Samuel Soubeyrand

Contributions to Statistical Plant and Animal Epidemiology

Document scientifique pour l'habilitation
à diriger des recherches en sciences

Date de soutenance
12 septembre 2016

Section CNU 2600

Mathématiques appliquées
et applications des mathématiques

Université d'Aix-Marseille
Faculté des Sciences

Jury

Bar-Hen Avner, Université Paris 5	Rapporteur
Dubois-Peyrard Nathalie, INRA	Rapportrice
Gabriel Edith, Université Avignon	Présidente
Gibson Gavin, Université Heriot-Watt, SCO	Examinateur
Lantuéjoul Christian, MinesParisTech	Membre invité
Parent Eric, AgroParisTech	Examinateur
Penttinen Antti, Université Jyväskylä, FIN	Examinateur
Pommeret Denys, Université Aix-Marseille	Rapporteur

Preface

Ce document, écrit à l'occasion de la demande d'Habilitation à Diriger des Recherches, dépeint les recherches que j'ai menées depuis mon doctorat obtenu en 2005. Dans cette synthèse de mes travaux, l'accent est mis sur les approches de modélisation et d'inférence statistique pour les besoins de l'exercice. Ceci dit, les motivations scientifiques qui sous-tendent mon travail émergent le plus souvent des domaines d'applications, et je m'efforce tout particulièrement de faire dialoguer, de la question au résultat, la statistique et l'épidémiologie. L'environnement scientifique offert par l'INRA est une des clefs de ce dialogue. Je remercie donc mes collègues de l'INRA, et tout particulièrement ceux de BioSP et du département SPE, pour leur disponibilité, les échanges fructueux que j'entretiens avec eux et les opportunités qu'ils m'offrent. Je remercie également les collègues hors INRA qui permettent notamment d'élargir l'horizon de mes recherches. Ma gratitude va aussi aux membres du jury qui ont consacré une partie de leur temps pour évaluer mon travail. Enfin, un grand merci à mes proches, notamment à Julie, pour leur soutien.

Avignon,

Samuel Soubeyrand
Septembre 2016

Université d'Aix-Marseille

ATTESTATION DE REUSSITE AU DIPLOME

Le Président de l'Université atteste que

l' Habilitation à Diriger des Recherches EN SCIENCES
a été décernée à

Monsieur SAMUEL SOUBEYRAND
né le 17 septembre 1978 à AVIGNON (084)

au titre de l'année universitaire 2015/2016

Date de soutenance	: 12 septembre 2016
Etablissement soutenance	: UNIVERSITE D'AIX-MARSEILLE
Jury	: Mme EDITH GABRIEL, Président du jury, MAITRE DE CONFERENCES UNIVERSITE AVIGNON M. AVNER BAR-HEN, Membre du jury, PROFESSEUR DES UNIVERSITES UNIVERSITE PARIS 5 Mme NATHALIE DUBOIS PEYRARD, Membre du jury, DIRECTEUR DE RECHERCHE INRA M. GAVIN GIBSON, Membre du jury, PROFESSEUR DES UNIVERSITES UK Heriot-Watt University, Edinburgh M. ERIC PARENT, Membre du jury, INGENIEUR ENGREF M. ANTTI PENTTINEN, Membre du jury, PROFESSEUR EMERITE UNIVERSITE JYVASKYLA / FINLANDE M. DENYS POMMERET, Membre du jury, PROFESSEUR DES UNIVERSITES UNIVERSITE D'AIX-MARSEILLE
Section CNU	: 2600 - Mathématiques appliquées et applications des mathématiques

Faculté des Sciences AMU

Fait à Marseille, le 14 septembre 2016

Université d'Aix-Marseille AMU

Yvon BERLAND

N° étudiant : 15039458

Avis important: Il ne peut être délivré qu'un seul exemplaire de cette attestation. Aucun duplicata ne sera fourni.

Contents

Abbreviations

2D, 3D: two- and three-dimensional (space)
AIC: Akaike's information criterion
ABC: Approximate Bayesian computation
ANR: French national research agency
BioSP: Biostatistics and spatial processes research unit
BMSE: Bayesian mean square error
BvM: Bernstein-von Mises (theorem)
CB-MAP: Contrast-based maximum *a posteriori*
CB-posterior distribution: Contrast-based posterior distribution
CCF: Circular correlation function
EBBM: Energy balance model with memory
FMDV: Foot-and-mouth disease virus
GDM: Group dispersal model
GLMM: Generalized linear mixed model
GLM: Generalized linear model
GRP: Gaussian random process
HDF: Horizontal dispersal function
HMC: Hamiltonian Monte-Carlo (algorithm)
HTS: High-throughput sequencing
HYSPLIT: Hybrid Single Particle Lagrangian Integrated Trajectory (model)
IDM: Independent dispersal model
INRA: French national institute for agricultural research
KZN: Kwa-Zulu Natal (eastern province of South Africa)
LPP: Local posterior probability
MCEM: Monte-Carlo expectation-maximization (algorithm)
MCMC: Markov chain Monte-Carlo (algorithm)
MLE: Maximum likelihood estimate
MPLE: Maximum pseudo-likelihood estimate
MRCA: Most recent common ancestor
MSE: Mean square error
PDE: Partial differential equation

PEP: Point estimates of parameters
PMSE: Partial mean square error
PODS: Pseudo-observed data set
PPM: Pine processionary moth
Sd.: Standard deivation
SEIR: Susceptible-exposed-infectious-removed
SPE: *Plant-health and environment* division of INRA
UK: United Kingdom
USA: United States of America
VDF: Vertical dispersal function

1

Introduction

Foreword

May 2, 2016

I wish to apologize to members of the jury because some of them would have preferred to read this manuscript in French, and the others would have preferred to read this manuscript in better English. However, I hope this text is clear enough to allow all the jury members to assess my ability to supervise research.

1.1 Short biography

In early January 2016, my older daughter (5 years old) explained my job to another child by using her words (and her body movements, which cannot be reproduced here): "My father is a researcher. He's like a dog, which smells and finds." Obviously, this analogy is over-simplistic, and I have substantial work to do to make my daughter understand what a researcher is and what kind of researcher I am. This document, which was written to obtain the *habilitation à diriger des recherches* (i.e. the accreditation to supervise research), illustrates what kind of researcher I am: a researcher who carries out his own research and who contributes to the research of colleagues; a researcher who tends to explore various fields, techniques and issues, but who is consistently interested in recurrent topics.

My research has been strongly influenced by my affiliation, since 2002, to INRA (the French national research institute for agricultural research), which is a hotspot for multidisciplinary science. It has also been influenced by my early education (I liked to put my thinking cap on to understand issues in mathematics, and also in history, physics, theology, sociology, etc.) and by my university curriculum: I have an undergraduate degree in mathematics and physics (*classe préparatoire*, Aix-en-Provence, 1996-1999), a License in

economical science (Univ. Rennes 1, 2000-2001), a Master in statistics and information analysis (ENSAI, Rennes, 1999-2002), a Master in fundamental mathematics and applications with a specialization in statistics (Univ. Rennes 2, 2001-2002) and a Doctorate in biostatistics (Univ. Montpellier 2, 2002-2005). My research has also been influenced by my time spent at the University of Chicago with Michael Stein (master internship), at the Plant Epidemiology research unit of INRA (Grignon) with Ivan Sache, at the Ludwig-Maximilians University with Leonhard Held, at the University of Jyväskylä with Antti Penttinen, at the University of Glasgow with Daniel Haydon, and at the BioSP research unit of INRA (Avignon) with Joël Chadœuf (during my doctoral period) and all the other members of BioSP, who contribute to establish a stimulating environment.

These influences (and a few professional opportunities) led me to carry out research in spatial and spatio-temporal statistics applied to plant and animal epidemiology.

1.2 Statistics for plant (and animal) epidemiology

Epidemiology can be briefly described as the study of the development of disease in populations (this short description encompasses human, animal and plant epidemiology). Disease is a broad term, which includes a huge variety of disorders. Here, I focus on infectious diseases caused by pathogens such as fungi, viruses and bacteria. For such diseases, human, animal and plant epidemiology share the same general concepts and mechanisms (e.g. transmission, incubation, basic reproduction number, co-evolution, etc.) and can be tackled in very similar ways from the point of view of process modeling and data analysis[1].

An early demonstration of the utility of the spatial and quantitative analysis of data in epidemiology was made by Snow (1855) in his study *On the Mode of Communication of Cholera*. In the mid-19th century, Snow identified impure water as a vector for cholera: he mapped fatal cholera cases in Soho (London; see Figure 1.1), noted the spatial clustering of these cases and identified the water pump from Broad Street as a potential source of the outbreak (unknown particles were observed with a microscope in the water supplied by the Broad Street pump, and when this pump was closed, the local epidemic stopped). Snow carried out a larger-scale analysis of deaths from cholera (see Table 1.2 and Figure 1.3). A larger rate of mortality was observed in subdistricts where water was supplied by the Southwark and Vauxhall Water Company whose water was contaminated by sewage.

Snow's investigation on cholera shows how spatial and quantitative analysis of data contributes to the understanding of epidemics affecting humans.

[1] It has however to be noted that some aspects of plant epidemiology are distinctive from human and animal epidemiology and lead to specific challenges in modeling plant diseases (Cunniffe et al., 2015).

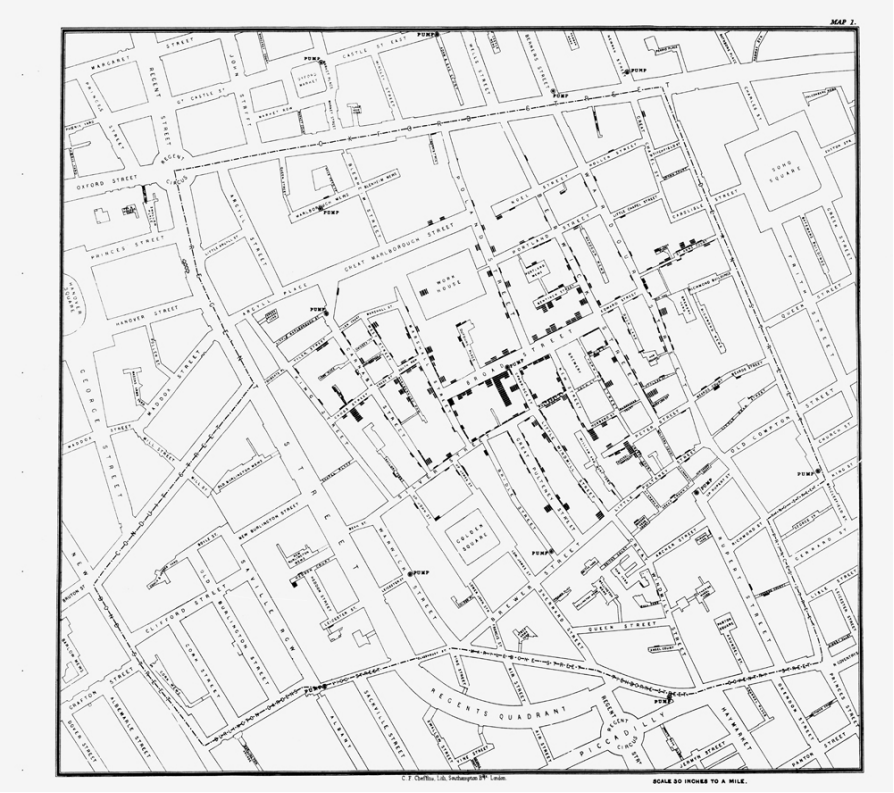


Fig. 1.1. Map showing the deaths from cholera in Broad Street, Golden Square, and the neighborhood (Soho, London), from 19th August to 30th September 1854. A black bar for each death is placed in the location of the house in which the fatal attack took place. This map also indicates the locations of water pumps to which the public had access. Original map from Snow (1855).

This statement holds for epidemics affecting plants as well. The study of diseases of plants is an old science. For example, Theophrastus (c. 372 – c. 287 BC) provided a written testimony on plant diseases in his *Enquiry into Plants* (Theophrastus, 1916, translated by Hort). The quantification of epidemics in plant populations had expanded much later, around the mid-20th century, especially with the development of theoretical models describing the dynamics of diseases in time and/or space; see Frantzen (2007, chap. 1) and Strange (2003, chap. 3). Thus, a current of thought called *theoretical plant epidemiology* emerged, using works in mathematical biology as a foundation (e.g. Kermack and McKendrick, 1927), and it led to original studies such as those on the effect of crop heterogeneity on the spread of diseases (Gilligan, 2008; Jeger, 2000). Meanwhile, the use of data and accompanying statistical

Sub-Districts.	Population in 1851.	Deaths from Cholera in 1853.	Deaths by Cholera in each 100,000 living.	Water Supply.
St. Saviour, Southwark	19,709	45	227	
St. Olave . . .	8,015	19	237	
St. John, Horsleydown	11,360	7	61	
St. James, Bermondsey	18,899	21	111	
St. Mary Magdalen .	13,934	27	193	
Leather Market .	15,295	23	153	Southwark and Vauxhall Water Company only.
Rotherhithe* . .	17,805	20	112	
Wandsworth .	9,611	3	31	
Battersea . . .	10,560	11	104	
Putney . . .	5,280	—	—	
Camberwell . .	17,742	9	50	
Peckham . . .	19,444	7	36	
Christchurch, Southwk.	16,022	7	43	
Kent Road . .	18,126	37	204	
Borough Road . .	15,862	26	163	
London Road . .	17,836	9	50	
Trinity, Newington .	20,922	11	52	
St. Peter, Walworth .	29,861	23	77	
St. Mary, Newington .	14,033	5	35	Lambeth Water Company, and Southwark and Vauxhall Company.
Waterloo (1st part) .	14,088	1	7	
Waterloo (2nd part) .	18,348	7	38	
Lambeth Church (1st part . . .	18,409	9	48	
Lambeth Church (2nd part) . . .	26,784	11	41	
Kennington (1st part)	24,261	12	49	
Kennington (2nd part)	18,848	6	31	
Brixton . . .	14,610	2	13	
Clapham . . .	16,290	10	61	
St. George, Camberwell	15,849	6	37	
Norwood . . .	3,977	—	—	Lambeth Water Company only.
Streatham . . .	9,023	—	—	
Dulwich . . .	1,632	—	—	
First 12 sub-districts .	167,654	192	114	Southwk. & Vaux.
Next 16 sub-districts .	301,149	182	60	Both Companies.
Last 3 sub-districts .	14,632	—	—	Lambeth Comp.

