9.1),

$$M_{0;0/1} = F_{0;0/1} = \frac{1}{2} N_{0/1} = \frac{21}{150} \Sigma_0. \tag{9.7}$$

These newborns (that count $M_{0;0/1}$ or $F_{0;0/1}$) will survive and will become 1 year old deers (males or females, respectively) the following hunting season $1/2$ with the probability

$$p_{0;0/1} = \frac{15}{M_{0;0/1}} = \frac{15}{F_{0;0/1}}. \tag{9.8}$$

Now the survival probabilities for the season $0/1$ are known: for $p_{0;0/1}$ we take (9.8) and all the others p_i, $i = 1, \ldots, 10$ are equilibrium ones from Table 9.5. They allow to predict the population state on April 1 the following season $1/2$, i.e. we can calculate $M_{i;1/2}$ and $F_{i;1/2}$, $i = 1, \ldots, 10$ from (9.3) and (9.4), respectively.

For all the following hunting seasons $j/j+1$, $j = 1, \ldots, 9$, the consideration is similar. Namely, from the previous step we first count the whole population on April 1 the hunting season $j/j+1$,

$$\Sigma_j = \sum_{i=1}^{10} \left(M_{i;j/j+1} + F_{i;j/j+1} \right).$$

Then, following the model (9.1), we estimate the number of newborns that year

$$M_{0;j/j+1} = F_{0;j/j+1} = \frac{1}{2} N_{j/j+1} = \frac{21}{150} \Sigma_j.$$

Finally, we model the survival probabilities of the newborns as

$$p_{0;j/j+1} = \frac{15}{M_{0;j/j+1}} = \frac{15}{F_{0;j/j+1}}.$$

This probability in conjunction with p_i, $i = 1, \ldots, 10$, from Table 9.5 gives $M_{i;j+1/j+2}$ and $F_{i;j+1/j+2}$, according to (9.3) and (9.4).

Note that the equilibrium survival probability for the newborns is

$$p_0^{eq} = \frac{15}{21}, \tag{9.9}$$

using (9.2) and (9.7), which coincides with the equilibrium p_0 from Table 9.5.

Taking the initial state (9.6), the newborns survival probabilities are plotted in Fig. 9.2, for the period 2015–2025.

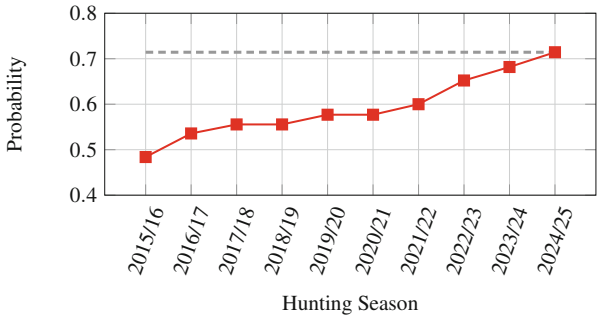

Fig. 9.2 The survival probability for the newborns by a hunting season. Dashed line represents the equilibrium probability p_0^{eq}

9.4 The Game Management Plan 2015–2025 According to the New Proposed Model

We present our new proposed model for the management strategy on an example. For the initial state we take the real data for the hunting season 2015/2016 from (9.6). Using survival probabilities for 1 year old deers and older from Table 9.5 and formulas (9.3) and (9.4), we get population state for 2 years old and older deer the following hunting season 2016/2017:

$$M_{2;2016/2017} = \tfrac{10}{15} * 23 = 15,\ M_{3;2016/2017} = \tfrac{8}{10} * 15 = 12,\ \ldots,\ M_{10;2016/2017} = 0 * 5 = 0,$$
$$F_{2;2016/2017} = \tfrac{10}{15} * 24 = 16,\ \ F_{3;2016/2017} = \tfrac{8}{10} * 17 = 14,\ \ \ldots,\ \ F_{10;2016/2017} = 0 * 9 = 0.$$

We calculate the sum of the whole population

$$\Sigma_0 = \sum_{i=1}^{10} M_{i;2015/2016} + F_{i;2015/2016} = 219.$$

Then we predict the number of newborns according to (9.1)

$$N_{2015/2016} = 2\frac{21}{150}\Sigma_0 = 62.$$

Therefore, from (9.7) we get the prediction of the number of newborns

$$M_{0;2015/2016} = F_{0;2015/2016} = \frac{1}{2}N_{2015/2016} = 31.$$

Finally, the survival probability for the newborns is calculated from (9.8),

$$p_{0;2015/2016} = \frac{15}{31}.$$

Using (9.3) and (9.4) we obtain the deer population 1 year old for the following season 2016/2017:

$$M_{1;2016/2017} = 31 * \frac{15}{31} = 15, \quad F_{1;2016/2017} = 31 * \frac{15}{31} = 15.$$

This completes the data for the hunting season 2016/2017. Proceeding iteratively we build the Game Management Plan.

In Table 9.6 we present our results. For simplicity, we show only data for the number of the individuals before hunting for each year. The usual table (as the one for equilibrium, Table 9.2) used by "Vorovo" hunting area can be easily deduced.

From Table 9.6 it can be seen that we control the population mainly by controlling the lower diagonal part. We reach the equilibrium state by significantly reducing the population of newborns. This strategy has a great benefit—it gives the possibility of a good deer selection. When the optimal number of 1 year old deer is reached the second year (in this case 2016/2017) and successively for all the following years, it starts to propagate the equilibrium property, due to the equilibrium survival probabilities. From the other side, the upper diagonal part does not influence population dynamics (as the lower part does), so we leave high equilibrium probabilities. Leaving deers growing up in the upper diagonal part has a strong advantage—the number of candidates for the hunting trophies is high over the years.

In Fig. 9.3 we represent our model for the Game Management Plan for the period 2015–2025, by plotting the number of male deers and comparing it with their optimal number.

Table 9.6 State of the population before hunting for the period 2015–2025. The initial state (9.6) and equilibrium state represented in Table 9.1 are framed

Number of individuals before hunting	Sex	Age of individuals 0	1	2	3	4	5	6	7	8	9	10	Sum
2015/16	M	31	23	15	18	7	8	7	9	5	5	7	$\Sigma_0 = 219$
	F	31	24	17	14	15	8	10	6	12	9	0	
2016/17	M	28	15	15	12	18	7	8	6	8	4	0	$\Sigma_1 = 199$
	F	28	15	16	14	14	15	8	9	5	10	0	
2017/18	M	27	15	10	12	12	18	7	7	5	7	0	$\Sigma_2 = 193$
	F	27	15	10	13	14	14	15	7	8	4	0	
2018/19	M	27	15	10	8	12	12	18	6	6	4	0	$\Sigma_3 = 191$
	F	27	15	10	8	13	14	14	13	6	7	0	
2019/20	M	26	15	10	8	8	12	12	16	5	5	0	$\Sigma_4 = 187$
	F	26	15	10	8	8	13	14	12	11	5	0	
2020/21	M	26	15	10	8	8	8	12	11	14	4	0	$\Sigma_5 = 183$
	F	26	15	10	8	8	8	13	12	10	9	0	
2021/22	M	25	15	10	8	8	8	8	11	9	12	0	$\Sigma_6 = 175$
	F	25	15	10	8	8	8	8	11	10	8	0	
2022/23	M	23	15	10	8	8	8	8	7	9	8	0	$\Sigma_7 = 162$
	F	23	15	10	8	8	8	8	7	9	8	0	
2023/24	M	22	15	10	8	8	8	8	7	6	8	0	$\Sigma_8 = 156$
	F	22	15	10	8	8	8	8	7	6	8	0	
2024/25	M	21	15	10	8	8	8	8	7	6	5	0	$\Sigma_9 = 150$
	F	21	15	10	8	8	8	8	7	6	5	0	

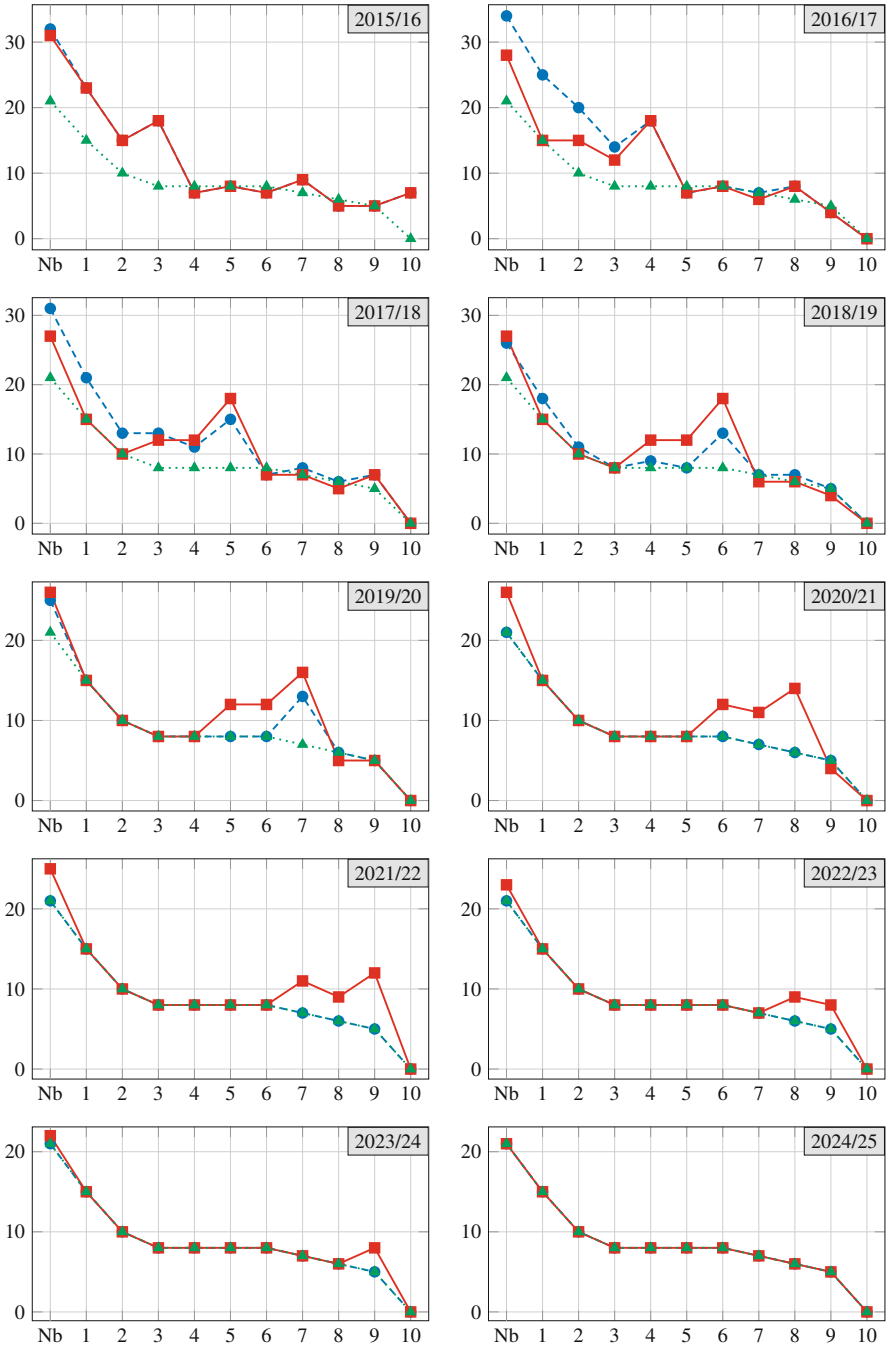

Fig. 9.3 The number of male deers by their age according to the new proposed Game Management Plan (—■—) versus the current one (- -●- -) and their optimal number (···▲···), over the hunting seasons 2015–2025. Nb stands for newborns

References

1. Bolton, M.: Conservation and the Use of Wildlife Resources. Chapman and Hall (1997)
2. Game Management Plans from the hunting area “Vorovo” in Serbia, for periods 2009–2019 and 2015–2025, private communication.
3. Annual Plans from the hunting area “Vorovo” in Serbia, for hunting seasons 2010/11, 2011/12, 2012/13, 2013/14 and 2014/15, private communication.

Chapter 10
Efficient Parameter-Dependent Simulation of Infections in a Population Model

Filippo Terragni

10.1 Introduction

Infectious diseases have been one of the major causes of mortality throughout history, which is still motivating researchers of different fields. Mathematical models can provide important insights into the evolution of epidemiological processes, the spread and course of infections, and the transmission dynamics in a host population [3]. Indeed, the obtained results are often promising with relation to comparing, planning, implementing and optimizing various programs of detection, prevention, therapy and control.

The propagation of infections in a host population depends on many factors, including the structure of the latter, the interactions among the organisms and the type of internal dynamics that is taken into account. In this article, a simple SIR model is considered, in which organisms are classified into susceptible, infectious and recovered. A time-dependent, seasonal transmission rate drives the epidemic and spatial heterogeneity is introduced by a diffusion process of organisms in the three groups. In addition, proliferation and death are modeled. There is a huge amount of works in the literature related to this framework. They can be either analytical or numerical-experimental studies and generally deal with different models depending on the assumptions or the aspect they are focused on (see [5, 7] and references therein). The considered deterministic model is not intended to be new but is regarded as a simple paradigm to which different numerical resolution approaches can be applied. Indeed, epidemiological models typically depend on a (quite large) set of parameters whose values influence the local spread of the infection. Thus, a numerical method (discretizing the model) that aims at simulating

F. Terragni (✉)
G. Millán Institute for Fluid Dynamics, Nanoscience and Industrial Mathematics, Department of Mathematics, Universidad Carlos III de Madrid, Leganés, Spain
e-mail: filippo.terragni@uc3m.es

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_10

the dynamics of the biological system for various key scenarios, especially for large time, tends to involve heavy computational resources. In this paper, an alternative approach is discussed, which consists in developing a reduced order model of the problem to improve the efficiency of a parameter-dependent study.

Reduced order models have been extensively used in various scientific fields and applications (see [2, 13, 16] and references therein), since they are able to reduce the complexity of a given problem while preserving a satisfactory accuracy, which makes them suitable to be applied for real-time simulations involving a huge number of degrees of freedom. In other words, these methods take advantage of the existing physical (or biological) laws to identify correlations among the system states and express the latter in terms of few essential features (modes), thus reducing the degrees of freedom of the problem and the computational resources of associated simulations. In the second part of the present work, a model reduction technique based on high-order singular value decomposition [6] is used to derive a low dimensional description of the epidemiological model on the basis of previously generated data for few values of the parameters. Then, a combination of this approach with some interpolation method will allow to approximate solutions for other, new values of the parameters in an efficient way, hence providing an effective mathematical tool for a fast simulation and analysis of the infection effects in the host population.

Thus, modeling has here a twofold goal. First, a mathematical description of the biological setting must be found. Secondly, a reduced order model for the identified equations is implemented to simplify the problem and its analysis. This work illustrates, by means of a simple paradigm, how some reduction techniques can be successfully used to perform a real-time control of a system response, which is a common task in science and engineering.

The remaining of the paper is organized as follows. The epidemiological model is deduced and described in Sect. 10.2, while the effects of some key parameters on the population dynamics are discussed in Sect. 10.3 via direct numerical simulation. Section 10.4 is devoted to the introduction of the high-order singular value decomposition and the obtained numerical results are shown in Sect. 10.5. Concluding remarks appear in Sect. 10.6.

10.2 The Epidemiological Model

A basic SIR model [1, 9] is constructed to describe the dynamics of a host population that is divided into three groups, namely the *susceptible* organisms (which can contract the disease), the *infectious* organisms (which are infected and can transmit the disease) and the *recovered* organisms (which healed). The population is sufficiently large and the disease is highly infective as to justify a continuous approximation. Thus, the three classes are represented by nonnegative densities $S(x,t)$, $I(x,t)$ and $R(x,t)$, respectively, which depend on the spatial location x and the time instant t. Indeed, the horizontal dimension of the population

habitat is assumed to be much larger than the other two, as to justify a one-dimensional model in space. Furthermore, all demographic changes balance so that the total population density

$$N = S(x,t) + I(x,t) + R(x,t) \tag{10.1}$$

is constant, namely $N \in \mathbb{R}^+$.

Basic Dynamics A SIR model is a 'two-step' process of first contracting the infection and then recovering from it. The first step is based on the assumption that organisms from the susceptible group get the disease from organisms of the infectious group at a *transmission rate* given by $\beta(t)SI$, where $\beta(t)$ is a time-dependent continuous function. The latter incorporates temporal external factors affecting the infection rate, namely the seasonality of some diseases (the dynamics of seasonally driven epidemics have been largely studied; see, e.g., [8, 20]). The second step assumes that organisms from the infectious group move to the recovered group at a rate proportional to a constant $\gamma > 0$. Thus, $1/\gamma$ is the average infectious period conditional on survival. Recovered organisms get permanent immunity against the infection, which is reasonable, e.g., for many childhood diseases whose spread can be studied via SIR models.

Proliferation and Death Only susceptible organisms are born into the host population, which occurs at rate $\mu_B N$, where μ_B is a positive constant. On the other hand, the death process in the three groups is driven by rate coefficients $\mu_S > 0$, $\mu_I > 0$ and $\mu_R > 0$, respectively. Thus, e.g., $1/\mu_S$ is the life expectancy of susceptible organisms. In order to maintain the total population density N constant, births are assumed to balance deaths by supposing that $\mu \equiv \mu_B = \mu_S = \mu_I = \mu_R$.

Mobility Organisms of the three groups move in the habitat according to the same diffusion process with coefficient $\nu > 0$ (this is a standard way to include spatial effects, see [15]). For simplicity, mobility of infectious organisms is supposed not to be affected by the disease.

Finally, demographic changes of a different type exhibit a longer characteristic timescale than the epidemic duration, the effect of any other structure in the host population (due to age, sex, variability of infectivity, ...) is neglected and no incubation or latency period is taken into account. Summarizing, all assumptions above provide a parameter-dependent model describing transmission, recovery, growth, decay and diffusion of the organisms in the three groups. Formulation of the corresponding equations is attained by the law of mass action and Fick's second law, which yields a reaction-diffusion system of three coupled PDEs given by

$$\frac{\partial S}{\partial t} = \nu \frac{\partial^2 S}{\partial x^2} - \beta(t)SI + \mu N - \mu S\,, \tag{10.2}$$

$$\frac{\partial I}{\partial t} = \nu \frac{\partial^2 I}{\partial x^2} + \beta(t)SI - \gamma I - \mu I\,, \tag{10.3}$$

$$\frac{\partial R}{\partial t} = \nu \frac{\partial^2 R}{\partial x^2} + \gamma I - \mu R\,, \tag{10.4}$$

for $0 < x < L$ and $t > 0$. Here, it is assumed that the population habitat is closed and bounded ($L \in \mathbb{R}^+$) and the system is isolated (no immigration or emigration), hence there is no flux of organisms across the boundaries of the domain, which is expressed by homogeneous Neumann boundary conditions for the three densities. The system (10.2)–(10.4) can be written in nondimensional form by scaling $S = NS^*$, $I = NI^*$, $R = NR^*$, $x = Lx^*$ and $t = \tau t^*$, being τ a characteristic timescale of the problem. Thus, after setting

$$\nu^* = \frac{\nu\,\tau}{L^2}\,, \qquad \beta^*(t) = \beta(t)\tau\,N\,, \qquad \mu^* = \mu\tau\,, \qquad \gamma^* = \gamma\,\tau\,, \tag{10.5}$$

and dropping asterisks, the dimensionless epidemiological model reads

$$\frac{\partial S}{\partial t} = \nu \frac{\partial^2 S}{\partial x^2} - \beta(t) SI + \mu(1 - S)\,, \tag{10.6}$$

$$\frac{\partial I}{\partial t} = \nu \frac{\partial^2 I}{\partial x^2} + \beta(t) SI - (\gamma + \mu)\, I\,, \tag{10.7}$$

$$\frac{\partial R}{\partial t} = \nu \frac{\partial^2 R}{\partial x^2} + \gamma I - \mu R\,, \tag{10.8}$$

for $0 < x < 1$ and $t > 0$, together with the boundary conditions

$$\frac{\partial S}{\partial x}(0,t) = \frac{\partial S}{\partial x}(1,t) = 0\,, \quad \frac{\partial I}{\partial x}(0,t) = \frac{\partial I}{\partial x}(1,t) = 0\,, \quad \frac{\partial R}{\partial x}(0,t) = \frac{\partial R}{\partial x}(1,t) = 0\,, \tag{10.9}$$

for $t > 0$. Note that Eq. (10.8) may be omitted and the recovered group density could be calculated from the scaled counterpart of the conservation law in Eq. (10.1), namely $R(x,t) = 1 - S(x,t) - I(x,t)$.

10.3 Effects of Key Parameters

In order to solve the epidemiological system constructed in Sect. 10.2, spatial derivatives are discretized by using second-order centered finite differences, while time discretization is performed by means of a second-order scheme based on the combination of Crank-Nicolson (for linear terms) and Adams-Bashforth (for nonlinear terms) methods [4]. The final discretized system will be referred to as the numerical solver. Note that, in all simulations below, values $\Delta x = 0.001$ and $\Delta t = 0.0005$ are selected for the spatial and time steps, respectively, in order to guarantee convergence and stability of the implemented scheme. Now, numerical experiments

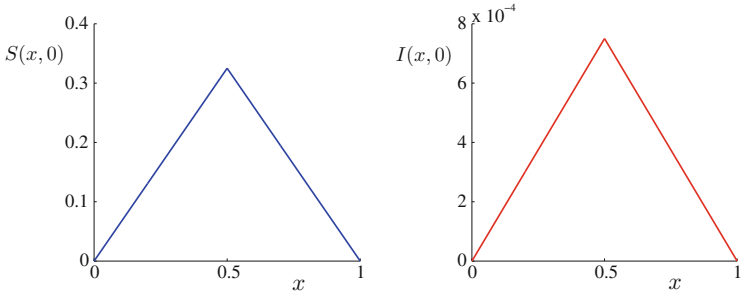

Fig. 10.1 Initial spatial distribution for the susceptible (left) and infectious (right) densities

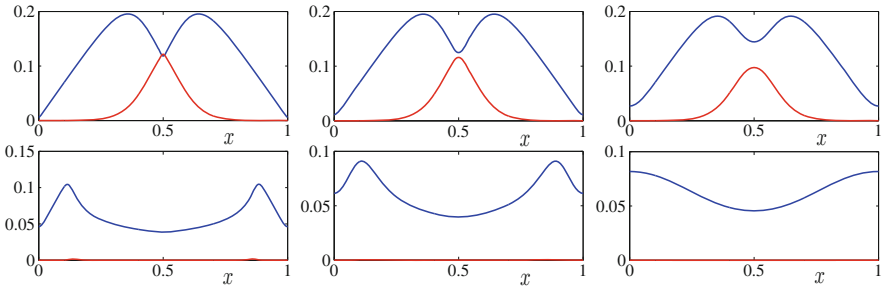

Fig. 10.2 Time evolution of $S(x,t)$ (blue) and $I(x,t)$ (red) for $t = 0.1$ (upper plots) and $t = 1$ (lower plots), when considering a diffusion coefficient equal to $\nu = 10^{-4}$ (left), $\nu = 10^{-3}$ (middle) and $\nu = 10^{-2}$ (right)

are conducted to examine the effects of several parameters on the population dynamics. The nonnegative initial densities for the susceptible and infectious groups are shown in Fig. 10.1, being $I(x,0)$ significantly smaller than $S(x,0)$. On the other hand, a sinusoidal transmission function $\beta(t) = 270\,(1 + \delta\,\sin(2\pi t))$ is considered, thus modeling the impact of periodic external factors on the time course of the infection (see, e.g., [20] and references therein). In other words, organisms are more likely to be infected during certain periods, which is reflected by enhanced interactions. Finally, some default values for the model parameters are set, namely $\nu = 0.001$, $\mu = 0.04$, $\gamma = 24$ and $\delta = 0.1$.