* A part of Rotherhithe was supplied by the Kent Water Company; but there was no cholera in this part.

Fig. 1.2. Table providing counts of deaths from cholera in sub-districts on the south side of the Thames in London. The table also indicates companies supplying water for each sub-district. Original table from Snow (1855).

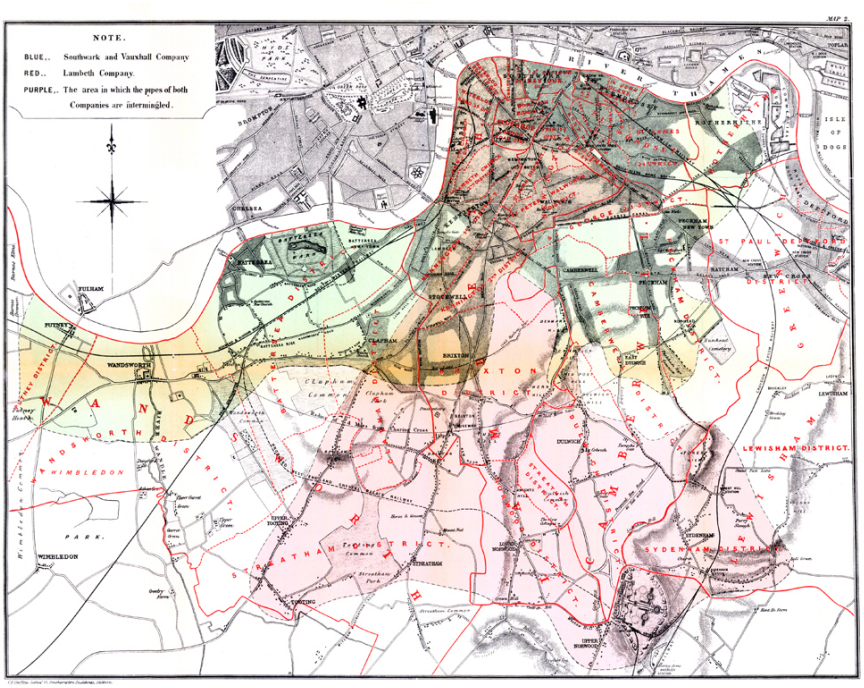


Fig. 1.3. Map showing the boundaries of the Registrar-General's districts on the south side of the Thames in London, and the water supply of those districts. Original map from Snow (1855).

methods has also contributed to gain insight into processes involved in plant epidemics; a precursory illustration of this set of approaches is the estimation and interpretation of plant disease dispersal gradients (Gregory, 1945, 1968).

During the last decades, significant advances have been made in statistical epidemiology in general and the statistical analysis of plant epidemics in particular. For example, generalized linear mixed models, survival analysis and decision analysis have led to testing existing hypotheses and addressing new questions (Scherm et al., 2006). More recently, the combination of a mechanistic vision of epidemics, a probabilistic vision of observation processes and a statistical approach for inferring model parameters and latent variables has led to re-exploring the link between theory and data in plant epidemiology (e.g. see Gibson, 1997; Soubeyrand et al., 2009c).

This brief overview gives only an idea of the vertiginous corpus of methods and results which have been established in quantitative analysis for plant epidemiology. Since the beginning of my PhD studies, I have participated in the development of this corpus and tried to bring original ideas by carrying out research at the interplay between statistics, modeling, probability, plant

epidemiology and, occasionally, animal epidemiology. Carrying out such multidisciplinary research led me to be a researcher in applied statistics. From a publication perspective, this means writing articles for journals at the interface between statistics and applied fields or for journals in other scientific fields besides statistics. However, these articles may include advanced methodological developments. For instance, my article published in *Theoretical Population Biology* (Soubeyrand et al., 2008a) provides and characterizes a new auto-correlation function for circular Gaussian random processes. Since publication, this work has been cited in the statistical literature, namely in *Bernoulli* by Gneiting (2013) and Cheng and Xiao (2016), and in the *Journal of the American Statistical Association* by Porcu et al. (2015).

Beyond my personal situation, it is interesting to see how application fields can lead researchers in applied statistics to investigate new inference algorithms, new spatial models, new testing procedures, etc. This is part of the iterative process of research: in any scientific field, once new results have been stated, one may be interested in refining them or understanding the discrepancies between the results and reality. This in turn leads to new models and methods. This iterative process led me, for instance, to propose new point process models for particle dispersal and a new form of approximate Bayesian computation (ABC).

1.3 My toolbox

In my research practice, I am not focused on a given methodology, but I exploit diverse statistical and modeling tools and explore some of them in depth. The main tools I have used are spatial and spatio-temporal point processes, continuous-time Markov and semi-Markov processes, state-space models and estimation algorithms.

I have used spatial and spatio-temporal point processes mainly for describing the dispersal of particles that propagate plant diseases (a point in these point processes can represent the deposit location of a particle). For this application, inhomogeneous Poisson point processes, Cox point processes, and inhomogeneous Neyman-Scott point processes are particularly relevant because their inhomogeneous intensity functions can model the spatial and temporal heterogeneity of the risk of infection. The heterogeneity of the risk is due to sources of infection, which have non-uniform spatio-temporal patterns.

I have mostly exploited continuous-time Markov and semi-Markov processes to build genetic-space-time and individual-based models of epidemics caused by fast-evolving pathogens. The (semi-)Markov property leads to tractable models, in terms of estimation, despite the complex dependence structure due to the interplay in the models of genetics, space and time.

State-space models can be found in most of my works because they form a flexible modeling tool to address hidden processes (i.e. influential processes for which no explanatory variable is available), scale change (e.g. between

the process scale and the data scale) and data heterogeneity (in this case, different types of data can be modeled conditional on a single unobserved process model). I have recurrently considered a specific class of state-space models, namely the mechanistic-statistical models, which combine a *process model* built in a mechanistic way and a *data model* of the observation process.

My vision of estimation is more opportunistic than founded on dogmas. Thus, when the model is rather simple and no prior information is available, I apply maximum likelihood estimation or, more generally, minimum contrast estimation. In contrast, when there is an informative prior knowledge about parameters or when the model incorporates latent variables, which generate a complex dependence structure in the model, then I adopt the Bayesian approach. In the latter case, the models I deal with generally require the use of numerical tools such as MCMC (Markov chain Monte-Carlo) algorithms or ABC.

In the following chapters, I more marginally exploit other modeling and statistical tools, for instance, circular Gaussian processes modeling anisotropy functions, discrete-time Markov processes modeling the vertical dispersal of particles, cylinder-based models providing a concise representation of group dispersal, partial differential equations (PDE) providing a concise representation of population dynamics, convergence analyses providing the asymptotic behavior of estimators, and randomization procedures allowing for the construction of tests adapted to specific case-studies.

1.4 Contents of the manuscript

In Chapter 2, I illustrate how spatial Poisson point processes and other tools of stochastic geometry, such as spatio-temporal point processes and object-based models, can be exploited to model, infer and simulate processes depending on the dispersal of particles. This chapter is introduced with an aggregative approach for constructing dispersal models, which are both based on a fine-scale description of the dispersal dynamics and adapted to larger-scale data classically collected in plant epidemiology. Then, I present my work on anisotropic dispersal models and group dispersal models. I conclude this chapter with two examples of multi-year epidemics analyzed with modeling and inference tools presented throughout Chapter 2.

Chapter 3 presents my work on genetic-space-time models, which are used to infer transmission trees using spatio-temporal epidemiological and genetic data. These models combine a spatio-temporal dynamics of the pathogen, and an evolutionary model for the evolution of genetic sequences of the pathogen. Estimation of model parameters and latent variables is carried out in the Bayesian framework via approximate MCMC algorithms. This approach was applied to infer transmission trees for foot-and-mouth outbreaks and a rabies endemic dynamic.

Chapter 4 addresses mechanistic-statistical models, which combine a *process model* for the dynamics under study and a *data model* for the observation process. Such models incorporating stochastic *process models* are introduced in Chapter 2, but Chapter 4 focuses on PDE-based mechanistic-statistical models. This approach is presented and applied to simulated and real-life case studies concerning biological invasions (epidemics can be viewed as a particular type of biological invasions) and long-term climatic dynamics.

In Chapter 5, I present three methodological works concerning parameter estimation without likelihood. Such estimation procedures (which may circumvent difficulties encountered in the implementation of likelihood-based approaches) can be particularly valuable when one aims to fit realistic, spatio-temporal, epidemiological models to data[2]. Thus, in this chapter, I explore the consequences of replacing the likelihood in the Bayesian formula of the posterior distribution with a function of a contrast, I present an algorithm for optimizing the distance between functional summary statistics in ABC, and I present a study of the weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and its implications in ABC.

The last chapter of this document, Chapter 6, provides complementary information concerning my work. First, it gives a snapshot of other contributions that have not been introduced in Chapters 2–5. Then, it gives information about supervision, teaching, networks and projects I have been involved in. Finally, it is concluded by a section about my perspectives of research.

[2] Fitting realistic, spatio-temporal, epidemiological models to data is often a difficult task because one generally has to handle, for example, latent processes, spatial dependencies, and heterogeneity in data.

2

Stochastic geometry applied to particle dispersal studies

Author's references: Allard and Soubeyrand (2012), Bourgeois et al. (2012), Bousset et al. (2015), Mrkvička and Soubeyrand (2015), Rieux et al. (2014), Soubeyrand et al. (2007b), Soubeyrand et al. (2007c), Soubeyrand et al. (2007d), Soubeyrand et al. (2008a), Soubeyrand et al. (2008b), Soubeyrand et al. (2009b), Soubeyrand et al. (2009c), Soubeyrand et al. (2011), Soubeyrand et al. (2014b), Soubeyrand et al. (2015).

Plant diseases due to fungi such as rusts and powdery mildew are mainly spread through the dissemination of microscopic particles called spores, which are released by wind gusts from symptomatic plants (Ingold, 1971; Rapilly, 1991). Characterizing the dissemination of spores contributes to understanding the dynamics of epidemics, assessing disease impacts on crop growth and crop yield, and designing control strategies. Spatial point processes (Diggle, 1983; Illian et al., 2008; Stoyan et al., 1995) naturally emerge in this context for modeling the spatial pattern of the deposit locations of spores. For instance, a typical situation consists of assuming that (i) the spores are emitted by one or several point sources in the 2D plane, (ii) transports of particles are mutually independent, and (iii) dispersal distances separating, in a given direction, the source locations and the deposit locations of particles are drawn from a decreasing probability density function. Under these assumptions, the spatial pattern of deposit locations in the 2D plane can be modeled by a spatial Poisson point process with an inhomogeneous intensity function. This process, which is a classical tool of stochastic geometry, is a basic component included, explicitly or implicitly, in many spatial dispersal models and spatio-temporal propagation models representing the dynamics of airborne plant diseases.

In this chapter, we aim to illustrate how spatial Poisson point processes and other tools of stochastic geometry such as spatio-temporal point processes and object-based models can be exploited to model, infer and simulate processes depending on the dispersal of particles. In this context, inference is usually based on standard methods and algorithms, e.g. maximum likelihood with a Nelder-Mead or an MCEM algorithm, and Bayesian estimation with an MCMC algorithm.

Section 2.1 describes an aggregative approach to building dispersal models adapted to data classically collected in plant epidemiology. This aggregative

approach is based on a fine-scale description of the dispersal dynamics that is derived to obtain larger-scale models of observed processes. Section 2.2 presents a series of anisotropic dispersal models constructed to characterize dispersal capacities of particles as a function of the direction. Section 2.3 introduces group dispersal models, which relax the independence hypothesis often assumed for the transports of particles. Sections 2.4 and 2.5 give examples of medium and large spatial-scale, multi-year epidemics analyzed with modeling and inference tools presented along this chapter.

2.1 An aggregative approach to build dispersal models

2.1.1 Summary of the approach

Wind-borne dispersal of particles can be studied at various scales: within a few square centimeters as well as between continents. By considering dispersal from a mechanistic perspective, we show in this section how to develop specific but coherent models for dispersal processes observed at different scales: *specific* because each model is tailored for a given situation, *coherent* because all models stem from a single base model. For this purpose, we build a model at a fine scale, i.e. a scale at which describing the sources of variations is natural, inherent and intuitive. Then, models at larger scales are built based on the fine-scale model, using an approach similar to the multi-scale modeling approach developed in physics where a macroscopic model is derived from a microscopic model (Weinan and Engquist, 2003; Weinan et al., 2003). Thus, explicit links between model structures at the fine scale and at each specific scale can be exhibited, and parameter estimations corresponding to different scales can be compared.

Here, the fine-scale model describes the probabilistic behavior of the presence/absence of the disease on small-scale susceptible units. The model includes the effects of spatially unstructured and structured covariates (e.g. due to genotype, physiology, climate) affecting the infectiousness of the infectious units and the receptivity of the susceptible units. Then, the fine-scale model can be scaled up to build larger-scale models adapted to observations.

2.1.2 Fine-scale model

Assumptions

We focus on the spread of a plant disease between two dates corresponding to the beginning and the end of an epidemic cycle. The disease of interest is transmitted via particles, which can be either specialized cells (spores), whole organisms (bacteria), or structures embedding pathogens (pollen grains, insects).

We assume that the variability of the disease cycle duration is negligible, and that a common starting point in time exists for the transmission from all infectious plants.

We assume that at the starting point the infectious plant units are detectable, and that they remain infectious during the cycle. At the end of the cycle, we assume that the newly infected plant units, thereafter called infected units, are detectable. The newly infected units are not infectious during the cycle.

We assume that the rules governing the transmission mechanisms are the same in all the spatial domain we are looking at.

Plants or plant units are considered as points in space marked by a qualitative sanitary status: either healthy, exposed (i.e. infected but not infectious) or infectious. No new plant unit is generated during the study period.

From a temporal point of view, time is discrete, each time step corresponding to the beginning of a cycle.