The first study illustrates how diffusion influences the evolution of the spatial distributions of S and I on the short timescale. Figure 10.2 highlights that small values of the diffusion coefficient ν produce nonhomogeneous densities, with sharp spatial regions of infection outbreaks (left plot). Indeed, if susceptible and infectious groups remain in local areas, the mechanism of spreading the disease may be less effective. As ν increases (middle and right plots), especially at late stages, diffusion tends to smooth out large spatial variations and densities become more homogeneous. In other words, the organisms mixing leads to global effects on the entire habitat.

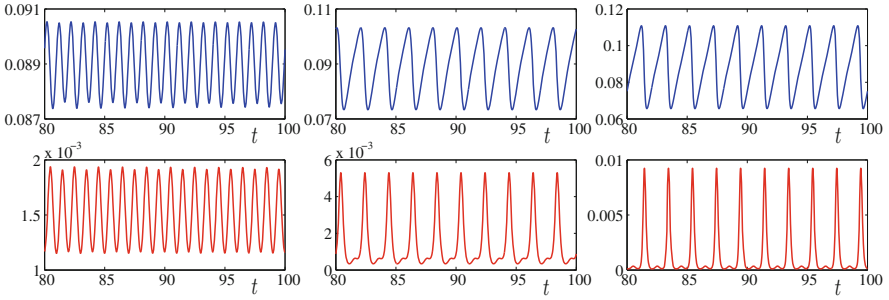

Fig. 10.3 Large-time dynamics of $S(x = 0.5, t)$ (upper plots) and $I(x = 0.5, t)$ (lower plots) over the interval $80 < t < 100$ for different values of the seasonality strength δ, namely $\delta = 0.05$ (left), $\delta = 0.10$ (middle) and $\delta = 0.25$ (right)

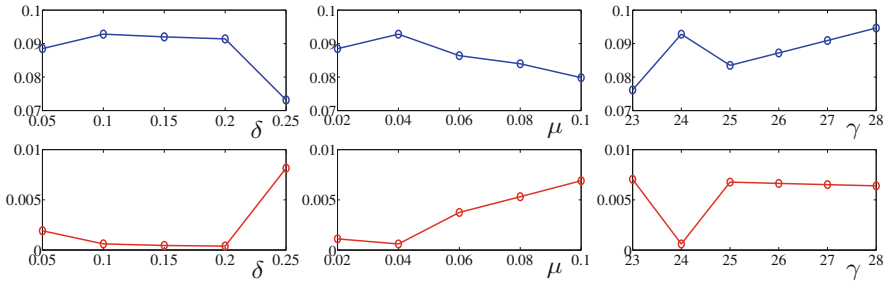

Fig. 10.4 Variability of $S(x = 0.5, t = 95.5)$ (upper plots) and $I(x = 0.5, t = 95.5)$ (lower plots) for different values of the model parameters δ (left), μ (middle) and γ (right)

The strength δ of the transmission function periodicity is another key parameter of the model that considerably affects the population dynamics. Even if its influence is clearly visible on a short timescale, more interesting effects appear if one performs a large-time numerical study, since then both densities S and I exhibit various transitions over a small range for δ. Figure 10.3 shows the asymptotic dynamics of $S(x = 0.5, t)$ (upper plots) and $I(x = 0.5, t)$ (lower plots) for three different values of δ, namely $\delta = 0.05$, 0.10 and 0.25, respectively. In the three cases, trajectories are clearly periodic of different periods. Note that a further increase of δ would lead to fairly complex temporal behaviors. In other words, the time course of the disease may largely change depending on the oscillation intensity of the external factors affecting the transmission rate. It is worth mentioning that the time span considered for this analysis has been calibrated as to capture the system attractors for all sets of parameter values taken into account in this work.

Similar simulations are run for different values of the parameters μ and γ. In Fig. 10.4, their effect (and that of δ) on $S(x = 0.5, t = 95.5)$ and $I(x = 0.5, t = 95.5)$ is shown. Observe that the fixed values for x and t are just an example to illustrate the outcomes after a long period of infection at a representative location within the habitat. Selected ranges evidence certain 'critical' values for each

parameter associated with a strong variability of both densities; some of them will be used to test the methodology described in the next section.

The results detailed so far show that an exhaustive analysis of the various effects due to the model parameters represents a nontrivial task. Moreover, the large CPU time required to simulate a comprehensive set of scenarios can make a numerical study somewhat hard. Indeed, integration of the proposed epidemiological model for each set of parameter values until the attractor (namely, for $0 < t < 100$) takes about 8 CPU minutes (all simulations are run on a desktop PC, with an Intel i5—3.1 GHz microprocessor and 4 GB RAM). Thus, an alternative technique (other than direct numerical simulation) can be useful in order to reduce the effort of the mentioned task.

10.4 HOSVD Combined with Interpolation

An alternative approach is described in this section, suitable for an efficient analysis of the population dynamics involving many, different values of the key parameters discussed above. Specifically, the *high-order singular value decomposition* (HOSVD), which is one possible extension to tensors of singular value decomposition [6, 19], is taken into account. Indeed, HOSVD is able to identify the most coherent features in a multidimensional dataset (i.e., a tensor) by looking for correlations among data in each direction, while eliminating redundancies. Thus, it models the relevant information in terms of few *modes*. Such low dimensional description can be applied to tensors containing values of the densities S and I for few values of some independent variables, chosen among time and the key parameters of the model. The final goal will be providing an accurate approximation for S and I corresponding to *new* values (i.e., not used to generate the initial datasets) of the independent variables, by using only the information stored in the original tensors. A proper combination of HOSVD with some interpolation technique will help to accomplish this objective. Some applications of HOSVD to different fields can be found, e.g., in [10–12, 14, 17].

For the sake of clarity, illustration of the model reduction technique is done for third-order tensors, being a generalization to higher dimension completely similar. For a third-order tensor $\mathrm{T} = (T_{ijk}) \in \mathbb{R}^{I \times J \times K}$, the HOSVD is a decomposition of the form

$$T_{ijk} = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \sum_{i_3=1}^{r_3} \sigma_{i_1 i_2 i_3} \, u_{i i_1} \, v_{j i_2} \, w_{k i_3} \,, \tag{10.10}$$

where the involved terms are defined as follows. Given the (symmetric, positive definite) 'correlation' matrices with elements

$$B^{(1)}_{lm} = \sum_{j,\,k} T_{ljk} \, T_{mjk} \,, \quad B^{(2)}_{lm} = \sum_{i,\,k} T_{ilk} \, T_{imk} \,, \quad B^{(3)}_{lm} = \sum_{i,\,j} T_{ijl} \, T_{ijm} \,, \tag{10.11}$$

then r_1, r_2 and r_3 in Eq. (10.10) are the ranks of $\mathrm{B}^{(1)}$, $\mathrm{B}^{(2)}$ and $\mathrm{B}^{(3)}$, respectively, while $u_{\cdot i_1}$, $v_{\cdot i_2}$ and $w_{\cdot i_3}$ are their respective orthonormal eigenvectors (*high-order modes*) associated with the positive eigenvalues. The square roots of the latter sorted in descending order, namely $\lambda^{(1)}_{i_1}$, $\lambda^{(2)}_{i_2}$ and $\lambda^{(3)}_{i_3}$, respectively, are called *high-order singular values*. Note that the high-order modes along the three tensor directions are the proper orthogonal decomposition modes of the associated fibers (i.e., the sets of vectors obtained by fixing two of the indexes of the tensor; see [14]). The elements of the *core tensor* are defined by

$$\sigma_{ijk} = \sum_{i_1=1}^{I} \sum_{i_2=1}^{J} \sum_{i_3=1}^{K} T_{i_1 i_2 i_3}\, u_{i_1 i}\; v_{i_2 j}\; w_{i_3 k}\,. \tag{10.12}$$

Now, a low dimensional approximation of the original tensor T can be obtained by retaining in the three directions only the first $s_1 \leq r_1$, $s_2 \leq r_2$ and $s_3 \leq r_3$ modes, which are associated with the largest singular values and thus account for the most relevant 'information' in the data, namely

$$T_{ijk} \;\approx\; \tilde{T}_{ijk} = \sum_{i_1=1}^{s_1} \sum_{i_2=1}^{s_2} \sum_{i_3=1}^{s_3} \sigma_{i_1 i_2 i_3}\, u_{i i_1}\, v_{j i_2}\, w_{k i_3}\,. \tag{10.13}$$

The root mean square (RMS) truncation error of this approximation can be (a priori) bounded as

$$\|\mathrm{T} - \tilde{\mathrm{T}}\|_{\mathrm{RMS}} \;\leq\; \sqrt{\frac{1}{IJK}\left(\sum_{i_1=s_1+1}^{r_1} \left(\lambda^{(1)}_{i_1}\right)^2 + \sum_{i_2=s_2+1}^{r_2} \left(\lambda^{(2)}_{i_2}\right)^2 + \sum_{i_3=s_3+1}^{r_3} \left(\lambda^{(3)}_{i_3}\right)^2 \right)}, \tag{10.14}$$

where $\|\,\mathrm{T}\,\|^2_{\mathrm{RMS}} = \sum_{i,j,k} T^2_{ijk}/(IJK)$. The error estimate in Eq. (10.14) is used to decide the number of modes to be retained by fixing a desired threshold. In practice, the way s_1, s_2 and s_3 are selected in truncated HOSVD in order to minimize the error bound defined in Eq. (10.14) can be adapted to the application at hand (see, e.g., [14]). Indeed, if the elements of the tensor show redundancies along all directions, then the high-order singular values decay fast and a good description of T can be achieved in terms of few modes.

Note that a third-order tensor may be handled by standard singular value decomposition by isolating one of the indexes but, in this way, only redundancies along that direction are taken into account. On the other hand, when applied to a multidimensional database, HOSVD considers information as a whole (namely, globally), extracting the most linearly independent 'features' along all directions. Thus, HOSVD is much more efficient in providing an accurate and *compressed* approximation of tensors. This is quantified by the factor $IJK/(s_1 s_2 s_3 + s_1 I + s_2 J + s_3 K)$, which yields the ratio between the dimension of the original third-

order tensor T and the number of data stored in its approximation $\tilde{\mathrm{T}}$ as given in Eq. (10.13).

In the context of this work, a tensor will contain values of either density S or I for few values of some independent variables, as mentioned above. Now, HOSVD works as a 'separation of variables' method: it provides a description by a set of modes for each direction separately (with a global effect). Hence, approximating the desired density at a new value of either time or a model parameter can be done by interpolating the high-order modes retained in the corresponding direction and using the interpolated values in Eq. (10.13), as shown in Sect. 10.5 (see [11, 14] for more advanced applications). In other words, integration of the epidemiological model by direct numerical simulation can be replaced by a few one-dimensional interpolations.

10.5 Numerical Results

The methodology outlined in Sect. 10.4 is first applied to study the influence of the diffusion coefficient ν at short time on the densities of susceptible and infectious groups. Two tensors are constructed, namely $S_{ijk} = S(x_i, t_j, \nu_k)$ and $I_{ijk} = I(x_i, t_j, \nu_k)$, by storing the values of S and I, respectively, at all mesh points used to discretize the problem as in Sect. 10.3 ($i = 1, \ldots, 1001$), for all possible combinations of few time instants in the interval $1 < t < 2$ ($j = 1, \ldots, 6$) and few values of ν in the range $10^{-4} \le \nu \le 10^{-2}$ ($k = 1, \ldots, 9$). Here, an almost uniform distribution of t_j and ν_k is taken into account. It is worth observing that a slightly different choice of these values would still provide good results, while optimizing their selection (according to the variability of the densities) would certainly improve the outcome of the technique. All data are computed by the numerical solver introduced in Sect. 10.3, with initial conditions as in Fig. 10.1 and default values for non-involved parameters.

Now, HOSVD is applied to both tensors, which are finally approximated as in Eq. (10.13) by means of s_1, s_2 and s_3 modes in the three directions (their values will be given below). Thus, high-order modes $u_{\cdot i_1}$ form a basis of a linear subspace for the spatial distributions of S or I, while the other two sets of modes account for the effects of t and ν. Relying on such information, the spatial structure of S or I can be approximated at new (i.e., not used to generate the original tensors) values of time and parameter ν, say $1 < t^* < 2$ and $10^{-4} \le \nu^* \le 10^{-2}$, by applying the formula

$$\mathrm{T}(x_i\,,\, t^*, \nu^*) \;=\; \sum_{i_1=1}^{s_1} \sum_{i_2=1}^{s_2} \sum_{i_3=1}^{s_3} \sigma_{i_1 i_2 i_3}\, u_{i i_1}\, v^*_{i_2}\, w^*_{i_3}\,. \tag{10.15}$$

Here T stands for either S or I, $v^*_{i_2}$ is the interpolated value at t^* of the i_2-th mode for time and $w^*_{i_3}$ is the interpolated value at ν^* of the i_3-th mode for parameter ν. In all simulations below, interpolation will be performed by cubic splines [18].

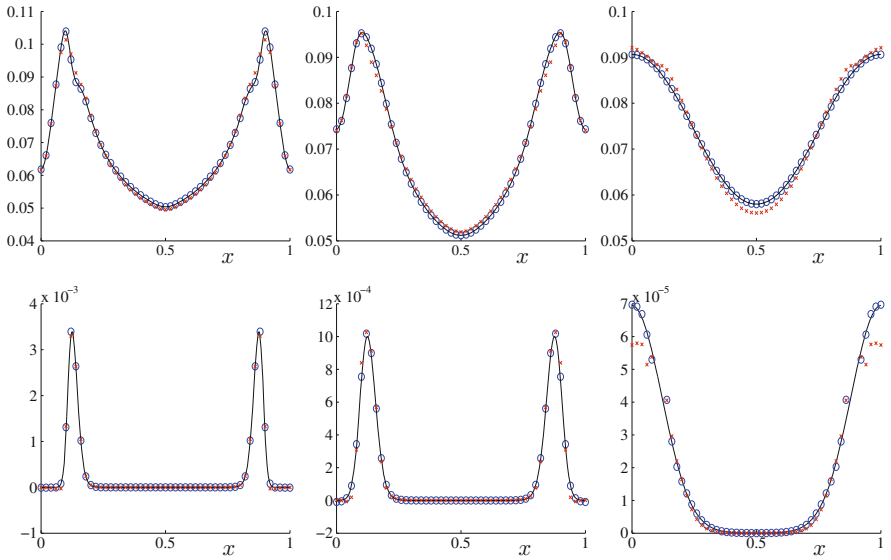

Fig. 10.5 Approximation via HOSVD plus interpolation of S (upper plots) and I (lower plots) at $t^* = 1.3$, and $\nu^* = 0.0002$ (left), $\nu^* = 0.0009$ (middle) and $\nu^* = 0.009$ (right). Approximated solutions with different sets of modes (circles and crosses; see text for details) are compared to those provided by direct numerical simulation (solid lines)

An example of the obtained results is shown in Fig. 10.5, where the spatial distribution of the susceptible and infectious densities (upper and lower plots, respectively) is approximated at the new time instant $t^* = 1.3$ and the new values of the diffusion coefficient $\nu^* = 0.0002, 0.0009$ and 0.009 (from left to right plots). In all cases, except for that corresponding to lower-right plot, only $s_1 = 10$, $s_2 = 5$ and $s_3 = 5$ modes in the three directions are retained, providing approximations (blue circles) versus the reference solutions computed via direct numerical simulation (solid lines) within relative RMS errors of 10^{-3} for S and $4 \cdot 10^{-2}$ for I in the worst case. Note that maximum errors are fairly small also. In addition, further reducing the number of modes for t and ν to $s_2 = s_3 = 2$ still yields an acceptable description of the dynamics in most test cases (red crosses). In the lower-right plot, slightly larger sets of modes must be used to get good results (namely, $s_1 = 15$, $s_2 = 6$ and $s_3 = 9$; see blue circles). Figure 10.6 illustrates a completely similar scenario for $t^* = 1.7$ and the same values of ν^* as in Fig. 10.5 (the numbers of involved modes are the same and errors behave as above). It is remarkable that most of the solutions used to construct the original tensors have a fairly different spatial structure in comparison with the approximated distributions for S and I in Figs. 10.5 and 10.6. In the presented results, s_1, s_2 and s_3 have been selected as to guarantee a reasonable level of accuracy. However, as the relevance of high-order modes for the reconstruction is associated with the magnitude of the corresponding singular values, which can have a different behavior in each direction, the values of s_1, s_2

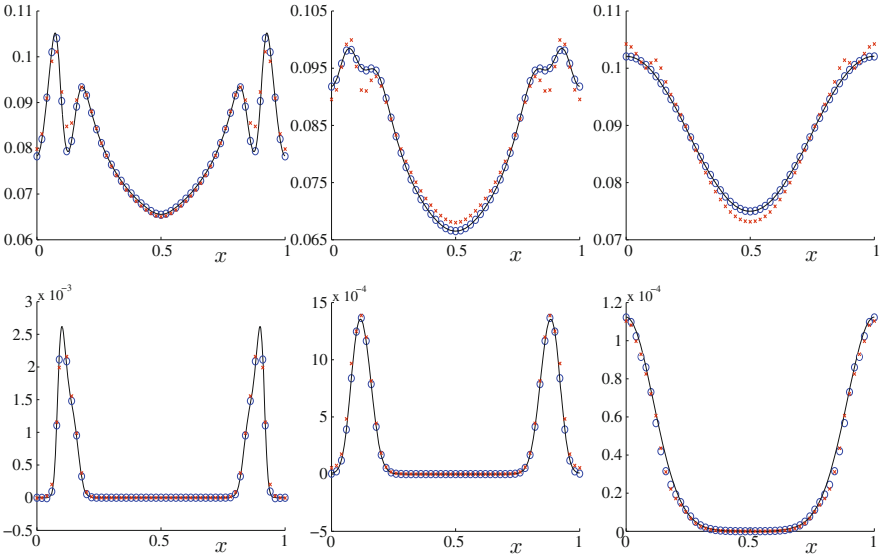

Fig. 10.6 Approximation via HOSVD plus interpolation of S (upper plots) and I (lower plots) at $t^* = 1.7$, and $\nu^* = 0.0002$ (left), $\nu^* = 0.0009$ (middle) and $\nu^* = 0.009$ (right). Approximated solutions with different sets of modes (circles and crosses; see text for details) are compared to those provided by direct numerical simulation (solid lines)

and s_3 could be optimized by means of the error estimate in Eq. (10.14). Finally, it is worth mentioning that the analysis discussed so far could be carried out for other model parameters also.

The next part of this section is intended to show the performance of HOSVD plus interpolation for the analysis of the population dynamics in the parameter space (δ, μ, γ) at large time, namely when an asymptotic state has been reached. For simplicity, so as to deal with third-order tensors, values $S(x = 0.5, t = 95.5)$ and $I(x = 0.5, t = 95.5)$ will be taken into account, with the goal of studying the effect of the epidemic at a representative point within the habitat at a late stage of the infection (different values of x and $t \gg 1$ would give similar results). Thus, two tensors $S_{ijk} = S(\delta_i, \mu_j, \gamma_k)$ and $I_{ijk} = I(\delta_i, \mu_j, \gamma_k)$ are constructed by storing the values of $S(x = 0.5, t = 95.5)$ and $I(x = 0.5, t = 95.5)$, respectively, for all possible combinations of few values of the involved parameters in the ranges $0.05 \leq \delta \leq 0.25$ $(i = 1, \ldots, 7)$, $0.02 \leq \mu \leq 0.1$ $(j = 1, \ldots, 5)$ and $23 \leq \gamma \leq 28$ $(k = 1, \ldots, 7)$. Taking advantage of the conclusions in Sect. 10.3, δ_i, μ_j and γ_k are concentrated where a strong effect on the variability of S and I is expected. Note that other approaches would be possible, for instance considering fifth-order tensors like $S_{ijklm} = S(x_i, t_j, \delta_k, \mu_l, \gamma_m)$ and looking for correlations in space, time and the three parameters. Indeed, for higher dimensions, the efficiency of HOSVD may be even better. All data are computed by the numerical solver introduced in Sect. 10.3, with initial conditions as in Fig. 10.1 and $\nu = 0.001$.

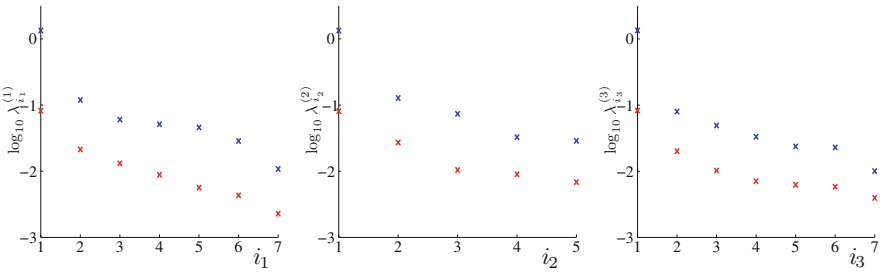

Fig. 10.7 Logarithm of the high-order singular values associated with tensors $S_{ijk} = S(\delta_i, \mu_j, \gamma_k)$ (blue) and $I_{ijk} = I(\delta_i, \mu_j, \gamma_k)$ (red) along the three directions (from left to right)

Table 10.1 Relative errors in the approximation of $S(x = 0.5, t = 95.5)$ and $I(x = 0.5, t = 95.5)$ at some representative points in the parameter space (δ, μ, γ) by means of HOSVD plus interpolation

δ^*	μ^*	γ^*	Relative error for S	Relative error for I
0.07	0.050	25.5	0.0007	0.0090
0.12	0.070	23.2	0.0114	0.0394
0.09	0.085	27.8	0.0006	0.0052
0.16	0.090	24.5	0.0010	0.0450
0.23	0.030	26.1	0.0623	0.0513

Tensors S_{ijk} and I_{ijk} are decomposed as in Eq. (10.13) by retaining s_1 modes $u_{\cdot i_1}$ for the strength of the transmission function periodicity, s_2 modes $v_{\cdot i_2}$ for the birth-death coefficient and s_3 modes $w_{\cdot i_3}$ for the recovery coefficient. The reconstruction step is then applied at new values of the involved parameters, say $0.05 \leq \delta^* \leq 0.25$, $0.02 \leq \mu^* \leq 0.1$ and $23 \leq \gamma^* \leq 28$, by means of a proper counterpart of Eq. (10.15). Figure 10.7 depicts the high-order singular values $\lambda^{(1)}_{i_1}$, $\lambda^{(2)}_{i_2}$ and $\lambda^{(3)}_{i_3}$ associated with both tensors, showing a qualitative similar behavior between susceptible and infectious densities. On the other hand, Table 10.1 stores the relative errors (versus the reference solutions computed via direct numerical simulation) in approximating $S(x = 0.5, t = 95.5)$ and $I(x = 0.5, t = 95.5)$ at some representative, new points in the parameter space (δ, μ, γ). For most test cases, they maintain around 10^{-3} and $4 \cdot 10^{-2}$ for S and I, respectively, in spite of the small number of parameter values used to generate the tensors (which convey a limited information on the system) and, consequently, the few available high-order modes. Regarding the number of the latter, $s_2 = 5$ and $s_3 = 6$ have been fixed for S in all simulations, while $s_2 = 4$ and $s_3 = 7$ for I. On the contrary, different values of s_1 have been chosen depending on the case in order to guarantee a good approximation. Indeed, the population dynamics show a high sensibility to parameter δ, especially for $\delta > 0.1$, as already discussed in Sect. 10.3. As a consequence, existing transitions are not fully captured by the few values considered in the original tensors, which also implies that some of the available modes in the

first direction are poorly interpolated. This is reflected in a worse approximation in test cases for $\delta > 0.1$ (see Table 10.1). On the other hand, note that a better *sampling* of the parameter space would lead to richer subspaces in the three directions and a more accurate description of the solutions. However, in order to perform this task, some additional knowledge about critical bifurcations of the system (requiring essential data to be taken into account) would be necessary. In this case, a clearer dependence of the errors on the number of retained modes would also be evident, which could help to construct a more efficient approximation strategy. All these issues are regarded as an extension of the discussed ideas and therefore beyond the scope of this article.