Epidemic spread is modeled by a three-step mechanism. First, particles are dispersed from each infectious plant or plant unit. Second, the accumulation of particles over a given susceptible unit defines a local infectious potential. Third, the susceptible unit becomes infected with a success probability depending on the local infectious potential.

Mathematical translation

Let x_i denote the location of the ith unit in the studied spatial domain. For a given time t, let $\delta_{it} = 1$ if the health status of unit i is observed at time t, $\delta_{it} = 0$ otherwise. Health status of unit i at time t is described by the binary variables $\mathbf{S}_{it}$, $\mathbf{E}_{it}$ and $\mathbf{I}_{it}$:

- $\mathbf{S}_{it} = 1$ if unit i is susceptible (i.e. healthy), $\mathbf{S}_{it} = 0$ otherwise,
- $\mathbf{E}_{it} = 1$ if unit i is exposed (i.e. infected but not infectious), $\mathbf{E}_{it} = 0$ otherwise, and
- $\mathbf{I}_{it} = 1$ if unit i is infectious, $\mathbf{I}_{it} = 0$ otherwise.

Particle dispersal from a given infectious unit i is described by the function $x \mapsto f(x - x_i)$, where x is any location in the study domain and f is a dispersal kernel, i.e. the probability distribution function of the deposit locations of particles emitted at the origin. Various parametric forms have been proposed for the dispersal kernel (e.g. see Austerlitz et al., 2004; Tufto et al., 1997, and the following sections), which is a key component of numerous propagation models in epidemiology and ecology.

The local infectious potential at location x and time t (viewed as a measure of the risk of infection of a susceptible host unit that would be located at x) is written as the following weighted sum (Mollison, 1977):

$$\lambda(x) = \sum_i c_i \mathbf{I}_{it} f(x - x_i), \tag{2.1}$$

where the contribution of each infectious unit depends on the spatial lag $x-x_i$ between the infectious unit and the target location x, and on the infection strength $c_i \geq 0$ of the infectious unit. Then, the probability of infection of a susceptible unit located at point x_j is described by a function depending on the local infectious potential:

$$P(\mathbf{E}_{j,t+1} = 1 \mid \lambda(x_j), \mathbf{S}_{jt} = 1) = g(\lambda(x_j)),$$

where g is a link function from $\mathbb{R}_+$ to $[0, 1]$. Interestingly, the form of g has not to be chosen arbitrarily, but it can be determined via additional mechanistic assumptions such as the ones proposed in the paragraph entitled *Examples of specifications* (see below).

If all infectious units are observed and if the observations are made at the beginning and the end of a cycle, parameter estimation can then be carried out by maximizing the following log-likelihood:

$$\sum_{\substack{j:\delta_{jt}\delta_{j,t+1}=1 \\ \mathbf{S}_{jt}=1}} \mathbf{E}_{j,t+1} \log\{g(\lambda(x_j))\} + (1 - \mathbf{E}_{j,t+1}) \log\{1 - g(\lambda(x_j))\}. \tag{2.2}$$

Depending on the shape of f, (2.2) is the log-likelihood of a generalized linear or nonlinear model with Bernoulli observation distribution (Harrell, 2013; Huet, 2004; McCullagh and Nelder, 1989).

Note that in (2.2) the sum is computed only for units j such that $\mathbf{S}_{jt} = 1$ because the other units, already infected at time t, do not bring information on the parameters in the framework of interest here. In Chapter 3, we will study situations leading to more complex likelihoods including more data, more processes and more parameters.

Examples of specifications

In practice, one must specify the nature of the infectious and susceptible units, the dispersal kernel f and the other components of the model. The list below provides typical specifications.

- Units can be agricultural plots, plants, leaves or other plant sections. The specified resolution determines what one means by *fine-scale* model.
- *Poisson specification.* Each infectious unit i spreads around its location a random number of particles, for example a Poisson number of particles with mean c_i. The locations of particles dispersed around infectious unit i are, for example, independently distributed from a 2D-exponential dispersal kernel:

$$x \mapsto f(x - x_i) = \frac{1}{2\pi\beta^2} \exp\left(-\frac{||x - x_i||}{\beta}\right), \tag{2.3}$$

where $||\cdot||$ is the Euclidean distance and $\beta > 0$ is called dispersal parameter[1]. Thus, the random field of particles generated by i is an inohomogeneous Poisson point process with intensity function $x \mapsto c_i f(x - x_i)$ defined over $\mathbb{R}^2$. Assuming that dispersal processes from different infectious units are independent, the random field of particles generated by all infectious units is an inhomogeneous Poisson point process whose intensity at point x is the local infectious potential $\lambda(x) = \sum_i c_i \mathbf{I}_{it} f(x - x_i)$.

- The argument in the dispersal kernel f is often the Euclidean distance, as in Equation (2.3), or a geographic distance, as in Sections 2.4 and 2.5. However, other types of arguments can be used depending upon the context. Indeed, f can be a function of the distance and the direction (see Section 2.2) if there is a prevailing wind for example. If the disease spreads through contacts between individuals, relations between individuals can be modeled by a network and distances on this network used as the argument of the dispersal function (Dargatz et al., 2005; Hufnagel et al., 2004; Parham and Ferguson, 2006).
- The susceptible unit, at the fine scale, can be an infinitesimal susceptible zone with area dx. The health status $\mathbf{E}_{j,t+1}$ is defined, in this case, by the presence or the absence of the disease at time $t+1$ on the susceptible unit j with area dx and location x_j. Under the *Poisson specification* made above, the area dx captures a Poisson number of particles with mean $\lambda(x_j)dx$. Assuming that particle attacks are independent and that an attack is successful (i.e. it leads to infection) with probability a_j, then j captures a Poisson number of successfully-attacking particles with mean $a_j\lambda(x_j)dx$, and the probability that j is infected satisfies:

$$P(\mathbf{E}_{j,t+1} = 1 \mid \lambda(x_j), \mathbf{S}_{jt} = 1) = g_j(\lambda(x_j)) = 1 - \exp\{-a_j\lambda(x_j)dx\},$$

which is equal to one minus the probability that j does not capture successfully-attacking particles. Here, the link function g_j depends on j because a_j is assumed to vary with j: $g_j : u \mapsto g_j(u) = 1 - \exp(-a_j u)$.

Introduction of covariates

Infection success depends on many local factors (Rapilly, 1991) such as plant characteristics (e.g. genotype, individual variations within a genetically homogeneous plantation, age, size), environmental variables (e.g. the soil and the climate, which can influence plant physiology), variations in source infectivity (some infectious plants may be more infectious than others because of a larger production of particles on this plant, or a larger local population of vectors for a vector-borne disease).

These factors can be introduced in the model in the effects a_j and c_i. These effects may depend on locations x_j and x_i, respectively, or may explicitly

[1] The multiplicative constant $1/2\pi\beta^2$ in Equation (2.3) ensures that f is a probability density function over $\mathbb{R}^2$.

depend on covariates (e.g. soil composition). Section 2.4 shows an example where the effects c_i are modeled as a log-normal random field. Section 2.5 shows an example where the effects a_j and c_i are modeled as deterministic and parametric functions of covariates characterizing susceptible and infectious units.

2.1.3 Deriving the fine-scale model to build models adapted to various disease-observation scales

The fine-scale model proposed above describes the presence/absence of a disease on infinitesimal units. In practice, various sorts of disease measures corresponding to various observation scales are encountered[2]. In the following, we show how the fine-scale model can be derived to obtain models adapted to the observation scale. It has however to be noted that the observation units are supposed to be small enough to consider that the local infectious potential is constant within any unit.

Counting the lesions on susceptible units

Consider a susceptible unit j with area s_j and central point x_j (to avoid additional notation, similar notation are used to denote infinitesimal units in the fine-scale model and the observation units in the larger-scale models). Suppose that each successfully-attacking particle generates a lesion on the susceptible unit, and that the success probability of any attack is constant and equal to a. By using the *Poisson specification* made above, j captures a Poisson number of particles with mean $s_j\lambda(x_j)$, and the number $N_{j,t+1}$ of lesions generated at time $t+1$ from the particles is then Poisson distributed with mean $s_j a\lambda(x_j)$, i.e.:

$$P(N_{j,t+1}=n)=\exp\{-s_j a\lambda(x_j)\}\frac{(s_j a\lambda(x_j))^n}{n!}. \tag{2.4}$$

Note that in this subsection and the following ones, the probabilistic conditioning is omitted to simplify notation; e.g. $P(N_{j,t+1}=n \mid \lambda(x_j), \mathbf{S}_{j,t}=1)$ is simply denoted by $P(N_{j,t+1}=n)$.

If lesions can be identified, then the disease measure can be lesion counts, and the log-likelihood used to estimate the parameters is:

$$\sum_{\substack{j:\delta_{jt}\delta_{j,t+1}=1\\ \mathbf{S}_{jt}=1}} \log P(N_{j,t+1}=n_{j,t+1})$$

$$= \sum_{\substack{j:\delta_{jt}\delta_{j,t+1}=1\\ \mathbf{S}_{jt}=1}} n_{j,t+1}\log\{s_j a\lambda(x_j)\} - n_{j,t+1}a\lambda(x_j) - \log(n_{j,t+1}!),$$

[2] A review on disease intensity measurements in plant epidemiology and their relationships was made by McRoberts et al. (2003).

where $n_{j,t+1}$ are the observed values of $N_{j,t+1}$, and the summation is performed on units observed at times t and $t+1$ (i.e. $\delta_{jt}\delta_{j,t+1} = 1$) and healthy at time t (i.e. $\mathbf{S}_{jt} = 1$).

Remark 1: the sum in this log-likelihood is computed only for healthy units at time t. However, already infected units at time t could also be considered in the log-likelihood. Indeed, they can be affected by particles dispersed from the infectious units and, consequently, they can bring information on the parameters. However, for taking into account this information, the autoinfection (i.e. the process of infection of a host by itself) must be modeled as well as its interaction with the alloinfection (i.e. the process of infection of a host by other hosts). This point is not tackled here.

Remark 2: here and thereafter, we assume that the observation units are small enough to consider that the local infectious potential is constant within any unit. To relax this assumption, and using the *Poisson specification*, the term $s_j a\lambda(x_j)$ should be replaced in Equation (2.4) by the integral of $x \mapsto a\lambda(x)$ over the area covered by j.

Measuring the infected areas of susceptible units

When lesions are hardly distinguishable, counting lesions is impossible and one relies on severity measures, the most classical one being the infected area on the susceptible unit, say S_{jt} for unit j at time t. Suppose that the area $S_{j,t+1}$ is a random variable depending on $N_{j,t+1}$ and s_j: $S_{j,t+1} = F(N_{j,t+1}, s_j)$, where F is a random function which may be selected empirically and/or based on mechanistic assumptions about the disease. For example $S_{j,t+1}$ can be derived from a spatial Boolean process (Stoyan et al., 1995; Molchanov, 1997) if lesions are assumed to be independent surface areas. The density probability function of $S_{j,t+1}$ is

$$p(S_{j,t+1}) = \sum_{N=0}^{\infty} h(S_{j,t+1} \mid N, s_j) \frac{(s_j a\lambda(x_j))^N}{N!} \exp(-s_j a\lambda(x_j))$$

where $h(\cdot \mid N, s)$ is the conditional density probability function of $F(N, s)$ given N and s. The log-likelihood is then:

$$\sum_{\substack{j:\delta_{jt}\delta_{j,t+1}=1 \\ \mathbf{S}_{jt}=1}} \log p(S_{j,t+1}).$$

Observing the presence/absence of the disease on susceptible units

The easiest way to measure the disease on a given susceptible unit is often to observe whether it is present or not on the unit. The absence of the disease at time $t+1$ corresponds to the event $\{\mathbf{S}_{j,t+1} = 1\}$, the presence of the disease at time $t+1$ corresponds to the event $\{\mathbf{S}_{j,t+1} = 0\}$. The disease is not on

unit j at time $t+1$ if no particle succeeds in infecting j, which occurs with probability $P(\mathbf{S}_{j,t+1}=1) = P(N_{j,t+1}=0) = \exp(-s_j a\lambda(x_j))$ because $N_{j,t+1}$ follows a Poisson distribution with mean $s_j a\lambda(x_j)$; see Equation (2.4). Thus $\mathbf{S}_{j,t+1}$ is Bernoulli-distributed with success probability $\exp(-s_j a\lambda(x_j))$.

In this case, we obtain the log-likelihood:

$$\sum_{\substack{j:\delta_{jt}\delta_{j,t+1}=1 \\ \mathbf{S}_{jt}=0}} (1-\mathbf{S}_{j,t+1}) \log\{1-\exp(-s_j a\lambda(x_j))\} - \mathbf{S}_{j,t+1} s_j a\lambda(x_j). \qquad (2.5)$$

This formula is similar to the log-likelihood (2.2), with $g_j(u) = 1-\exp(-s_j a u)$ depending on the unit characteristics s_j and a.

Counting the infected sub-units of susceptible units

Sometimes, the observation unit (e.g. a plant) is split into m_j sub-units (e.g. the leaves) and the disease measure is the number of infected sub-units M_{jt}. Let $\mathbf{S}_{jkt}$ denote the sanitary status of sub-unit k of unit j at time t. Suppose that unit j is completely healthy at time t, i.e. $\mathbf{S}_{jkt}=1$ for all $k=1,\ldots,m_j$. Following the paragraph above, $\mathbf{S}_{jk,t+1}$ is Bernoulli-distributed with probability $\exp(-s_{jk}a\lambda(x_j))$, where s_{jk} is the area of sub-unit k. All sub-units of unit j are assumed to be submitted to the same infectious potential $\lambda(x_j)$. In addition, $\mathbf{S}_{jk,t+1}$, $k=1,\ldots,m_j$, are independent because under the Poisson assumption the potential attacks of the sub-units are independent. This setting yields the following:

- In the case where the sub-unit areas are the same (i.e. $s_{jk} = s_j/m_j$), $M_{j,t+1}$ follows a binomial distribution with size m_j and success probability $p_j = 1-\exp(-s_j a\lambda(x_j)/m_j)$. Thus the log-likelihood is:

$$\sum_{\substack{j:\delta_{jt}\delta_{j,t+1}=1 \\ \mathbf{S}_{jt}=1}} \log\binom{m_j}{M_{j,t+1}} + M_{j,t+1}\log p_j - (m_j - M_{j,t+1})\log(1-p_j), \qquad (2.6)$$

  where $\binom{m}{M} = m!/\{M!(m-M)!\}$.
- In the case where the sub-unit areas are different and cannot be measured individually, one can for example consider the areas as independently and identically distributed with probability density function h_s. Then, $1-\mathbf{S}_{jk,t+1}$ is Bernoulli-distributed with success probability $p_j = \int_s \{1-\exp(-sa\lambda(x_j))\} h_s(s)ds$, $M_{j,t+1}$ follows a binomial distribution with size m_j and success probability p_j, and the log-likelihood can be written as in (2.6) by replacing p_j by its new expression.