The computational efficiency of the proposed model reduction technique is fairly high. Indeed, 100 large-time direct numerical simulations of the model would require more than 13 CPU hours, while HOSVD plus interpolation (once tensors are available) would need only very few CPU seconds for the approximation of all desired solutions.

10.6 Final Remarks

A SIR epidemiological model including diffusion of organisms, seasonal transmission, birth and death for the description of the dynamics of an infected population has been discussed. Numerical simulations of the associated reaction-diffusion system have been performed to study various scenarios for different ranges of parameter values. Although the implemented numerical solver may provide interesting insights into the dynamical behavior of the population groups, a systematic multiparameter analysis embracing a large number of test cases can be time consuming.

Thus, the model has been exploited as a simple paradigm to show that less standard numerical approaches to the problem are possible. Specifically, a reduction technique relying on HOSVD combined with interpolation has been applied to decrease the number of degrees of freedom of the discretized epidemiological system and consequently reduce the required computational effort. The proposed procedure led to quite good results, in terms of both accuracy and efficiency, allowing to satisfactorily approximate the population behavior under the effect of generic values of some key parameters. The approach turned out to be much faster than direct numerical simulation, especially for the analysis of large-time dynamics.

This work could be obviously improved and extended. For instance, a more careful selection of the parameter values generating the initial tensors would enhance the ‘quality’ (i.e., the conveyed information) of the modes, decrease their number, and reduce the involved truncation and interpolation errors. On the other hand, an extension to wider ranges of parameters or datasets with more than three dimensions would allow to perform a more complete parameter-dependent study of the epidemic with high efficiency, thus hopefully supporting in some way the design and assessment of vaccination strategies.

Acknowledgments The author would like to thank the students of his group at the ECMI Modelling Week 2013, held at Universidad Carlos III of Madrid on July 2013, for their contribution to this work. The latter has been supported by the Ministerio de Economía y Competitividad grant MTM2014-56948-C2-2-P and by the FEDER / Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación grant MTM2017-84446-C2-2-R.

References

1. Anderson, R. M., May, R. M.: Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford (1991)
2. Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**, 483–531 (2015)
3. Brauer, F., Castillo-Chavez, C.: Mathematical Models in Population Biology and Epidemiology. Springer-Verlag, New York (2001)
4. Cebeci, T.: Convective Heat Transfer. Springer-Verlag, Berlin & Heidelberg (2002)
5. Chowell, G., Sattenspiel, L., Bansal, S., Viboud, C.: Mathematical models to characterize early epidemic growth: a review. Phys. Life Rev. **18**, 66–97 (2016)
6. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. A. **21**, 1253–1278 (2000)
7. Funk, S., Salathé, M., Jansen, V. A. A.: Modelling the influence of human behaviour on the spread of infectious diseases: a review. J. R. Soc. Interface **7**, 1247–1256 (2010)
8. Keeling, M. J., Rohani, P., Grenfell, B. T.: Seasonally forced disease dynamics explored as switching between attractors. Physica D **148**, 317–335 (2001)
9. Kermack, W. O., McKendrick, A. G.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A **115**, 700–721 (1927)
10. Kreimer, N., Sacchi, M. D.: A tensor higher-order singular value decomposition for prestack seismic data noise reduction and interpolation. Geophysics **77**, 113–122 (2012)
11. Lorente, L. S., Vega, J. M., Velazquez, A.: Generation of aerodynamic databases using high-order singular value decomposition. J. Aircraft **45**, 1779–1788 (2008)
12. Lorente, L. S., Vega, J. M., Velazquez, A.: Compression of aerodynamic databases using high-order singular value decomposition. Aerosp. Sci. Technol. **14**, 168–177 (2010)
13. Lucia, D. J., Beran, P. S., Silva, W. A.: Reduced-order modeling: new approaches for computational physics. Prog. Aerosp. Sci. **40**, 51–117 (2004)
14. Moreno, A. I., Jarzabek, A. A., Perales, J. M., Vega, J. M.: Aerodynamic database reconstruction via gappy high order singular value decomposition. Aerosp. Sci. Technol. **52**, 115–128 (2016)
15. Murray, J. D.: Mathematical Biology. Springer-Verlag, New York (2002)
16. Quarteroni, A., Rozza, G. (eds.): Reduced Order Methods for Modeling and Computational Reduction. Springer International Publishing (2014)
17. Rajwade, A., Rangarajan, A., Banerjee, A.: Image denoising using the higher order singular value decomposition. IEEE T. Pattern Anal. **35**, 849–862 (2013)
18. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. Springer-Verlag, New York (2002)
19. Tucker, L. R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**, 279–311 (1966)
20. Uziel, A., Stone, L.: Determinants of periodicity in seasonally driven epidemics. J. Theor. Biol. **305**, 88–95 (2012)

Chapter 11
Optimising a Cascade of Hydro-Electric Power Stations

Marta Pascoal

11.1 Introduction

Hydro electricity is electricity produced from hydropower and is responsible for a good share of the world's total generated electricity. Most of this power comes from water stored in dams, usually coming from natural resources like rivers, rain or snow melts, which when released flows through a turbine activating a generator that produces electricity. The energy then produced depends on the volume of water that is released and the difference in height between the water starting and ending points. At times of less rain, or simply at high peak demands. there may be a shortage of water to turbine in the reservoirs, while electric power is still needed. To cope with such situations some power plants are capable of pumping water to higher reservoirs, which can be done when there is not enough water to be released when needed [2, 4–6].

A cascade system of hydro-electric power stations is a set of stations connected as in a network where water flows between some of them. Two examples are shown in Fig. 11.1. The triangles and circles in the plots represent the hydro station reservoirs and turbines, respectively. The straight lines between the power stations show the connection between them, whereas the blue arrows attached to each circle define in which direction(s) each turbine is able to pump water.

The purpose of this work is to model the operation of a branched cascade system like that depicted in Fig. 11.1b along 1 day, aiming at planning when each power station should release water downstream or pump it upstream. In this case, the turbines installed on hydro stations 3 and 4 have the ability of pumping water in

M. Pascoal (✉)
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal

Institute for Systems Engineering and Computers – Coimbra, Coimbra, Portugal
e-mail: marta@mat.uc.pt

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_11

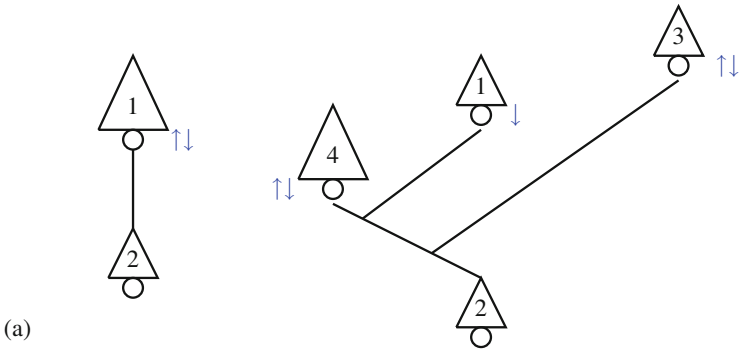

Fig. 11.1 Cascades with hydro-electric power stations. (**a**) Two hydro-electric power plants. (**b**) Four hydro-electric power plants

both directions, that is, both from hydro stations 3 or 4 downstream to hydro station 2, as well as from hydro station 2 upstream to hydro stations 3 or 4. Such daily plan is decided based on a forecast for the energy market prices and with the goal of maximising the daily profit. The problem is modelled as a nonlinear optimisation problem, which can be solved using a mathematical programming environment like AMPL [3] or Matlab.

11.2 Problem Formulation

The problem of optimising the branched cascade of hydro electric power plants in Fig. 11.1b aims at planning the daily water flow in the cascade, with the goal of maximising the profit related with the electric power generation. This value depends on several features of the system, like the power that is consumed when the water is pumped upstream, the power that is generated by the hydro stations when water is released downstream, and, last but not least, on the energy market price oscillations. The main characteristics of that system are described in the following. To simplify we begin by considering the system with only two hydro stations, depicted in Fig. 11.1a.

11.2.1 Two Power Plants Cascade Model

We assume the water flow plan for the power plants is defined hourly for 1 day and first consider the simple cascade in Fig. 11.1a. Two sets of indices are used

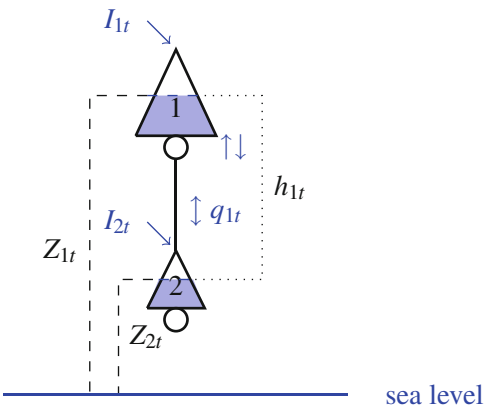

Fig. 11.2 Cascade with two hydro-electric power plants

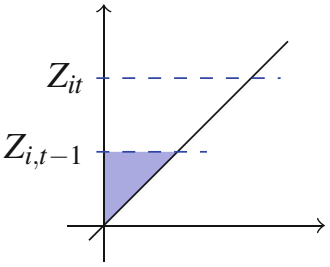

Fig. 11.3 Water reservoir

in the following, $I = \{1, 2\}$, which represents the set of power plants, and $T = \{1, 2, \ldots, 24\}$, associated with the hours of the day.

Water Level, Water Head and Water Volume

In order to characterise the system it is important to define the water level in each reservoir i with respect to the sea level, ξ, at instant t, denoted by Z_{it}, for $i \in I$, $t \in T$, and depicted in Fig. 11.2. The difference between the water levels of two reservoirs is related with the power that is produced by releasing water from one reservoir to the next. These amounts are called the water head of reservoir i at moment t, and are denoted by h_{it} and defined as

$$\begin{aligned} h_{1t} &= Z_{1t} - Z_{2t}, \quad t \in T \\ h_{2t} &= Z_{2t} - \xi, \quad t \in T \end{aligned} \tag{11.1}$$

Additionally, the water levels vary according to the water volume in the reservoir, denoted by V_{it}, for any $i \in I$ and $t \in T$. Assuming that these quantities are known and that the reservoir has approximately the shape of a cone as in Fig. 11.3, $Z_{it} - Z_{i,t-1}$ can be estimated as the volume of a solid of revolution depending on a constant r related with the width of the reservoir,

$$V_{it} - V_{i,t-1} = \int_{Z_{i,t-1}}^{Z_{it}} \pi\, r^2 y^2\, dy = \frac{\pi\, r^2}{3} \left(Z_{it}^3 - Z_{i,t-1}^3 \right),$$

and, therefore,

$$Z_{it} = \sqrt[3]{Z_{i,t-1}^3 + \frac{3}{\pi r^2}(V_{it} - V_{i,t-1})}.$$

In practice, and to simplify the calculations, the water levels are usually updated as

$$Z_{it} = Z_i^0 + \alpha_i (V_{it} - V_i^0)^{\beta_i}, \quad i \in I,\ t \in T \tag{11.2}$$

with α_i, β_i given parameters, dependent on the shape and the characteristics of the reservoir, and Z_i^0 the nominal water level in the reservoir, $i \in I$. Finally, the values V_i^0 are given constants representing the nominal water volumes that correspond to Z_i^0, for any $i \in I$ [7].