2.1.4 Implications

Deriving models adapted to data from a fine-scale model allows (i) the estimation of biologically relevant parameters, those defined in the fine-scale model,

(ii) and the comparison / combination of experiments performed at different scales[3].

Concerning point (i), for each constructed model, we have written a log-likelihood upon which the inference on the parameters can be based. In particular, inference on the parameters included in the infectious potential λ is possible in each case since λ appears in each expression of the log-likelihood. Moreover, each context offers the possibility to infer other parameters that are specific to the context: for example, the parameters which could link the receptor and source effects (a_j and c_i) to covariates (Section 2.1.2), or the parameters which could be involved in the random function F linking the lesion count to the infected area (Section 2.1.3).

The aggregative approach presented in this section is applicable / generalizable to different types of mechanisms, different types of data, different mathematical representations of the mechanisms, and different probabilistic representations of the observation processes. This point is the core of this chapter and Chapter 4, which deals with mechanistic-statistical modeling.

2.2 Anisotropic dispersal

2.2.1 Models

In the models under consideration here, deposit locations of particles emitted by a point source at the origin form a spatial Poisson point pattern with inhomogeneous intensity decreasing along radial directions. In addition the decrease along radial directions is anisotropic, i.e. it varies with respect to the direction.

The models of Klein et al. (2003), Stockmarr (2002) and Tufto et al. (1997) based on 3D spatial Brownian motions describing spore transports allow the introduction of anisotropy by adding trends to the horizontal components of the Brownian motions. Another set of approaches consists in incorporated von Mises functions (commonly used to describe the distributions of circular data; see Fisher, 1995) into dispersal kernels to achieve anisotropy. Thus, in the model of Herrmann et al. (2011); Wagner et al. (2004); Wälder et al. (2009), the Euclidean distance from the source is multiplied by a function of the direction, typically a von Mises function. The approach that we followed in Soubeyrand et al. (2007c, 2008a, 2009b) and Rieux et al. (2014) also uses von Mises functions, or more generally functions defined on the circle, but these functions are used to modify the parameter of the dispersal kernel as

[3] In an other framework, namely the mapping of weeds, we combined three types of weed data to interpolate the spatial intensity function of weeds; see Bourgeois et al. (2012). The three types of data are counts of weeds in small quadrats, counts censored by interval in large quadrats, and areas of high intensity of weeds. A unique intensity function governs the probabilistic laws of the three types of data.

well as the source strength. This approach, presented below, leads to a double anisotropy in the dispersal of particles.

Anisotropizing dispersal kernels

Consider a point source located at the origin of the planar space $\mathbb{R}^2$. Suppose that the deposit locations of particles form a Poisson point process with intensity at location $x \in \mathbb{R}^2$ proportional to the isotropic exponential dispersal kernel (introduced in Section 2.1.2):

$$f_{iso}(x) = \frac{1}{2\pi\beta_{iso}^2} \exp\left(-\frac{||x||}{\beta_{iso}}\right),$$

where $||\cdot||$ is the Euclidean distance and $\beta_{iso} > 0$ is called dispersal parameter. The multiplicative constant $1/2\pi\beta_{iso}^2$ ensures that f_{iso} is a probability density function over $\mathbb{R}^2$.

The isotropic kernel f_{iso} has been generalized by Soubeyrand et al. (2007c) into a doubly anisotropic exponential dispersal kernel:

$$f(x) = \frac{\alpha(\phi)}{\beta(\phi)^2} \exp\left(-\frac{||x||}{\beta(\phi)}\right), \tag{2.7}$$

where ϕ is the angle made by x, $\alpha(\cdot)$ is a circular probability density function (defined over $[0, 2\pi)$ and whose integral over $[0, 2\pi)$ is one) and $\beta(\cdot)$ is a positive circular function (defined over $[0, 2\pi)$). It can be easily verified that f is, like f_{iso}, a probability density function over $\mathbb{R}^2$. $\alpha(\phi)$ gives the density of deposit locations of particles in direction ϕ: the larger $\alpha(\phi)$, the more deposited particles in direction ϕ. $\beta(\phi)$ is the dispersal parameter in direction ϕ: the larger $\beta(\phi)$, the further in expectation particles are deposited in direction ϕ.

Other isotropic kernels can obviously be *anisotropized* in the same way. For example, Soubeyrand et al. (2009b) proposed to generalize the isotropic Gaussian kernel into:

$$f(x) = \frac{\alpha(\phi)}{\beta(\phi)^2} \exp\left(-\frac{||x||^2}{2\beta(\phi)^2}\right),$$

and the isotropic geometric kernel into:

$$f(x) = \frac{\alpha(\phi)(\gamma-1)(\gamma-2)}{\beta(\phi)^2} \left(1 + \frac{||x||}{\beta(\phi)}\right)^{-\gamma}.$$

where γ is an additional shape parameter in the geometric kernel which could also be replaced by a positive circular function. Other classical dispersal kernels *anisotropized* in the same way are presented in Rieux et al. (2014).

Specifying the anisotropy functions using von Mises functions

It was first proposed in Soubeyrand et al. (2007c) to specify α and β using von Mises functions, which are regular and unimodal functions defined over the circle:

$$\alpha(\phi) = \frac{1}{2\pi I_0(\sigma_\alpha)} \exp\{\sigma_\alpha \cos(\phi - \mu_\alpha)\}$$
$$\beta(\phi) = \frac{\beta_0}{2\pi I_0(\sigma_\beta)} \exp\{\sigma_\beta \cos(\phi - \mu_\beta)\}$$

where $\mu_\alpha \in [0, 2\pi)$ is the mean dispersal direction and $\sigma_\alpha \geq 0$ measures the dispersion around μ_α; $\mu_\beta \in [0, 2\pi)$ is the direction along which particles are deposited the furthest in expectation, $\sigma_\beta \geq 0$ measures the dispersion of dispersal distances around μ_β, and $\beta_0 > 0$ is a multiplicative constant measuring how far from the source particles are deposited; and $I_0(\sigma) = (2\pi)^{-1} \int_0^{2\pi} \exp\{\sigma(\theta - \mu)\} d\theta$. Obviously, other parametric forms than the von Mises form could be used in the same way, e.g. mixtures of von Mises functions, cardioid functions, wrapped Cauchy or normal functions; see Fisher (1995).

Specifying the anisotropy functions using circular Gaussian random processes

To take into account rougher anisotropies, Soubeyrand et al. (2008a) used circular Gaussian random processes (GRP) to specify the anisotropy functions. The anisotropy functions α and β are defined by:

$$\alpha(\phi) = \frac{1}{\lambda_0} \exp(Z_\alpha(\phi)) \tag{2.8}$$
$$\beta(\phi) = \exp(Z_\beta(\phi)), \tag{2.9}$$

where λ_0 is a multiplicative constant such that α is a density probability function over the $[0, 2\pi)$, Z_α and Z_β are the realizations of two independent stationary circular GRPs with means $\eta_\alpha \in \mathbb{R}$ and $\eta_\beta \in \mathbb{R}$, variances κ_α^2 and κ_β^2 and circular correlation functions (CCF) C_α and C_β.

In this work, one of the concerns was the roughness of the circular GRPs and, consequently, the shape of the CCFs. Therefore, several CCFs were proposed and a model selection procedure was applied to select the most appropriate CCF among the proposed ones. CCFs are generally obtained by using the chordal distance as the argument of a correlation function defined over $\mathbb{R}$: let C denote a valid correlation function over $\mathbb{R}$, then $\phi \mapsto C(2\sin(\phi/2))$ is a valid correlation function on the circle with radius one ($2\sin(\phi/2)$ is the chordal distance between two points belonging to the unit circle and separated by the angle ϕ). Two CCFs built in such a way were considered: the first one

($\phi \mapsto \exp\{-2\sin(\phi/2)/\alpha\}$, $\alpha > 0$) was obtained from the exponential correlation function and the other one ($\phi \mapsto \exp[-\{2\sin(\phi/2)/\alpha\}^{\gamma}]$, $\alpha > 0$, $\gamma > 0$) was obtained from the exponential-power correlation function. In addition, Soubeyrand et al. (2008a) built the following CCF without resorting to this technique:

$$C(\phi) = 1 - \sin^{\delta}(\phi/2), \qquad \forall \phi \in [0, 2\pi), \tag{2.10}$$

where $\delta \in (0, 2)$ is between zero and two to get a valid (positive definite) CCF. Since this new CCF was not obtained using the chordal distance as the argument of a correlation function defined on the line, its validity had to be checked (i.e. the positive definiteness of the CCF had to be shown). Using theory on positive definite functions presented in Sasváry (1994), Soubeyrand et al. (2008a) derived the Bochner's theorem for the circle that provides the class of positive definite functions defined over the circle. CCF (2.10) was shown to belong to this class if parameter δ was in the interval $(0, 2)$. Several properties of CCF (2.10) were obtained. In particular, this CCF is continuous but not differentiable at the origin. Thus, a GRP with CCF (2.10) is mean square continuous but not mean square differentiable. Therefore, such a GRP is a rather rough process.

2.2.2 Estimation

For particles whose sizes are measured in micrometers (e.g. spores, pollen grains) and particles which are not detected easily in fields, orchards or forests even if they are visible by eyes (e.g. seeds), we do not generally observe the point pattern formed by the deposited particles, but, we may observe numbers of particles collected in traps, numbers of symptoms on plants for diseases disseminated with spores, or presence–absence of seedlings. By following the aggregative approach described in Section 2.1, the observed data (e.g. counts) can be modeled like random variables whose distributions depend on the dispersal kernel. Then, a likelihood may be written and inference based on this likelihood may be performed.

For example, Soubeyrand et al. (2007c) considered counts of wheat plants infected by the yellow rust and fitted to data the kernel (2.7) incorporating von Mises functions by specifying a binomial observation process and using a Newton-Raphson algorithm to maximize the likelihood.

Exploiting the same data set, Soubeyrand et al. (2008a) fitted to data the kernel (2.7) incorporating circular GRPs by specifying a binomial observation process and using a Markov chain Expectation-Maximization algorithm (Wei and Tanner, 1990) to obtain maximum likelihood estimates of parameters and latent variables of their hierarchical model; see Figure 2.2.

Soubeyrand et al. (2009c) studied the spatio-temporal spread of powdery mildew infecting *plantago lanceolata*, the data consisting in presence patterns of powdery mildew in a set of about 4000 host plant patches. They included the kernel (2.7) incorporating von Mises functions in their spatio-temporal model

and fitted this model to data using a Markov chain Monte Carlo algorithm (Robert and Casella, 1999) to assess posterior distributions of parameters. This study is presented in Section 2.5.

Rieux et al. (2014) assessed the dispersal of spores of the wind-dispersed banana plant fungus *Mycosphaerella fijiensis* by estimating the parameters of several dispersal kernels (i.e. exponential, geometric, Wald and power-exponential) *anisotropized* with von Mises functions. In this case, data were counts of lesions on banana leaves. However, these counts were generally noised by lesions due to sources of spores located outside the experimental plot. Therefore, genetic analyses of subsets of lesions were performed to distinguish (i) lesions due to the source of spores voluntary introduced at the center of the experimental plot, and (ii) lesions due to exogenous sources of spores (the source had a specific genotype, say G_s, not represented among the exogenous sources). Thus, three types of lesion counts were available for each sampled leaf: the total number of lesions, the number of genotyped lesions which was generally lower than the total number of lesions, and the number of genotyped lesions with genotype G_s. Here the observation process was modeled using an hypergeometric distribution depending on the anisotropic dispersal kernel. Parameters were estimated by maximizing the likelihood, and the Akaike criterion was used to investigate the significance of the anisotropy and to select the best dispersal kernel.

2.2.3 Application

The yellow rust of wheat is an airborne plant disease caused by the fungus *Puccinia striiformis*. This fungus forms lesions on wheat leaves. The disease is spread by spores produced by the lesions and transported in the air mostly by wind and rain.

For gaining insight into the anisotropic spread of yellow rust in large field plots, the following experiment was carried out in 2002. Wheat plants infected with yellow rust were settled in a source plot ($2.5 \times 1.6\text{m}^2$) located more or less at the center of a 75,000m^2 field of healthy wheat plants. Five days later, infected leaves were counted for 187 trap plots ($1 \times 1\text{m}^2$) located at the nodes of a regular grid covering the field area. In each panel of Figure 2.1, the gray point indicates the location of the source plot, and figures are at the locations of the trap plots. On the left panel, the figures are the counts of infected leaves in the trap plots. On the right panel, the figures are leaf density levels in the trap plots; the levels rank from 1 to 7 and correspond to different total counts of leaves (see the correspondences at the top-left). Thus, for each trap plot $i \in \{1, \ldots, n = 187\}$, we observe the location $x_i \in \mathbb{R}^2$ of its center, the count $y_i \in \mathbb{N}$ of infected leaves, and the total count $q_i \in \mathbb{N}$ of leaves. By convention, the source plot is located at the origin.