Limits are imposed to the minimum and the maximum amount of water that can be stored in each reservoir, either by using constraints over the volume of water, or over the level of water, in each of them. The constraints

$$Z_i^{\min} \leq Z_{it} \leq Z_i^{\max}, \quad i \in I, t \in T, \tag{11.3}$$

model the latter situation, for $Z_i^{\min}$, $Z_i^{\max}$ given constants, $i \in I$.

The Water Flow Rate

The volume of water in a reservoir $i \in I$ is usually affected by inflows from natural resources, like rain, which are assumed to be estimated as I_{it}, water from incoming discharges on upstream reservoirs, q_{jt}, as well as water releases from the reservoir i itself to other downstream reservoirs, q_{it}, $i, j \in I, t \in T$, as illustrated in Fig. 11.2. Thus,

$$\begin{aligned} V_{1t} &= V_{1,t-1} + I_{1t} - q_{1t}, \quad t \in T \\ V_{2t} &= V_{2,t-1} + I_{2t} + q_{1t} - q_{2t}, \quad t \in T \end{aligned} \tag{11.4}$$

are the flow conservation constraints needed to model the amount of water stored at any moment at reservoirs 1 and 2, respectively.

It is assumed that the flow rate q_{it} is positive when the water is being pumped downstream, and negative if the water is being pumped upstream, $i \in I$, $t \in T$. Constraints that limit the flow rate at each reservoir are also necessary. In case of reservoirs that only pump water downstream (turbine) the bounds are

$$0 \leq q_{it} \leq q_i^0 \sqrt{\frac{h_{it}}{h_i^0}}, \quad i \in I,\ t \in T \tag{11.5}$$

and for the remaining reservoirs

$$\zeta_i \left(h_{it} - h_i^0\right) - q_i^0 \leq q_{it} \leq q_i^0 \sqrt{\frac{h_{it}}{h_i^0}}, \quad i \in I,\ t \in T \tag{11.6}$$

Here, q_i^0 represents the nominal amount of turbined water in the reservoir i in the first case or the amount of nominal pumped water in the reservoir i in the second, ζ_i is the pumping coefficient of the reservoir i, and h_i^0 is the nominal head of the reservoir i, $i \in I$ [7].

Power and Revenue

The goal of the problem is to find a distribution of the times for each hydro plant to pump water downstream (called turbining) or to pump water upstream (called pumping up) along the day, in order to maximise the profit resulting from the produced power. The hourly prices of energy are denoted by P_t and are assumed to be known, for $t \in T$. These values need to be combined with the electrical power that is produced and consumed by the power plants, which differs when water is only pumped downstream or when it can both be pumped downstream, thus producing power, and upstream, consuming it.

The power produced by the turbine of a hydroelectric station depends on the water flow, the height of the plant, the gravity acceleration, 9.8, and the equipment characteristics. A simple formula to model this quantity is

$$9.8 q_{it} \mu_i h_{it},$$

where μ_i is a parameter specific to turbine $i \in I$ that represents its efficiency in electricity production mode. In a more accurate model this value is also affected by an internal consumption factor ϕ_i, which limits the net plant power output, as well as makes the power output grow slower as the water flow grows. The new model defines the power produced when turbining as

$$9.8 q_{it} (h_{it} - \Delta h_{it}) \mu_i (1 - \phi_i),$$

where

$$\Delta h_{it} = \Delta h_i^0 \left(\frac{q_{it}}{q_i^0} \right)^2$$

represents friction losses when turbining or pumping, and Δh_i^0 and q_i^0 are nominal values, $i \in I$.

At a given moment, each power plant either turbines water producing revenue, pumps it upstream with a certain cost in the short term, or the system is idle and there is zero flow. The formulae for the value of the power output and the price for pumping are combined as follows:

$$r_{it} = \begin{cases} 9.8 q_{it} (h_{it} - \Delta h_{it}^T) \mu_i^T (1 - \phi_i) & \text{if } q_{it} \geq 0 \\ 9.8 q_{it} (h_{it} + \Delta h_{it}^P) \frac{1}{\mu_i^P (1 - \phi_i)} & \text{if } q_{it} < 0 \end{cases}, \quad i \in I,\ t \in T \tag{11.7}$$

where

$$\Delta h_{it} = \Delta h_i^0 \left(\frac{q_{it}}{q_i^0} \right)^2$$

represents friction losses when turbining (T) or pumping (P); both values expressed as a head loss. The nominal values Δh_i^0 and q_i^0 are constants specific to each turbine; the parameters μ_i^T and $1/\mu_i^P$ represent efficiencies of turbines in electricity production mode and pumping mode, respectively. The objective function is then given by

$$\sum_{t \in T} P_t \sum_{i \in I} r_{it}, \tag{11.8}$$

and the full formulation of the optimisation problem associated with the cascade depicted in Fig. 11.1a is

$$\begin{array}{lll}
\text{maximise} & \displaystyle\sum_{t \in T} P_t \sum_{i \in I} r_{it} & \\
\text{subject to} & h_{1t} = Z_{1t} - Z_{2t}, & t \in T \\
& h_{2t} = Z_{2t} - \xi, & t \in T \\
& Z_{it} = Z_i^0 + \alpha_i (V_{it} - V_i^0)^{\beta_i}, & i \in I, \quad t \in T \\
& V_{1t} = V_{1,t-1} + I_{1t} - q_{1t}, & t \in T \\
& V_{2t} = V_{2,t-1} + I_{2t} + q_{1t} - q_{2t}, & t \in T \\
& Z_i^{\min} \le Z_{it} \le Z_i^{\max}, & i \in I, \quad t \in T \\
& 0 \le q_{2t} \le q_2^0 \sqrt{\frac{h_{2t}}{h_2^0}}, & t \in T \\
& \zeta_1 \left(h_{1t} - h_1^0 \right) - q_1^0 \le q_{1t} \le q_1^0 \sqrt{\frac{h_{1t}}{h_1^0}}, & t \in T
\end{array} \tag{11.9}$$

11.2.2 Four Power Plants Cascade Model

The case of the four power plants cascade depicted in Fig. 11.1b can be seen as an extension of the previous one. In the following $I = \{1, 2, 3, 4\}$ stands for the set of four hydro stations.

According to the presented scheme, in this case two turbines are able to work both in upstream and in downstream pumping modes, namely those located at the hydro electric plants 3 and 4. Thus, the previous flow conservation constraints are replaced by the conditions

$$\begin{array}{l}
V_{2t} = V_{2,t-1} + I_{2t} + q_{1t} + q_{3t} + q_{4t} - q_{2t}, \quad t \in T \\
V_{it} = V_{i,t-1} + I_{it} - q_{it}, \quad i \in I - \{2\}, \quad t \in T
\end{array} \tag{11.10}$$

The formulation of the new optimisation model is as follows

$$
\begin{aligned}
\text{maximise } & \sum_{t\in T} P_t \sum_{i\in I} r_{it} \\
\text{subject to } & h_{it} = Z_{it} - Z_{2t}, && i \in I - \{2\}, \quad t \in T \\
& h_{2t} = Z_{2t} - \xi, && t \in T \\
& Z_{it} = Z_i^0 + \alpha_i (V_{it} - V_i^0)^{\beta_i} && i \in I, \quad t \in T \\
& V_{it} = V_{i,t-1} + I_{it} - q_{it}, && t \in T \\
& V_{2t} = V_{2,t-1} + I_{2t} + q_{1t} + q_{3t} + q_{4t} - q_{2t}, && t \in T \\
& Z_i^{\min} \le Z_{it} \le Z_i^{\max} && i \in I, \quad t \in T \\
& 0 \le q_{it} \le q_i^0 \sqrt{\frac{h_{it}}{h_i^0}} && i = 1, 2, \quad t \in T \\
& \zeta_i \left(h_{it} - h_i^0\right) - q_i^0 \le q_{it} \le q_i^0 \sqrt{\frac{h_{it}}{h_i^0}} && i = 3, 4, \quad t \in T
\end{aligned}
\tag{11.11}
$$

It is noted that the variables q_{it} are decision variables whereas V_{it}, Z_{it} and h_{it} depend somehow on the flow rates q_{it}, $i \in I, t \in T$. Like the formulation presented in the previous subsection, the problem that we want to solve (11.11), is a nonlinear optimisation problem. In fact, both the objective function in (11.8) and the sets of constraints (11.2), (11.5) and (11.6) are nonlinear.

11.3 Numerical Results

In [1] results of the implementation of the non-linear program (11.11) using the modelling language AMPL [3] are reported. Two cases were considered:

1. one where all the reservoirs started virtually empty, and
2. the other where all but the reservoir number 2 were empty and this one was almost full.

The input data of the model was provided by the company *REN - Redes Energéticas, S.A.*

The solution obtained for the first case had a small profit. Additionally, the increase/decrease in the flow rates followed the fluctuation in the energy prices, and pumping upstream appears in the optimal solution at times when the price is low, alternated by occasional pumping downstream when the price decreases. Usually the highest of the hydro plants is chosen as the sink of water pumped upstream.

The profit of the system was bigger in the second case and the optimal solution consisted mainly in pumping water downstream at maximum flow rate, as expected if no shortage of water occurs.

11.4 Concluding Remarks

The daily planning of a branched cascade of hydro power plants arranged as in Fig. 11.1b was modelled as a non-linear program. As concluding remarks it should be noted that it would be interesting in practice to extend the planning horizon to more than 1 day, and possibly include weekly patterns or seasonal characteristics. However, this will increase significantly the size of the problem. Also, this problem is associated with several natural phenomenon that are typically uncertain and, therefore, it would be most useful, and challenging, to handle it from a stochastic point of view.

Acknowledgments This work was partially supported by the Centre for Mathematics of the University of Coimbra—UID/MAT/00324/2019 and by the Institute for Systems Engineering and Computers—Coimbra—UID/MULTI/00308/2019, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

1. A. Ajne, B. Byukusenge, M. Cavalleri, D. Pettersson, M. Salvador Svaholm, *Optimising a complex hydroelectric cascade in an electricity market*, Report of the 26-th European Consortium for Mathematics in Industry Modelling Week, Technical University of Dresden, August 2012 (www.math.tu-dresden.de/essim2012/pdf/Project1.pdf accessed on April 15 2019)
2. M. Ferreira, A. Ribeiro, G. Smirnov, *Local minima of quadratic functionals and control of hydro-electric power stations*, Journal of Optimization Theory and Applications **165**:985–1005, 2015
3. R. Fourer, D. Gay, B. Kernighan. *A modeling language for mathematical programming*, Management Science, **36**:519–554, 1990
4. M. Gardini, A. Manicardi, *Hydropower optimization: an industrial approach*, arXiv preprint **1610.10057**, 2016
5. A. Korobeinikov, A. Kovacec, M. McGuinness, M. Pascoal, A. Pereira, S. Vilela, *Optimizing the profit from a complex cascade of hydroelectric stations with recirculating water*, MICS Journal **2**:111–133, 2010
6. A. Ribeiro, M. Guedes, G. Smirnov, S. Vilela, *On the optimal control of a cascade of hydro-electric power stations*, Electric Power Systems Research **88**:121–129, 2012
7. International Atomic Energy Agency, *Valoragua: a model for the optimal operating strategy of mixed hydrothermal generating systems*, Users' manual for the mainframe computer version. N. 4. (IAEA computer manual series, Vienna, 1992)

Chapter 12
Networks of Antennas: Power Optimization

Stéphane Labbé

12.1 Introduction of the Problem

In this text, we illustrate the process leading from a physical problem to an effective simulation. This process will display to three types of models. The first one, the most simple, will provide the opportunity to familiarise with the model and the existence of solutions to the problem, a second, more complex, will illustrate the necessity of numeric computations and at last we will give a complete formulation. The example we chose is the optimisation of a network of antennas. For the sake of simplification, the antennas taken into consideration will be assimilated to discrete dipolar systems. The goal of this study is to give an algorithm for antennas placement and power regulation. In a first part, we will focus on the modelling of the problem. We will start by setting the problem and choose the notations, and will, then, focus on the modelling of an antenna, explaining the link between the electromagnetic equations and a dipolar antenna. In the second, third and fourth parts, we will treat the three optimisation problems.