The count of infected leaves Y_i among the q_i leaves in trap plot i is supposed to be drawn from a binomial distribution with size q_i and probability $p(x_i)$:

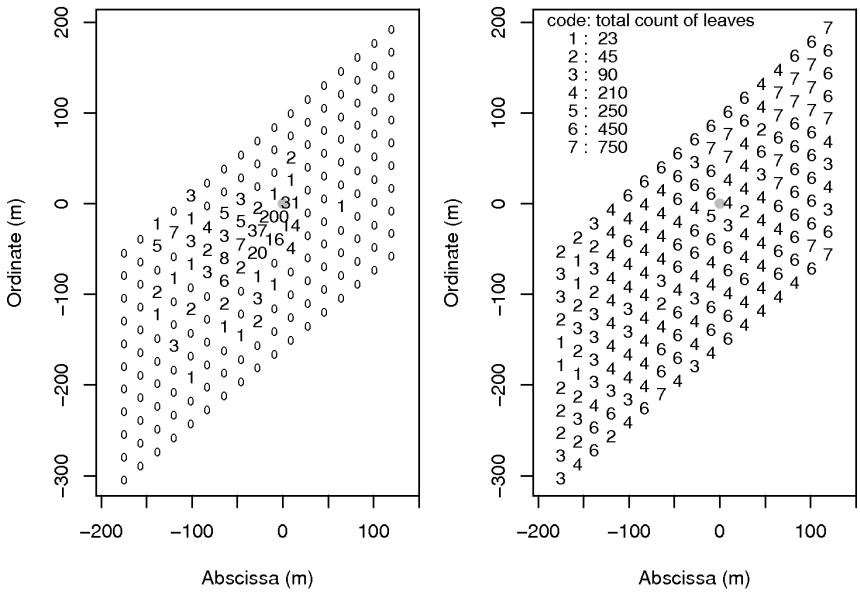


Fig. 2.1. Data maps. Left: counts of infected leaves in the 1m^2-trap plots. Right: levels of the leaf density in the trap plots. In each panel, the gray point indicates the location of the source plot.

$$Y_i \sim \text{Binomial}\{q_i, p(x_i)\}, \tag{2.11}$$

where

$$p(x) = 1 - \exp\{-\lambda_0 f(x)\}, \tag{2.12}$$

where f is the exponential anisotropic kernel given by Equation (2.7). The link function $u \mapsto 1 - \exp(-u)$ is obtained by applying the aggregative approach introduced in Section 2.1 leading to a model whose response variable is the count of infected sub-units in a sampling unit[4]. In this application, the anisotropy functions incorporated into f were specified using circular Gaussian random processes (GRP; see Equations (2.8–2.9)) characterized by circular correlation functions (CCF) satisfying Equation (2.10). For assessing the suitability of this anisotropy model, we compared it to three other models: the anisotropic model including von Mises functions, the model including

[4] See specifically the paragraphs entitled *Observing the presence/absence of the disease on susceptible units* and *Counting the infected sub-units of susceptible units*.

two GRPs with exponential CCFs, and the model including two GRPs with exponential-power CCFs (this model has two additional parameters compared to the three other models). The four models were compared using the Akaike's information criterion (AIC, Burnham and Anderson, 2002), the quadratic (or Brier) score and the spherical score (Gneiting and Raftery, 2005). The AIC is based on the likelihood function and is penalized by the number of parameters; the lower the AIC, the more suitable the model. The quadratic and the spherical scores are based on the probability distributions which are predicted for the observed variables; the higher these scores, the more suitable the model.

Table 2.1 shows the values of the criteria which were obtained for the four models. The modeling of the disease spread is clearly improved when GRPs are used instead of von Mises functions for modeling α and β. Among the models including the GRPs, the one with CCFs (2.10) is the more suitable. It is even slightly better than the model including the GRPs with exponential-power CCFs (which has more parameters).

Table 2.1. Model comparison. Number of parameters (N), value of the log-likelihood (loglik), Akaike's information criterion (AIC), quadratic score and spherical score obtained when λ and μ are proportional to von Mises functions, or when they are functions of GRPs with various CCFs.

Model	N	loglik	AIC	quadratic	spherical
Von Mises	6	-280.2	572.4	-61.2	148.4
GRP with CCF given by Eq. (2.10)	6	-104.2	220.4	-46.4	156.0
GRP with exponential CCF	6	-112.5	237.0	-49.3	154.1
GRP with exponential-power CCF	8	-104.2	224.4	-46.6	155.9

Figure 2.2 shows Monte Carlo estimates of the anisotropy functions (left panels) and the probabilities for wheat leaves to be infected. These plots highlight the irregularity of the particle dispersal and the resulting disease spread.

If the global trend, i.e. the deterministic component, of the spread is commonly associated with the mean wind direction and speed (Aylor, 1990; McCartney and Fitt, 2006), the local fluctuations, i.e. the stochastic component, are still poorly understood. Finely describing the irregular patterns of particle dispersal and disease spread should be valuable for better understanding the processes underlying these phenomena.

The advantage of the model proposed here is that local fluctuations are estimated and, consequently, can be analyzed together with meteorological variables such as wind, turbulence and humidity (meteorological measurements are not available in the experiment analyzed above). Turbulence, which is involved in release, escape from the canopy, transport and deposit of spores (Aylor, 1999; Aylor and Flesch, 2001), is especially expected to play a role in

the irregularity of dispersal patterns. Turbulence is for instance a key component to describe the random paths of spores in the Lagrangian stochastic simulation model (Aylor and Flesch, 2001; Wilson, 2000). In practice, we could plan in future works to analyze the possible link between strong wind gusts and the peaks displayed in Figure 2.2. It seems however difficult to link a given wind gust with a given peak. Instead of focusing on such specific events, it may be more pragmatic to link irregularity characteristics of the model output with general characteristics of meteorological variables. For example we could plan to analyze the possible link between the frequency of changes in wind direction and the roughness of the intensity function α and the mean distance function β (the roughness is measured by the parameter of the correlation function).

Characterizing the irregularity of the dispersal should also be valuable in the modeling of epidemics. An epidemic is a nonlinear system made of many propagation events, as the one studied in this paper, repeated in space and time. It is important to catch the stochasticity of single propagation events because it can affect the global dynamics of the epidemic (Rohani et al., 2002). It would be especially interesting to investigate what sort of patterns will be obtained after several generations given the irregularity of the intensity function α and the mean distance function β, and their discrepancy.

2.2.4 Side topic 1: sequential sampling for estimating anisotropy

Anisotropy is observed in dispersal patterns occurring for a wide range of biological systems. While dispersal models more and more often incorporate anisotropy, the sampling schemes required to collect data for validation usually do not account for the anisotropy of dispersal data. In Soubeyrand et al. (2009b), using the anisotropic model presented in Section 2.2.1, we carried out a study aimed at recommending an appropriate sampling scheme for anisotropic data. In a first step, we showed with a simulation study that prior knowledge of dispersal anisotropy can be used to improve the sampling scheme. One of the main guidelines to be proposed is the orientation of the sampling grid around the main dispersal directions. In a second step, we proposed a sequential sampling procedure used to automatically build anisotropic sampling schemes adapted to the actual anisotropy of dispersal.

2.2.5 Side topic 2: 3D anisotropy

In most of the propagation studies in plant epidemiology, the spread of the disease is represented in the 2D-horizontal plane. In Soubeyrand et al. (2008b), we analyzed the spread of a disease in a wheat field where observations were made at different times, at different locations in the horizontal plane, and at different heights, i.e. leaf layers, in the vegetal cover. Here, the vertical dimension was viewed as a discrete space consisting of the ground, the different

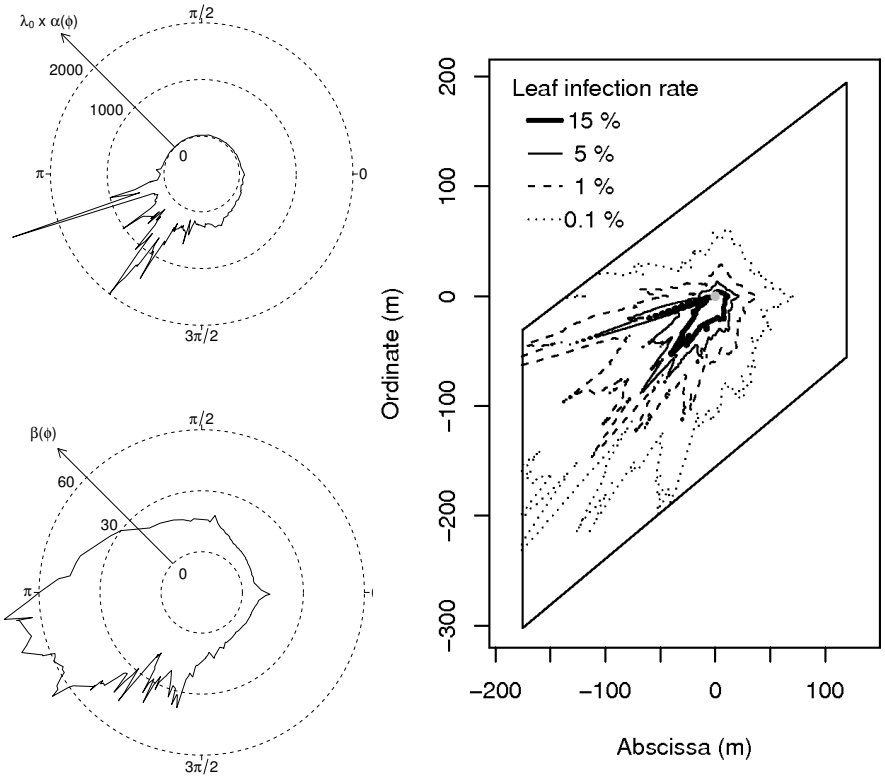


Fig. 2.2. Left: Monte-Carlo estimates of the anisotropy functions α (up to a multiplicative constant λ_0; Eq. (2.8)) and β (Eq. (2.9)) based on circular GRPs. Monte-Carlo estimates of the resulting probabilities (in %) for wheat leaves to be infected (Eq. (2.12)).

leaf layers[5], and the air above the vegetal cover. To analyze the horizontal and vertical spread of the disease, we built dispersal kernels in the 3D space. These dispersal kernels inherently incorporate anisotropy because the structure of the space is different in the horizontal and the vertical dimensions.

[5] It has to be noted that for adult wheat plants, the leaf layers are usually unambiguously identified even if the plants continue to grow.

Combining vertical and horizontal dispersal functions

Let i and j denote two host units whose locations in the horizontal plane $\mathbb{R}^2$ are x_i and x_j and whose locations in the vertical space {1,2,...,K} are z_i and z_j, where K is the number of host layers in the vertical dimension (layer 1 is the bottom host layer; layer K is the top host layer).

We modeled the dispersal function of particles $(i, j) \mapsto p(i, j)$ by combining a horizontal dispersal function (HDF) f and a vertical dispersal function (VDF) $v(\cdot, \cdot)$. The HDF f describes the transport of particles in the horizontal plane above the top layer and is analogous to 2D dispersal kernels presented in the previous sections. The quantity $f(x_j - x_i)$ is the probability density for a particle that reached the air at x_i to be definitely re-introduced into the cover at x_j. The VDF $v(\cdot, \cdot)$ governs the transports of particles in the vertical direction between the host layers, the ground (denoted by G) below layer 1, and the air (denoted by A) above layer K. In particular, if i and j are located at the same site in the horizontal plane (i.e. $x_i = x_j$), $v(z_i, z_j)$ is the probability for a particle released by unit i at layer z_i to be deposited on unit j at layer z_j. Besides, $v(z_i, \mathrm{A})$ (resp. $v(z_i, \mathrm{G})$) is the probability for a particle released by unit i at leaf layer z_i to reach the air above layer K (resp. the ground).

Combining f and v yields the following expression for the dispersal function p:

$$p(i,j) = \begin{cases} v(k_i, k_j) & \text{if } x_i = x_j \\ v(k_i, \mathrm{A}) f(x_j - x_i) \frac{v(K,k_j)}{1 - v(K,\mathrm{A})} & \text{if } x_i \neq x_j. \end{cases} \tag{2.13}$$

The term[6] $v(K, k_j)/\{1 - v(K, \mathrm{A})\}$ is the probability for a particle which is re-introduced into the cover at location z_i to be deposited on unit j at layer k_j.

We proposed two constructions for v which do not require physical or biological input variables (in contrast with the sophisticated model proposed by Koizumi and Kato, 1991), but offer more flexibility than the one-parameter vertical kernel of Djurle and Yuen (1991). The first construction for v was inspired by the Beer-Lambert law, which is used in optics to assess the intensity of the light after passing through a material. The second construction was based on a discrete Markov chain. In both constructions, disease severity is assumed to be locally constant in the horizontal plane and, consequently, ingoing and outgoing horizontal flux of particles at a given layer are assumed to be equal.

[6] In this term, the first argument of the numerator $v(K, k_j)$ is K because the particle is re-introduced by above. The denominator appears because x_j is the location where the particle is definitely re-introduced into the host layers and, consequently, the probability for the particle initially at leaf layer K to be deposited at leaf layer k_j is conditional on the fact that the particle cannot reach again the air above the cover.

Here we only show how to construct the VDF v using a discrete Markov chain. Particles are assumed to move both up and down until they are deposited on a host layer or absorbed by the air A or by the ground G. We assume that particle movements obey a stationary Markov chain, where a particle can be in one of the following states:

1. at layer k in $\{1, \ldots, K\}$, but not deposited on a host unit;
2. at layer k in $\{1, \ldots, K\}$ and deposited on a host unit (a star will be used to denote these states);
3. in the air A above the top layer;
4. deposited on the ground G.

State A, state G and states where the particle is deposited on a host unit are absorbing states. Different specifications for the transition probabilities between the non-absorbing and absorbing states may be proposed; Figure 2.3 shows an example using three parameters. Once the transition probabilities are specified, the expression of v is obtained by computing the limiting transition probabilities of the absorbing states conditional on the initial state.

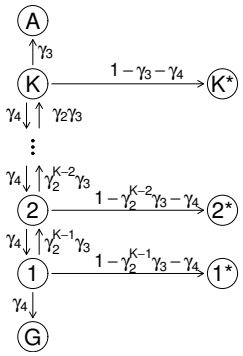


Fig. 2.3. Markov chain used to construct a model for the vertical dispersal function (VDF) of particles. A star is used to mark the absorbing layer states. The transition probabilities for this Markov chain are defined with three parameters γ_2, γ_3 and γ_4, which have to satisfy the following constraints: $0 < \gamma_2, \gamma_3, \gamma_4 < 1$ and $\gamma_3 + \gamma_4 < 1$. Heuristically for this specification, γ_4 is related to the gravity force which is supposed to be constant whatever the leaf layer and, because $\gamma_2 < 1$, ascending of particles is most probable at upper leaf layers than at lower leaf layers.

Application

The 3D dispersal kernels introduced above were incorporated into a spatio-temporal model of the spread of yellow rust (a fungal disease) in a wheat field. This model was developed to analyze experimental data shown in Figure 2.4 (top). A source of disease was settled at the center of a healthy wheat field

and the disease severity (i.e. the proportion of the sporulating (or infectious) surface on wheat leaves) was measured across time and at different leaf layers. Since wheat plants grown during the sampling period, the disease was measured at the nodes of a time-varying 3D-grid (at sampling time 5, leaf layer 1 disappeared and leaf layers 3 and 4 were generated).

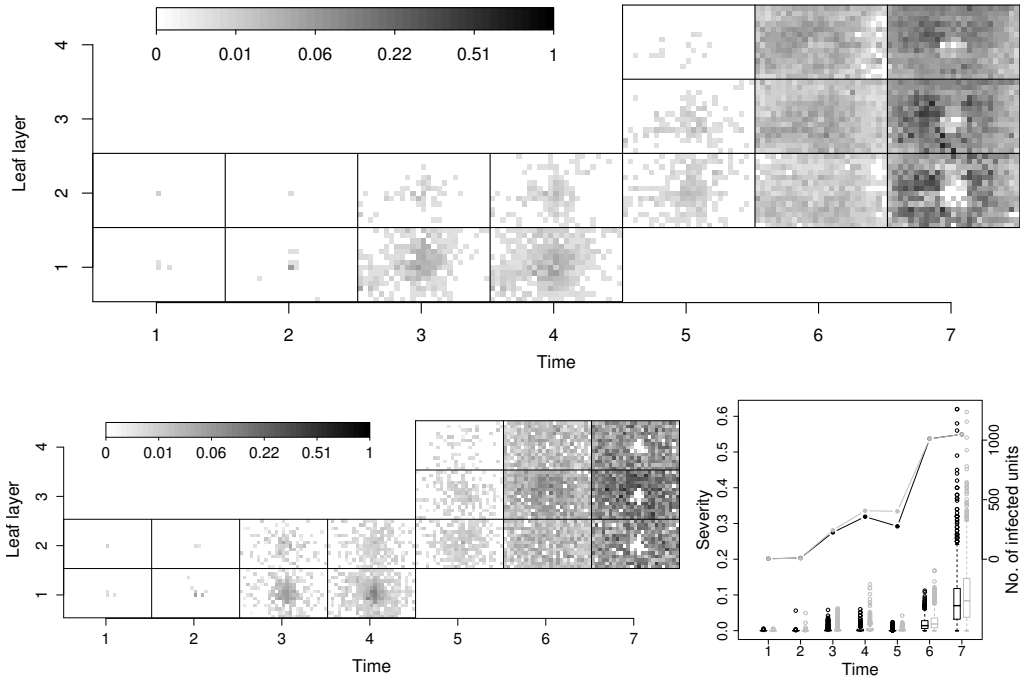


Fig. 2.4. Top panel: spatio-temporal evolution of the disease severity. Each rectangle provides, for a given time and a given leaf layer, the spatial variation of the disease severity. Bottom left: simulation of an epidemic using the estimated parameters, using the real data at time one as the initial state, and preserving the space-time structure of the vegetal cover. Bottom right: Evolutions in time of the severities (box-plots), and the number of infected units (lines). These elements are in black for the real data set and in grey for the simulated data set.

The spatio-temporal model is a two-stage model describing the joint distribution of the occurrence (new infection) and the severity of the disease: occurrence of the disease on a leaf is modeled at the first stage; severity of the disease is modeled at the second stage given disease occurrence. The model for disease occurrence was built with the aggregative approach of Section 2.1 (see specifically paragraph *Observing the presence/absence of the disease on susceptible units*) and depends on an infection potential where source strengths coincide with severities observed in the past. The model for disease severity is

more empirical: it is defined as a zero-inflated beta GLM whose explanatory variables include occurrence variables and the infection potential.

Parameter estimation was carried out with maximum likelihood and model selection (to choose the most appropriate horizontal and vertical dispersal functions) was performed with the Akaike criterion (AIC). The smallest AIC value was obtained with the VDF based on the Markov chain shown in Figure 2.3 and with the Cauchy HDF satisfying:

$$f(x) = \frac{1}{2\pi\gamma_1^2}\left(1 + \frac{||x||^2}{\gamma_1^2}\right)^{-3/2}.$$

Figure 2.4 (bottom left) shows a simulated epidemic under the estimated parameters. Qualitatively, this simulation reproduces some aspects of the real epidemic quite well. In particular, (i) the overall temporal trend of the disease spread (stagnation at time two, strong increase at time 3 and so on), (ii) the scattered spatial pattern of the disease at the first time steps, and (iii) the decrease in disease severity at the centre of the plot at time seven are well reproduced. Figure 2.4 (bottom right) provides a quantitative comparison between the simulated epidemic and the real epidemic shown in Fig. 2.4 (top). It shows the temporal variation of the severities (box-plots) and the number of infected units (lines) for both the real and simulated data (resp. in black and grey). Similar patterns are observed. Additional results carried out on series of simulations are provided by Soubeyrand et al. (2008b).

2.3 Group dispersal

2.3.1 Doubly inhomogeneous Neyman-Scott point process

Group dispersal occurs when several particles are released because of a wind gust, transported in the air into a more or less limited volume and deposited over a more or less limited area; see Figure 2.5.

In propagation models for airborne plant pathogens and plants, deposit locations of particles are usually assumed to be independently and identically drawn under the dispersal kernel. If group dispersal occurs, then the independence assumption is not valid anymore. To represent group dispersal, Soubeyrand et al. (2011) resorted to a hierarchical structure of dependence: at the first stage of the hierarchy, groups are independently dispersed; at the second stage, particles within each group are dispersed independently but conditionally on the group transport. The resulting model can be viewed as a Neyman-Scott point process[7] (Illian et al., 2008) with double inhomogeneity:

[7] A homogeneous Neyman-Scott point process is obtained by drawing a stationary Poisson point process, which forms the parent process, and by drawing clusters of daughter points around parents, where the cluster sizes are random and

(i) an inhomogeneity in the locations of cluster centers and (ii) an inhomogeneity in the spread of cluster points (here, a *cluster* is formed by the deposit locations of a *group* of particles released simultaneously by the same source)[8]. The model of Soubeyrand et al. (2011) and some of its properties are described in what follows. One of the properties, namely the concentration of particles, is not commonly studied in point process theory but is especially relevant in dispersal studies.

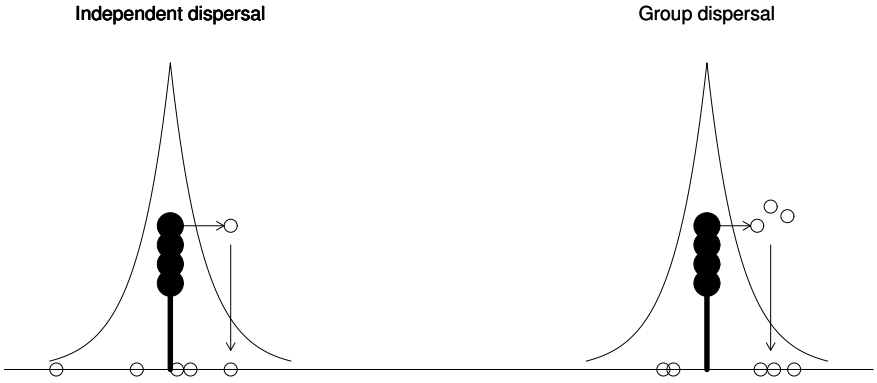


Fig. 2.5. Schematic representation of independent and group dispersal from a source of particles, e.g. an infected plant (black shape) releasing fungal spores (open circles). Independent dispersal: particles independently released and transported. Group dispersal: particles simultaneously released, and settling at different but positively correlated locations. The dispersal kernel in each case is represented by the solid curve.

Group dispersal model (GDM)

Consider a single point source of particles located at the origin of the planar space $\mathbb{R}^2$. The deposit location vector X_{jn} of the n-th particle of group j ($j \in \{1, \ldots, J\}$ and $n \in \{1, \ldots, N_j\}$) is assumed to satisfy

$$X_{jn} = X_j + B_{jn}(\nu||X_j||),$$

where $X_j = E(X_{jn} \mid X_j)$ is the final location vector of the center of group j, B_{jn} is a centered Brownian motion describing the relative movement of

the daughter points are scattered independently and with identical distribution around their respective parents. The Neyman-Scott process is formed by the daughter points only.

[8] The group dispersal model can also be viewed as a doubly stochastic point process model, also called spatial Cox process (Illian et al., 2008).

the n-th particle in group j with respect to the group center, ν is a positive parameter and $||\cdot||$ denotes the Euclidean distance.

The random variables J, N_j, X_j and the random processes $\{B_{jn} : n = 1, \ldots, N_j\}$ are mutually independent. The number of groups J is Poisson distributed with mean value λ. The N_j are independently drawn from the counting distribution p_{μ,σ^2} defined over $\mathbb{N}$ with mean and variance parameters μ and σ^2, respectively. The group center locations X_j are independently and identically drawn from the probability density function (p.d.f.) $f_{X_j} : \mathbb{R}^2 \mapsto \mathbb{R}_+$ (the consequence of this assumption is the inhomogeneity in the locations of cluster centers in the Neyman-Scott process). The function f_{X_j} can be characterized by features usually associated with classical dispersal kernels: for instance, a more or less steep decrease at the origin, a more or less heavy tail, and a more or less anisotropic shape. The Brownian motions B_{jn} defined over $\mathbb{R}^2$ are centered, independent and with independent components. They are stopped at time $t_j = \nu||X_j||$. The distance between the source and the location X_j is used, up to the scaling parameter ν, as a time surrogate. Thus, the further a group is transported, the most the particules forming the group are spread with respect to the group center (the consequence of this assumption is the inhomogeneity in the spread of cluster points in the Neyman-Scott process). The value of ν determines the strength of the relative spread from the group center. It follows that $B_{jn}(\nu||X_j||)$ follow independent and centered normal distributions with variance matrices $\nu||X_j||I$ where I is the 2×2 identity matrix. Figure 2.6, which shows a simulation of the GDM, clearly illustrates the existence of groups whose extents increase with distance from the point source.

Moments and generation of foci

Soubeyrand et al. (2011) specifically studied the ability of the model to generate secondary foci in a spatio-temporal context, that is to say when the group dispersal is repeated generation after generation (particles dispersed from a source become sources from which new particles are dispersed). This characteristic of the GDM was investigated by studying theoretical moments and analyzing simulated dynamics. Below, we only present the study of moments.

A first understanding of the ability of the GDM to generate multiple foci was achieved by studying the moments of the number of particles $Q(x + dx)$ deposited in the infinitesimal surface $x + dx$ centered around x. The expectation, variance and covariance satisfy:

$$\begin{aligned} E\{Q(x+dx)\} &= \lambda\mu f_{X_{jn}}(x)dx \\ V\{Q(x+dx)\} &= \lambda[\mu f_{X_{jn}}(x)dx + (\sigma^2+\mu^2-\mu)E\{\phi_{\nu,X_j}(x)^2\}(dx)^2] \\ \mathrm{cov}\{Q(x_1+dx), Q(x_2+dx)\} &= \lambda(\sigma^2+\mu^2-\mu)E\{\phi_{\nu,X_j}(x_1)\phi_{\nu,X_j}(x_2)\}(dx)^2, \end{aligned}$$

where $f_{X_{jn}}$ is the probability density function of X_{jn}, that is to say the dispersal kernel of particles, which is equal to:

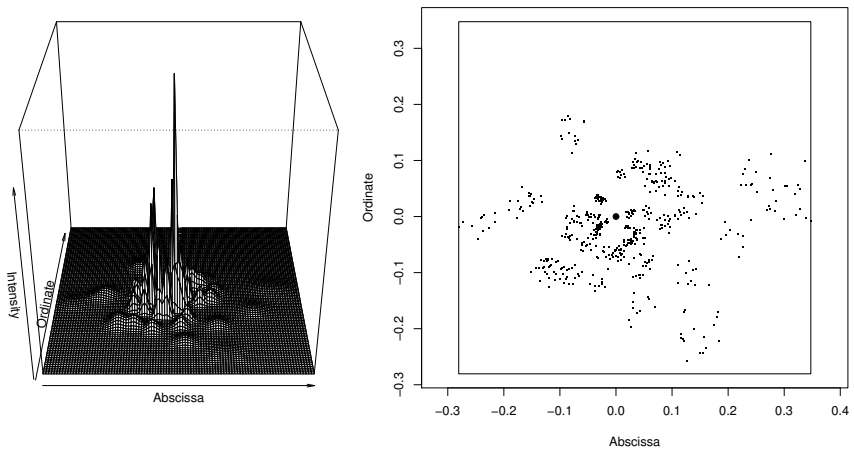


Fig. 2.6. Simulation of the group dispersal model with a single point source. Left: conditional density of the deposit location of a particle whose group it belongs to is unknown given the locations of the group centers. Right: deposit locations of particles obtained under the density shown on the left; the large dot indicates the location of the point source.

$$\begin{aligned} f_{X_{jn}}(x) &= \int_{\mathbb{R}^2} f_{X_{jn}|X_j}(x \mid y) f_{X_j}(y) dy \\ &= \int_{\mathbb{R}^2} \phi_{\nu,y}(x) f_{X_j}(y) dy, \end{aligned}$$

with $\phi_{\nu,y}(x) = \frac{1}{2\pi\nu||y||} \exp\left(-\frac{(x-y)'(x-y)}{2\nu||y||}\right)$, $E\{\phi_{\nu,X_j}(x)^2\} = \int_{\mathbb{R}^2} \phi_{\nu,y}(x)^2 f_{X_j}(y) dy$ and $E\{\phi_{\nu,X_j}(x_1)\phi_{\nu,X_j}(x_2)\} = \int_{\mathbb{R}^2} \phi_{\nu,y}(x_1)\phi_{\nu,y}(x_2) f_{X_j}(y) dy$. It has to be noted that for a counting distribution over $\mathbb{N}$ characterized by mean $\mu > 0$ and variance σ^2, the quantity $\sigma^2 + \mu^2 - \mu$ is positive; it is zero if and only if $\mu = 0$ (which implies that $\sigma^2 = 0$) or $(\mu, \sigma) = (1, 0)$. This implies that the covariance given above for the GDM is non-negative.