12.2 A Network of Antennas: Modelling

The first work to perform in order to model a situation, is to set the problem in mathematical terms. The built model will enable the study and optimisation of the situation parameters. This first milestone of the modelling and simulation process is very important and must be carefully treated. The key of the modelling process

S. Labbé (✉)
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP (Institute of Engineering Univ. Grenoble Alpes), Grenoble, France
e-mail: stephane.labbe@univ-grenoble-alpes.fr

E. Lindner et al. (eds.), *Mathematical Modelling in Real Life Problems*,
Mathematics in Industry 33, https://doi.org/10.1007/978-3-030-50388-8_12

is to answer the question: "what do you want to do?". This question is, not only about what we are modelling but also about what we want to do with this model. Is this work will be exploited in order to forecast the behaviour of a system, to understand and enhance a modelling or to compute specific parameters? In our case, the objective is clear: how to optimise the topology of an antenna network in order to provide a given signal strength on a given territory. The accuracy of the physical hypothesis is not required, then we can simplify the model with the assumption that antennas are dipoles and the signal is not harmonic in time. These antennas, in finite number, are set on a collection of points in space. The set of antennas will induce a resultant power in the whole space, the question we want to tackle here is how to optimise the number of antennas, their position and power to ensure that, in a given part of the space, the resultant power would stay between a minimum and a maximum. The questioning is quite common even if simplified here: how to optimise a network to certify that the power of the signal is sufficient to ensure its good functioning and sufficiently small to guarantee the safety of the system for users?

12.2.1 The Global Problem: Setting of a Mathematical Model

We define Σ as set of elements of $\mathbb{R}^3$, the locations of the dipoles. This set indexes the dipoles; the set of dipoles is $M = (\mu_x)_{x\in\Sigma}$, subset of $S^2 = \{u \in \mathbb{R}^3 || u| = 1\}$, and the set of powers $P = (p_x)_{x\in\Sigma}$, subset of $\mathbb{R}$. Moreover, we set Ω a subset of $\mathbb{R}^3$, the location where measures have to be performed and $(\underline{m}, \overline{m}) \in \mathbb{R}^2$, respectively the minimum and maximum required power. Given f, from $\mathbb{R}^3 \times S^2 \times \mathbb{R}^3$ into $\mathbb{R}^3$, the function evaluating the power emitted by an antenna of power 1 in a given place of the space. Hence, for the whole space, except a small ball around the antenna position, we set:

$$\forall x \in \mathbb{R}^3 \setminus \bigcup_{y\in\Sigma} B(y, \varepsilon), \text{ for every } \varepsilon \in \mathbb{R}_*^+,$$

$$F_\varepsilon(\Sigma, M, P)(x) = \left| \sum_{y\in\Sigma} p_y \ f(y, \mu_y, x) \right|,$$

where $F_\varepsilon(\Sigma, M, P)(x)$ is the power developed by the network on a given point x. The problem now is to optimise the power P, the directions of dipoles M and the locations Σ, of dipolar antennas, in order to ensure on optimal power in $\Omega_\varepsilon = \Omega \setminus \bigcup_{y\in\Sigma} B(y, \varepsilon)$, it is to say a power such that, locally, $\underline{m} \leq F_\varepsilon(\Sigma, M, P)(x) \leq \overline{m}$.

To tackle this problem, we define the set of admissible solutions:

$$\mathscr{A}_\varepsilon = \left\{ (\Sigma, M, P) \subset \mathbb{R}^3 \times S^2 \times \mathbb{R} | \forall x \in \Omega_\varepsilon, \ \underline{m} \leq F_\varepsilon(\Sigma, M, P)(x) \leq \overline{m} \right\}.$$

Here, the question is: how to find the best triplet (Σ, M, P) and, what does means best?

First, we would be tempted to redefine, more exactly to restraint, the problem by freezing one or more parameters. For X subset of (Σ, M, P), we set $\mathscr{A}_\varepsilon(X)$ the function $\mathscr{A}_\varepsilon$ where the values X have been fixed. Hence, the problem to solve is: find (Σ, M, P) in $\mathscr{A}_\varepsilon$ such that

$$\sum_{x \in P} p_x \text{ is minimal.} \tag{12.1}$$

This condition is motivated, in terms of modelling, by the goal to minimise the required energy needed to obtain an admissible network.

As we see, the problem is complex and several bottlenecks will be encountered

- Does solutions exist?
- If a solution exists, is this solution unique?
- Can we compute explicitly this solution?

12.2.2 Power of an Antenna

In this section, we will focus on the power of a single antenna. The model we chose for the antenna is the dipolar one. In this approximation, we focus on a stationary problem but we could imagine a dynamical version based upon the complete Maxwell equations. For our purpose this level of precision in the model will be useless. For a complete and clear description of the physic of electromagnetism, see the book of J.D. Jackson [3]. The complete Maxwell equations are

$$\begin{cases} \dfrac{\partial D}{\partial T} - \operatorname{curl} H = 0 & \dfrac{\partial B}{\partial T} + \operatorname{curl} E = 0 \\ \operatorname{div} D = \varepsilon_0 \rho & \operatorname{div} B = 0 \\ D = \varepsilon_0 E & B = \mu_0 (H + M) \end{cases} \tag{12.2}$$

where D and E characterises the electric field, B and H the magnetic field, μ_0 and ε_0 physical constants, ρ the distribution of electric charges and M the distribution of magnetic moments.

In this study, as evoked above, we focus on the stationary part of the system and more exactly the magnetic part. Then, we will work on equations (12.2) (first line, first column and third line, second column):

$$\varepsilon_0 \frac{\partial E}{\partial t} - \operatorname{curl} H = 0, \quad B = \mu_0 (H + M).$$

Now, let perform a formal dimension study of the equation. In order to do so, we set, for $(\overline{e}, \overline{h}, \overline{\mu}, \overline{t}, \overline{x})$, positive real, dimension factors:

$$E = \overline{e}e,\ H = \overline{h}h,\ M = \overline{\mu}m,\ T = \overline{t}t,\ X = \overline{x}x.$$

Then, the equation in their dimensionless version becomes

$$\frac{\varepsilon_0 \overline{e}}{\overline{t}} \frac{\partial e}{\partial t} - \frac{\overline{h}}{\overline{x}} \mathrm{curl}_x\ h = 0, \quad \overline{b}b = \mu_0(\overline{h}h + \overline{\mu}m),$$

moreover we have

$$\frac{\overline{b}}{\overline{t}} \frac{\partial b}{\partial t} + \frac{\overline{e}}{\overline{x}} \mathrm{curl}_x\ e = 0.$$

The process we engage in order to obtain dimensionless equation implies, from the previous equations

$$\overline{\mu} = \overline{h},\ \mu_0 \overline{h} = \overline{b},\ \frac{\varepsilon_0 \overline{e}}{\overline{t}} = \frac{\overline{h}}{\overline{x}},\ \frac{\overline{b}}{\overline{t}} = \frac{\overline{e}}{\overline{x}}.$$

This leads to

$$\overline{h} = \frac{\overline{e}\,\overline{t}}{\mu_0 \overline{x}} = \frac{\varepsilon_0 \overline{x}\,\overline{e}}{\overline{t}}$$

then

$$\frac{\overline{x}^2}{\overline{t}^2} \varepsilon_0 \mu_0 \frac{\partial e}{\partial t} - \mathrm{curl}_x\ h = 0, \quad b = h + m.$$

We notice (see for example [3]) that $\varepsilon_0 \mu_0 = \dfrac{1}{c^2}$, where c is the speed of light, and we have

$$\left(\frac{\overline{x}}{c\overline{t}}\right)^2 \frac{\partial e}{\partial t} - \mathrm{curl}_x\ h = 0, \quad b = h + m, \quad \mathrm{div}_x(h + m) = 0. \tag{12.3}$$

Under the hypothesis that $\eta = \frac{\overline{x}}{c\overline{t}}$ is small compared to 1, we obtain formally the following approximated system

$$\mathrm{rot}_x h = 0, \quad \mathrm{div}_x h = -\mathrm{div}_x m. \tag{12.4}$$

To understand this equation and determine its solution, we use several theoretical elements developed, for example, in [2, 4]. As we will not focus on this problem in this article, we next summarize the ideas, with no theoretical details, used in order

to solve this problem. First, from the equation $\mathrm{rot}_x h = 0$, we deduce that, up to a constant, there exists φ, a function form $\mathbb{R}^3$ into $\mathbb{R}$, such that $\nabla_x \varphi = h$. This step leads to a Laplace equation:

$$\triangle_x \varphi = -\mathrm{div}_x m,$$

whose solution exists and is unique when m is a sum of Diracs like in our case; but, better than the uniqueness of the solution, we have a representation formula for this solution, using the Green kernel on $\mathbb{R}^3$:

$$\forall x \in \mathbb{R}^3 \backslash \{0\},\ G(x) = \frac{1}{4\pi |x|},$$

solution of the equation $\triangle_x G = \delta_0$ [4]. Thanks to this formula, we can give the expression of the solution of equation (12.4)

$$h = -\nabla_x \mathrm{div}_x \left(G * m\right),$$

where $*$ designates the two entries operator of convolution (see for example [4]). In our case, we can explicit this solution, in particular, using the properties of the convolution, we focus on $\nabla_x \mathrm{div}_x G(x-y)$

$$\nabla_x \mathrm{div}_x G(x-y) = \begin{pmatrix} \partial^2_{x_1} G(x-y) & \partial_{x_1}\partial_{x_2} G(x-y) & \partial_{x_1}\partial_{x_3} G(x-y) \\ \partial_{x_1}\partial_{x_2} G(x-y) & \partial^2_{x_2} G(x-y) & \partial_{x_1}\partial_{x_2} G(x-y) \\ \partial_{x_1}\partial_{x_3} G(x-y) & \partial_{x_2}\partial_{x_3} G(x-y) & \partial^2_{x_3} G(x-y) \end{pmatrix},$$

for (i, j) in $\{1, 2, 3\}^2$, if $i \neq j$

$$\partial_{x_i}\partial_{x_j} G(x-y) = \frac{-1}{4\pi}\left(3\frac{(x_i - y_i)(x_j - y_j)}{|x-y|^5}\right),$$

if $i = j$

$$\partial^2_{x_i} G(x-y) = \frac{-1}{4\pi}\left(3\frac{(x_i - y_i)^2}{|x-y|^5} - \frac{1}{|x-y|^3}\right),$$

then

$$\nabla_x \mathrm{div}_x G(x-y) = \frac{1}{|x-y|^3}\left(-\mathrm{Id} + 3\frac{(x-y)^t\ (x-y)}{|x-y|^2}\right),$$

Hence, using the fact that δ_0 is a neutral element for the convolution, we obtain, for a magnetic moment $m\delta_0$

$$\forall x \in \mathbb{R}^3 \backslash B(0,\varepsilon), \quad h(m)(x) = \frac{1}{4\pi |x|^3}\left(m - 3\frac{x^t x}{|x|^2}m\right).$$

In our case, we are interested on vertical antennas, then $m = (0,0,1)^t$ and on the measure at ground level, it is to say for $x = (x_1, x_2, 0)$. Finally, with these hypothesis we obtain

$$\forall x \in \mathbb{R}^3 \backslash B(0,\varepsilon), \quad h(m)(x) = \frac{1}{4\pi |x|^3}\begin{pmatrix}0\\0\\1\end{pmatrix}.$$

In what follows, we can only consider the third component of the magnetic field, $h_{d,3}$, which corresponds to the local power. This result can be directly generalised to the networks of antennas defined at the beginning of the modelling description and give $f\ \forall x \in \mathbb{R}^3 \backslash \bigcup_{y\in\Sigma} B(y,\varepsilon)$, for every $\varepsilon \in \mathbb{R}_*^+$

$$F_\varepsilon(\Sigma, M, P)(x) = \left|\sum_{y\in\Sigma} p_y \frac{m_y}{4\pi |x-y|^3}\right|.$$

In particular, using the previous results we set

$$\forall m \in \mathbb{N}, \alpha \in \mathbb{R}, W_\alpha^m = \left\{v \in \mathscr{D}'(\mathbb{R}^3), \forall \lambda \in \mathbb{N}^3, 0 \le \lambda \le m, (1+r^2)^{\frac{\alpha - m + |\lambda|}{2}} D^\lambda v \in L^2(\mathbb{R}^3)\right\},$$

where $\mathscr{D}'(\mathbb{R}^3)$ is the space of distributions (see [1]), D^λ denotes the partial derivative application. The topological dual space of W_α^m is $W_{-\alpha}^{-m}$.