These moments can be used to compare point patterns obtained under the GDM with $\mu = 1$ and the GDM with $\mu = 1$ and $\sigma = 0$, which is an independent dispersal model (IDM; every group has size exactly equal to one). The extra-variance and the positive spatial covariance characterizing the GDM induce the occurrence of aggregates in space, while such aggregates are not expected under the IDM ($V_{\mathrm{IDM}}\{Q(x + dx)\} = \lambda f_{X_{jn}}(x)dx$ and $\mathrm{cov}_{\mathrm{IDM}}\{Q(x_1 + dx), Q(x_2 + dx)\} = 0$). These aggregates are at the origin of secondary foci visible in a spatio-temporal context (i.e. when deposited particles become sources of particles at the next time step) without resorting to heavy-tailed dispersal kernels or to spatial heterogeneity.

Furthest particle and concentration of particles

The distribution of the furthest deposited particle is of interest because it determines the spreading speed (or invasion speed) in a spatio-temporal context: the more concentrated the particles, the lowest the spreading speed.

Let R^{max} denote the distance from the source to the furthest deposited particle:

$$R^{max} = \max\{R_{jn} : j \in \mathcal{J},\ n \in \mathcal{N}_j\},$$

where $R_{jn} = ||X_{jn}||$ is the distance between the source (at the origin) and the n-th deposited particle of group j, $\mathcal{J} = \{1, \ldots, J\}$ if $J > 0$ and the empty set otherwise, and $\mathcal{N}_j = \{1, \ldots, N_j\}$ if $N_j > 0$ and the empty set otherwise. By convention, if no particle is dispersed ($J = 0$ or $N_j = 0$ for all j), then $R^{max} = 0$.

Under the GDM, the distribution of the distance between the origin and the furthest deposited particle is zero-inflated and satisfies:

$$\begin{aligned} P(R^{max} = 0) &= \exp\left[\lambda\{p_{\mu,\sigma^2}(0) - 1\}\right] \\ f_{R^{max}}(r) &= \lambda f_{R_j^{max}}(r) \exp\{\lambda(F_{R_j^{max}}(r) - 1)\}, \qquad \forall r > 0, \end{aligned}$$

where $f_{R_j^{max}}$ is the p.d.f. of the distance $R_j^{max} = \max\{R_{jn} : n \in \mathcal{N}_j\}$ between the origin and the furthest deposited particle of group j, and $F_{R_j^{max}}$ is the corresponding cumulative distribution function ($F_{R_j^{max}}(r) = P(R_j^{max} = 0) + \int_0^r f_{R_j^{max}}(u)du$). The distribution of R_j^{max} is zero-inflated and satisfies:

$$\begin{aligned} P(R_j^{max} = 0) &= p_{\mu,\sigma^2}(0) \\ f_{R_j^{max}}(r) &= \int_{\mathbb{R}^2} f_{R_j^{max}|X_j}(r \mid x) f_{X_j}(x)dx \\ &= \sum_{q=1}^{+\infty} q p_{\mu,\sigma^2}(q) \int_{\mathbb{R}^2} f_{R_{jn}|X_j}(r \mid x) F_{R_{jn}|X_j}(r \mid x)^{q-1} f_{X_j}(x)dx, \qquad \forall r > 0. \end{aligned}$$

where $f_{R_{jn}|X_j}$ is the conditional distribution of R_{jn} given X_j satisfying:

$$\begin{aligned} f_{R_{jn}|X_j}(r \mid x) &= 2r \int_0^{r^2} h_1(u, x) h_2(r^2 - u, x)du, \\ h_i(u, x) &= \frac{f_i(\sqrt{u}, x) + f_i(-\sqrt{u}, x)}{2\sqrt{u}}, \quad \forall i \in \{1, 2\}, \\ f_i(v, x) &= \frac{1}{\sqrt{2\pi\nu||x||}} \exp\left(-\frac{(v - x^{(i)})^2}{2\nu||x||}\right), \quad \forall i \in \{1, 2\}, \end{aligned}$$

with $x = (x^{(1)}, x^{(2)})$ and $F_{R_{jn}|X_j}(r \mid x) = \int_0^r f_{R_{jn}|X_j}(s \mid x)ds$.

The material provided above allows the analytic study of the probability that R^{max} is larger than a distance $r > 0$:

$$P(R^{max} \geq r) = \int_r^{+\infty} f_{R^{max}}(s)ds.$$

It especially follows that for every GDM and IDM characterized by the same dispersal kernel for the particles and the same expected number of dispersed particles $\lambda\mu$, the furthest particle under the GDM has less chance to be at a distance greater than any $r > 0$ than the furthest particle under the IDM. Therefore, the population of particles is expected to be more concentrated under the GDM than under the IDM. In other words, the average expansion speed under the GDM is expected to be lower than the average expansion speed under the IDM.

2.3.2 Estimation

In Mrkvička and Soubeyrand (2015), we proposed an MCMC algorithm for Neyman-Scott point processes with inhomogeneous intensity of parent points and inhomogeneous spread of daughter points around their parents. In these processes, both inhomogeneity are described by parametric functions. The group dispersal model presented above is a special case of such a process. Our MCMC algorithm is an adaptation of the algorithm proposed by Mrkvička et al. (2014) for Neyman-Scott point processes with a single inhomogeneity in the intensity of parent points. Briefly, the algorithm consists of updating the process of parent points by using the Birth-Death-Move algorithm described in Møller and Waagepetersen (2003), and updating the model parameters by using a Metropolis-Hastings sampler.

2.3.3 Application

Most often, the spread of airborne plant diseases has been studied at the field, landscape and regional scales (Papaïx et al., 2014; Soubeyrand et al., 2008b, 2009c). It has been more marginally studied at the host scale as in Lannou et al. (2008), who quantified autoinfection (autoinfection is defined as the reinfection of an infected plant by contaminants released by this plant). Autoinfection largely determines the rate of host colonization by the pathogen during polycyclic epidemics and, consequently, the development of epidemics in time and space at larger scales.

In Lannou et al. (2008), we investigated autoinfection for the brown rust of wheat (*Puccinia triticina*) that forms lesions on wheat leaves, from which spores are dispersed, usually by wind (spores are the contaminating particles). Then, spores that are deposited on wheat leaves (the source leaf itself or other leaves) can generate new lesions if conditions are favorable to the development of the disease. Lannou et al. (2008) studied autoinfection for the brown rust of wheat by observing infected leaves with one mother lesions and the resulting set of daughter lesions. Figure 2.7 shows one of these leaves and the locations of all the daughter lesions carried by this leaf. It was checked that this leaf

was sufficiently far from other infected leaves to reasonably assume that the observed daughter lesions only resulted from the mother lesion on the same leaf.

Studying the dispersal at the leaf scale offers the opportunity to directly observe a point pattern, and not only aggregated data such as those discussed in Section 2.1.3. Thus, we fitted the group dispersal model[9] described in Section 2.3.1 to the point pattern shown in Figure 2.7. In the model, the distribution p_{μ,σ^2} of the number of daughter points per parent point was simply the Poisson distribution with mean μ (and $\sigma^2 = \mu$). the dispersal kernel f_{X_j} for parent points was simply the 2D-exponential dispersal kernel given by Equation (2.3). In addition, the prior distributions for the parameters were set to wide uniform distributions.

The point pattern shown in Figure 2.7 is formed by 229 points. The *Mathematica* code that was developed being designed to handle point patterns observed in a rectangular window W, we only considered a subset of the observed point pattern by considering the observation window $W = [-1.80, 1.80] \times [-0.22, 0.17]$ drawn in Figure 2.7. W was chosen such that it was roughly included in the leaf surface and it contained almost all observed lesions (224 locations of lesions are within W). The MCMC algorithm was applied with 10^6 iterations, the first 5×10^4 iterations were discarded as burn-in, every 10th iteration were used for calculation of posterior characteristics.

Table 2.2 provides posterior characteristics of parameters. Estimates that are provided must be interpreted with caution. Indeed, in this case study, dispersal of spores is observed at a small scale and, consequently, our conclusions only concern short-distance dispersal. Thus, parameter estimates only inform the processes that governs the autoinfection of the leaf. They are not representative of processes leading to medium- and long-distance dispersal.

Given these restrictions in the interpretation of parameter estimates, the estimation of β, whose posterior mean is 0.44 and whose 95%-posterior interval is [0.29,0.69], implies that the mean dispersal distance of group centers is about 0.22, i.e. about the half of the maximum width of the leaf. Thus, a non-negligible amount of spores dispersed at the scale of the leaf is lost and/or is deposited on other leaves, especially the spores dispersed in the x_2 direction. This is corroborated by the fact that the product $\lambda\mu$ indicates that about 900 spores that could potentially generate lesions were dispersed at the leaf scale whereas only 229 led to observed lesions.

The estimation of μ indicates that, at the leaf scale, groups of lesions with posterior mean size equal to 7.0 are formed. This corroborates the impression

[9] In Mrkvička and Soubeyrand (2015), we also fitted an independent dispersal model to data and we tested the goodness-of-fit using a rank envelope test (Myllymäki et al., 2015; Mrkvička et al., 2015). The null hypothesis was rejected at the risk level 0.05. We applied the same procedure to assess the fit of the group dispersal model to data. In this case, the null hypothesis was not rejected.

given by Figure 2.7 where we can visually detect clusters of points, especially on the right hand side of the leaf.

The estimation of ν indicates that unidimensional standard deviation $\sqrt{\nu||x||}$ of the distances between the points of a group and their group center is about 0.07 at a distance $||x|| = 0.5$ from the mother lesion, 0.10 at a distance $||x|| = 1.0$, and 0.12 at a distance $||c|| = 1.5$ ($||x||$ is the distance between the mother lesion and the group center —or parent point). This approximately coincides with the clusters of points that are guessed on the right hand side of the leaf on Figure 2.7.

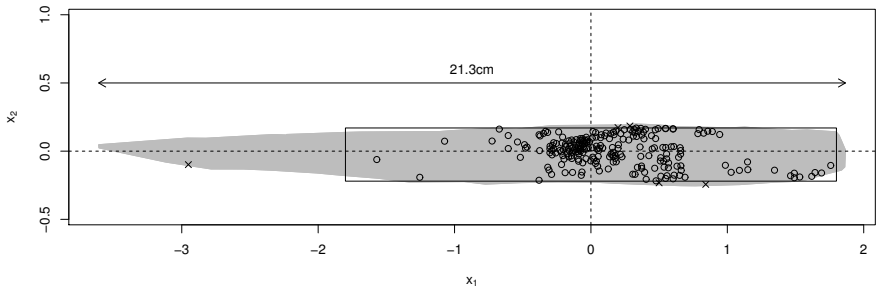


Fig. 2.7. Locations (open circles and crosses) of brown rust lesions on a wheat leaf (grey shape). The pathogen lesions were formed after the dispersal of spores emitted from one mother lesion located at the intersection of the two dashed lines. The estimation algorithm was applied to the point pattern within the rectangle (open circles; 224 points), whereas the lesions outside the rectangle (crosses; 5 points) were not used.

Table 2.2. Estimates of parameters of the group dispersal model fitted to the locations of pathogen lesions on the plant leaf shown in Figure 2.7.

Parameter	β	λ	μ	ν
Posterior mean	0.44	132	7.0	0.010
Posterior median	0.43	127	6.9	0.010
95%-posterior interval	[0.29,0.69]	[71,237]	[3.5,10.6]	[0.004,0.015]

2.3.4 Side topic 1: doubly non-stationary cylinder-based model

Soubeyrand et al. (2014b) propose an alternative and concise way of modeling group dispersal which allowed theoretical investigation of group dispersal in fragmented habitat.

In Section 2.3.1, the particles of a group form a cluster of points, the number of particles in the group is randomly and independently distributed, and the scatter of the particles of the group increases in expectation with the average dispersal distance of the group particles. Here, to get a more concise representation of a group, we propose to model it as a cylinder whose volume is proportional to the size of the group and whose base area increases with the dispersal distance of the group center, which is the cylinder center. This construction leads to a dispersal model that is a doubly nonstationary cylinder-based model[10]. Figure 2.8 (left) shows a realization of this model.

For populations studied in ecology and epidemiology, the habitat is often fragmented (Hanski and Gaggiotti, 2004). Here, we consider and compare two different habitats: (i) a uniform habitat where all points in space are equally favorable to the settlement of the population; (ii) a fragmented habitat modeled by a Boolean model. In case (ii), the Boolean model has to satisfy one particular property: the space covered by the Boolean model has to include the location of the source of particles (e.g. a plant which acts like a source of seeds automatically lies in the plant population habitat). Thus, in case (ii), the habitat is modeled by a conditional Boolean model. Figure 2.8 (center) shows a realization of this model.

The group dispersal model is simply the cylinder-based model for the uniform habitat (Figure 2.8, left) but it is the product of the cylinder-based model and the Boolean model for the fragmented habitat (Figure 2.8, right).

In Soubeyrand et al. (2014b), we describe the models introduced above and we derive their properties. These properties concern the first and second order moments of the random surfaces generated by the models, the probability of population vanishing and the spatial extent of the dispersal. The two latter characteristics, which are studied here because of the biological context underlying our models, are usually not analyzed in stochastic geometry and led to our original theoretical developments, especially in the case of the fragmented habitat. The formula which are provided should allow the study of the interaction between group dispersal and habitat fragmentation. Thus, in future work, we expect to compare the vanishing probabilities and the spatial extents of the product model when the fragmentation of the Boolean model representing the habitat varies.

[10] Such a cylinder-based model can also be viewed as a marked point process where the point are the cylinder centers and the marks are the volume and the base area of the cylinder.