Theorem 12.1 *There exists φ in W_0^1, unique up to a constant, such that $\nabla_x \phi$ is solution of the equation $rot_x h = 0$.*

Then, we can use this theorem to analyse the equation $\mathrm{rot}_x h = 0$: there exists a unique φ in W_0^1, up to a constant, such that $\nabla_x \varphi = h$, which implies

$$\mathrm{div}_x \nabla_x \varphi = \triangle_x \varphi = -\mathrm{div}_x m.$$

If m was in $L^2(\mathbb{R}^3; \mathbb{R}^3)$, we use the following theorems

Theorem 12.2 *The operator $\triangle_x$ is an isomorphism from W_1^1 into $W_1^{-1} \perp \mathbb{R}$ where $W_1^{-1} \perp \mathbb{R} = \{f \in W_1^{-1}, (f,1) = 0\}$.*

Lemma 12.1 *Given f in $L^2(\mathbb{R}^3)$, compactly supported, then div f is an element of $W_1^{-1} \perp \mathbb{R}$.*

$\frac{\overline{x}^2}{\overline{t}^2} = \frac{1}{\varepsilon_0 \mu_0} = c^2$, where c is the speed of light.

12.3 Two Fixed Antennas and Yet, Problems...

Let us begin with the case of two antennas. The power transmitted by this system, considering two antennas at distance λ with $M = \{e_3, e_3\}$[1] and $P = \{p, p\}$, is the following, for ε, p and h in $\mathbb{R}_*^+$ with $h > \varepsilon$

$$F_\varepsilon(\{(0,0,h),(\lambda,0,h)\}, M, P)(x) = p\frac{1}{4\pi|x-y_1|^3} + p\frac{1}{4\pi|x-y_2|^3}.$$

Here, to simplify our problem, we focus on what happens at ground level: given R in $\mathbb{R}_*^+$, $R > \lambda$, $\Omega = \{x \in \mathbb{R} | x \cdot e_3 = 0, |x - \frac{\lambda}{2}| \leq R\}$, then, by construction in this particular case, $\Omega_\varepsilon = \Omega$. Now, we re-write the power developed, setting $x = u \cdot e_1 + v \cdot e_2$:

$$\begin{aligned} F_\varepsilon(\{(0,0,2\varepsilon),(\lambda,0,2\varepsilon)\}, M, P)(x) &= G(u,v) \\ &= \frac{p}{4\pi}\left(\frac{1}{\sqrt{u^2+v^2+h^2}^3} + \frac{1}{\sqrt{(u-\lambda)^2+v^2+h^2}^3}\right), \end{aligned}$$

hence, we compute the gradient[2] of $G(u,v)$: $\forall (u,v) \in \mathbb{R}^2$,

$$\begin{aligned} \nabla G(u,v) = &\frac{3p}{4\pi}\left(\frac{u}{\sqrt{u^2+v^2+h^2}^5} + \frac{u-\lambda}{\sqrt{(u-\lambda)^2+v^2+h^2}^5}\right)\cdot d_1 + \\ &\frac{3p}{4\pi}\left(\frac{v}{\sqrt{u^2+v^2+h^2}^5} + \frac{v}{\sqrt{(u-\lambda)^2+v^2+h^2}^5}\right)\cdot d_2. \end{aligned}$$

Then, the critical points of G on $\mathbb{R}^2$ cancel this gradient and are: $(0,0)$, $(\lambda,0)$, $(\frac{\lambda}{2},0)$, the first and last are global maxima on $\mathbb{R}^2$ and the second is a local minimum. This property is obtained thanks to the fact that G is an element of $C^\infty(\mathbb{R}^2;\mathbb{R})$. Now, we take into account the place of interest of the solution in Ω, then, let find the global minimum of G on Ω. We can easily prove that, if another

[1] We set $(e_1, e_2, e_3) = \{(1,0,0),(0,1,0),(0,0,1)\}$.

[2] $(d_1, d_2) = \{(1,0),(0,1)\}$.

minimum than the one exhibited in $(\frac{\lambda}{2}, 0)$ exists, it must take place on the boundary of Ω. Then, let explore this boundary:

$$\forall (u, v, 0) \in \partial\Omega, \exists \alpha \in [0, \pi] \mid u = R\ \cos(\alpha) + \frac{\lambda}{2}, v = R\ \sin(\alpha),\ \text{then}$$

$$G(u, v) = G(\alpha) = \frac{p}{4\pi}\left(\frac{1}{\sqrt{(R\ \cos(\alpha) + \frac{\lambda}{2})^2 + R^2 \sin(\alpha)^2 + h^2}^3} + \frac{1}{\sqrt{(R\ \cos(\alpha) - \frac{\lambda}{2})^2 + R^2 \sin(\alpha)^2 + h^2}^3}\right),$$

by a direct computation of the critical points of $G(\alpha)$, we obtain that the global minimum is attained for $\alpha = \frac{\pi}{2}$ and $\frac{3\pi}{2}$. Now, the question is: can we ensure the admissibility of the solution? Here, admissibility means that on the domain of interest, the power developed by the antennas network is, on each point, comprised between $\underline{m}$ and $\overline{m}$:

$$\underline{m} \leq G(\frac{\pi}{2}) \leq F_\varepsilon(\{(0, 0, 2\varepsilon), (\lambda, 0, 2\varepsilon)\}, M, P)(0, 0, 0) \leq \overline{m}.$$

12.4 More Antennas and No Analytic Solving

Now, let add more antennas in the network. Typically, we have to choose parameters to move in order to ensure the suitability of the configuration. We could have no constrains on the number of antennas, their positions, orientations and power, but it will be too complex. In order to manage this problem, we will treat two versions, in these versions, we keep the same orientation for all the antennas, it is to say (0, 0, 1). This means in our case:

- Problem 1: Fixed number of antennas, fixed powers, the positions vary among a finite predetermined set of positions.
- Problem 2: Fixed ground positions of the antennas, but the power and the height of the antennas vary.
- Problem 3: Fixed power and heights of the antennas, the number of antennas is fixed and the positions vary freely.

Here, it will not be possible to treat analytically these cases in general case, we must use a powerful tool: the numerical optimisation (see for example [1]). In these three cases, the existence of at least a solution is almost every time guaranteed (classical proof), but not the uniqueness of this solution. In the first case,

a systematic exploration of the set of solution may be long but possible. Here is a glimpse for each of these three problems.

12.4.1 Problem 1

Imagine that n is the number of antennas and p the number of possible position, then the number of possibilities is given by $N(n, p) = A_p^n = \frac{p!}{(p-n)!}$. If p increases but n does not change, we see that a not so bad approximation of $N(n, p)$ would be $\tilde{N}(n, p) = p^n$. Then, when p becomes huge, if we want to compute $F_\varepsilon(\Sigma, M, P)(x)$, we must perform p computations of the function composed mainly by a sum of n real numbers. In this context, if n_+ is the elementary computation time, the total computation time becomes: $np^{n+1}n_+$. For example, if $p = 100$ and $n = 10$, on a computer whose performance is 3 GHz (10^9 Hz) and if we accept approximatively 100 cycles for n_+, we have a total computation time of:

$$T = 10.100^{11}.10^{-9}.100\,\mathrm{s} = 10^{16}\,\mathrm{s},$$

it means almost 317 millions of years! Even on the most powerful computer, 10^{17} flops, the computation time would be almost 11 days. This is not acceptable and in order to minimise the computation time, we could develop new efficient algorithms.

It is necessary to be careful as, for this problem, even if you find an algorithm sufficiently fast to performs computations, you do not know if the problem admits a solution; in fact, the constraints may not be fulfilled, in particular the maximum constraint if the power is too high but also the minimum constraint if the power is too low and the assigned position not sufficiently close.

12.4.2 Problem 2

Strangely, this problem is in fact simpler than the previous one. The positions are fixed but power and height of the antennas vary. The existence of a solution, like in the previous problem, is not guaranteed if the network of position is ill-prepared. Here, we can, starting from a well chosen configuration, apply a gradient method adapted to the constraint. Two problems appear: gradient method adapted to the constraint and good starting configuration. Here, the first step is to find a starting position. But, even if we find this kind of position, we are not sure to be able to attain the optimal solution. The projected gradient method will guarantee that the power decreases, respecting the constraints, but this decreases ensures that we arrive in a local minimum which is not necessarily the global minimum. The question is then: if we arrive in a local minimum, do we stop or do we try to find a better one in order to attain the global one. One algorithm is the so called simulated annealing,

this method is inspired from the technics of heating and cooling when injecting heat in a system.

The main principle of this algorithm is then the following: a gradient descent algorithm (projected in our case), perturbed regularly in order to push out of possible non optimal minima bowl.

12.4.3 Problem 3

This case is much more difficult than the previous ones, but quite similar to Problem 2. We could see it as a simple adjunction of a third dimension (vertical position of the antenna and height). Here, we can imagine to apply the previously described algorithm.

12.5 A Complex Situation

In fact, modelling of the antennas covering is much more complex and would require the resolution of Maxwell equations in "town" represented by volume with given electric permittivity and magnetic permeability. The models used effectively are combining, in order to accelerate the computation, a ray tracing part, using the classical geometrical light propagation and, when necessary, a complete electromagnetic resolution in order to catch the diffraction phenomena.

12.6 Conclusion

This text is not extensive but gives the tracks in order to treat a simplified version of an important optimisation problem. Nevertheless, in scientific literature, there exists several occurrences treating the wave propagation in complex areas. In particular, the perfect simulation of problem is almost impossible. The propagation of electromagnetic wave using the Maxwell model is highly dependant of the exact geometry and composition obstacles, mobile or fixed, the humidity ratio and many other parameters not manageable exhaustively. In this context, the modelling process describe gives an example of simplification in order to build a manageable problem in finite time. This process is essential and must be carefully documented in order to identify the simplification and prevent errors of interpretations of the results obtained by simulation.

Acknowledgments I thank Christophe Picard for his carefully reading this article and his pertinent remarks and propositions.

References

1. Allaire, G. and Craig A., *Numerical Analysis and Optimization*. Numerical Mathematics and Scientific Computation, OUP Oxford , 2007.
2. Blanchard P., Brüning E., *Schwartz Distributions. In: Mathematical Methods in Physics*. Progress in Mathematical Physics, vol 26. Birkhäuser, pp 27–45, Boston, MA, 2003.
3. Jackson J.D., *Classical Electrodynamics*, 2nd Ed. Wiley 1975.
4. Lions J.-L. et al, *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 2 Functional and Variational Methods*, Springer Berlin Heidelberg, 1999.