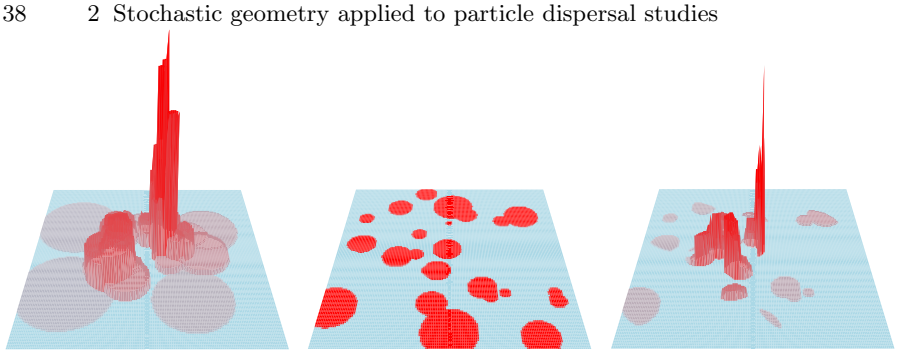


Fig. 2.8. Realizations of the cylinder-based model with source at the origin (left), realization of the conditional Boolean model (center) and product of the two previous realizations (right), which illustrates the group dispersal model in a fragmented habitat.

2.3.5 Side topic 2: group dispersal viewed from an evolutionary perspective

The question *Why to disperse?* has been extensively investigated from an evolutionary perspective, and these investigations generally established the significant advantage of the ability to disperse (Ronce, 2007; Rousset, 2012; Starrfelt and Kokko, 2012; Travis and Dytham, 2002). Beyond this first question, the amazing diversity of dispersal mechanisms that animals, plants, fungi, peat mosses and other organisms have developed leads to a second question: *How to disperse?* In Soubeyrand et al. (2015), we enrich the modeling framework presented in Section 2.3.1 to study the evolution of dispersal in asexual populations where reproducing individuals release particles and can adopt (by mutation) three strategies: independent movements of all particles, clump dispersal (i.e. clumps of particles attached together and settling at the same location), or group dispersal (i.e. groups of particles simultaneously released and settling at different but positively correlated locations). Using simulation experiments, we show (i) how the spatial limits and fragmentation of the species habitat shape the frequencies of the three strategies in the population and the sizes of groups and clumps, and (ii) the co-existence of the independent, clump and group dispersal strategies at the stationary state of the population dynamics.

2.4 Dispersal of phoma at the landscape scale

Current modeling of inoculum transmission from a cropping season to the following one often relies on the extrapolation of kernels estimated on data at short distances from punctual sources, because data collected at larger distances are scarce. In Bousset et al. (2015), we estimated the dispersal ker-

nel of *Leptosphaeria maculans*[11] ascospores from stubble left after harvest in the summer previous to newly sown oilseed rape fields. The estimation was based on data corresponding to counts of lesions observed in autumn during two successive seasons. In the model built to analyze these data, (i) source strengths are described by a log-Gaussian spatial process limited to source fields, (ii) infection potential in the following season is described by a convolution of source strengths and a power-exponential dispersal kernel, and (iii) data are assumed to follow counting distributions conditional on the log-Gaussian spatial process (for data collected in source fields) and on the convolution (for data collected in the target fields). Two data sets were collected from real farmer fields in 2009–2010 and in 2011–2012, respectively. We applied the Bayesian approach for model selection and parameter estimation. We obtained fat-tail kernels for both data sets. This estimation is the first from data acquired over distances of 0 to 1000 m, using several non-punctual inoculum sources. It opens the prospect of refining the existing simulators and developing disease risk maps.

2.4.1 Data

We observed two transmissions of phoma stem canker, from 2009 to 2010 and from 2011 to 2012, in two locations near Le Rheu (Brittany, France). For each transmission, we assessed disease intensity both on source fields and on target fields. A subset of the existing fields was observed, and observation points in observed fields were approximately regularly scattered to cover the whole fields (the sampling was higher near borders of some target fields in contact with source fields). Disease intensity was quantified by counting the number of phoma leaf spots seen within 1 minute on a $1m^2$ rectangle with dimension 0.5×2m and covered by oilseed rape. Data are represented across space in Figure 2.9.

2.4.2 Model

To infer the dispersal of the spores of *Leptosphaeria maculans*, we constructed a mechanistic-statistical model involving a mechanistic model of the dispersal and a probabilistic model of the observation processes. The following paragraphs present each of these two components.

Mechanistic model

Here, we introduce the two following model components: the strengths of pathogen sources in source fields and the infection potential that pathogen sources generate in target fields.

[11] *Leptosphaeria maculans* is a fungus infecting oilseed rape plants and causing the disease called phoma stem canker.

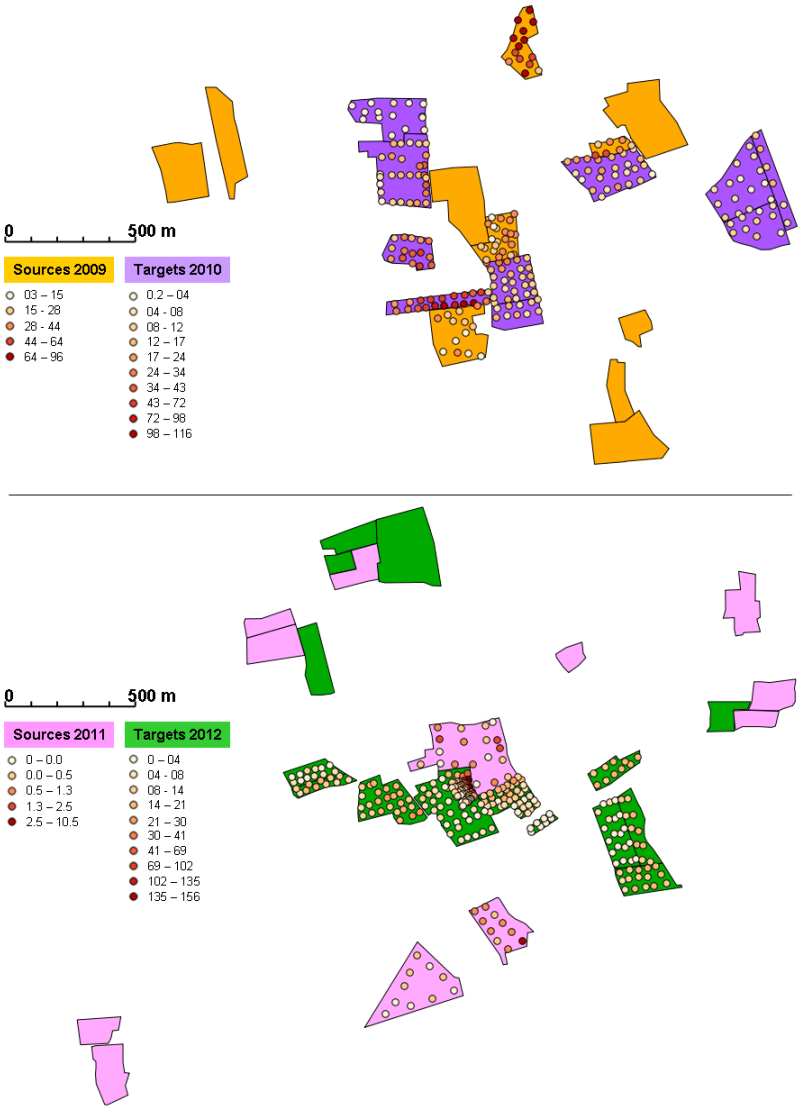


Fig. 2.9. Observed measurements of disease intensities (count of lesions per square meter) in source plots and target plots for the 2009–2010 transmission (top) and the 2011–2012 transmission (bottom). Circles represent sampling points, which are colored with respect to disease intensity. Disease has not been quantified in the fields containing no circles.

The source strengths in source fields are modeled by a log-Gaussian stationary spatial process Λ_S with an exponential power auto-covariance function (Yaglom, 1987, pp. 364-365); the subscript S means source. Thus, for a set of n sites $y_1, \ldots, y_n$ located in source fields,

$$\{\log \Lambda_S(y_1), \ldots, \log \Lambda_S(y_n)\} \sim \text{Normal}\{(\mu, \ldots, \mu), \Sigma\}$$

where $\mu \in \mathbb{R}$ is the mean parameter of the log-source strengths, Σ is their variance-covariance matrix whose element (j, j') is equal to $\sigma_1^2 \exp(-\sigma_2 ||y_j - y_{j'}||^{\sigma_3})$, σ_1^2, σ_2 and σ_3 are the variance parameter, the range parameter and the smoothness parameter of the covariance function, respectively, and $||\cdot||$ is the Euclidean distance in the space $\mathbb{R}^2$.

The infection potential, that gives a measure of the risk of infection in target fields, is modeled by a convolution Λ_T between the source strengths Λ_S and the exponential-power dispersal kernel f; the subscript T means target. Thus, for any site x located in a target field,

$$\Lambda_T(x) = \int_{\mathbb{R}^2} \Lambda_S(y) f(x-y) dy$$
$$f(x-y) = \frac{\beta_2}{\beta_1^2 \Gamma(2/\beta_2)} \exp\left\{ - \left(\frac{||x-y||}{\beta_1} \right)^{\beta_2} \right\},$$

where $\Lambda_S(y) f(x-y)$ represents the contribution of the pathogen source located at site y to the potential infection at site x, and where β_1 and β_2 are the scale and shape parameters of the dispersal kernel.

Models of the observation processes

Disease intensities in source fields noted $Y_{S,i}$, $i = 1, \ldots, n_S$ (i.e. counts of phoma leaf spots seen within 1min on 1m^2), are assumed to be drawn under independent Poisson distributions with means proportional to the source strengths at sites $x_{S,1}, \ldots, x_{S,n_S}$ where the intensities were measured:

$$Y_{S,i} \underset{\text{indep.}}{\sim} \text{Poisson}\{\alpha_1 \Lambda_S(x_{S,i})\}$$

where $\alpha_1 > 0$ is a proportionality parameter. This model is obtained by assuming that phoma leaf spots form a log-normal spatial Cox process and by collecting counting data in small sampling windows instead of observing the points.

Disease intensities in target fields noted $Y_{T,i}$, $i = 1, \ldots, n_T$ (which are also counts of phoma leaf spots seen within 1min on 1m^2), are assumed to be drawn under independent negative-binomial distributions with means proportional to the infection potentials at sites $x_{T,1}, \ldots, x_{T,n_T}$ where the intensities were measured:

$$Y_{T,i} \underset{\text{indep.}}{\sim} \text{Negative-binomial}\{\alpha_2 \Lambda_T(x_{T,i}), \theta\}$$

where $\alpha_2 > 0$ is a proportionality parameter and $\theta > 0$ is an over-dispersion parameter. Here, the negative-binomial distribution is used to counterbalance the regularization due to the convolution defining Λ_T. Indeed, for disease intensities in source fields, the combination of the log-Gaussian randomness in Λ_S and the Poisson randomness leads to a given level of stochasticity. In contrast, Λ_T, which is defined as a convolution between Λ_S and K, has a lower level of variability than Λ_S. To counterbalance this, we used a negative binomial distribution whose realizations are more variable than those of a Poisson distribution with the same mean[12].

2.4.3 Estimation

The model parameters and latent variables were estimated via a MCMC algorithm with Metropolis-Hastings sampler. The vector of unknown parameters is $(\mu, \sigma_1, \sigma_2, \sigma_3, \beta_1, \beta_2, \alpha_1, \alpha_2, \theta)$ (it has to be noted that we arbitrarily fixed $\alpha_1 = 1$ because α_1 and α_2 are not both identifiable.). The latent variables correspond to values of Λ_S at the nodes of a regular square grid limited to source fields. This discretization was performed to handle a finite vector of latent variables in the MCMC and to easily approximate the convolution Λ_T. The MCMC algorithm that we developed was roughly similar to those developed by Diggle et al. (1998) and Bourgeois et al. (2012) for hierarchical spatial models with latent Gaussian vectors. For each MCMC algorithm that we ran, 10^6 iterations were performed, the first 50000 iterations were discarded for the burn-in, and the chain was sub-sampled every 500 iterations to get a posterior sample.

2.4.4 Results

Posterior estimates of the dispersal kernel for the two data sets are shown in Figure 2.10. Both kernel estimates have similar values at distances up to 100m. At larger distances, values issued from the 2010 kernel are higher than values from the 2012 kernel. Figure 2.10 also shows examples of kernels used in Lô-Pelzer et al. (2010) for different wind speeds, selected in the range of observed hourly wind velocities at Le Rheu in autumn 2010 and 2012. The latter kernels tend to underestimate dispersal at short distances for increasing wind speeds. On the contrary, for wind speed from 1 to 20m/s, kernels fall between the two estimated in our study.

[12] This probabilistic trick is a parsimonious manner to obtain the same level of stochasticity in source and target fields. Alternative approaches related to biological processes are discussed in Bousset et al. (2015).

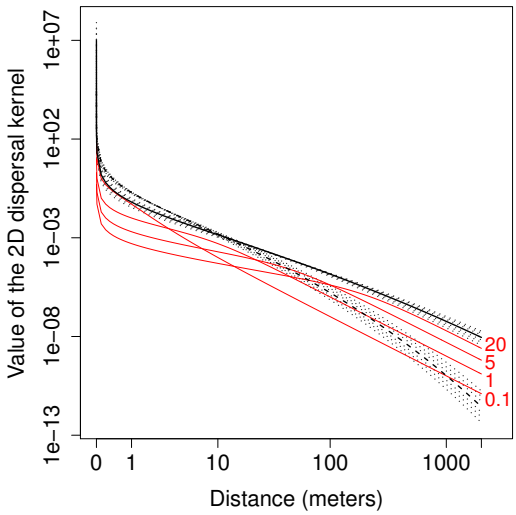


Fig. 2.10. Posterior median of the value of the two dimensional dispersal kernel at increasing distances for the 2009–2010 dataset (solid black line) and the 2011–2012 dataset (dashed black line). Dotted lines are the corresponding posterior quantiles of levels 0.025, 0.1, 0.25, 0.75, 0.9 and 0.975, from bottom to top lines. Red lines are examples of kernels used in the simulator of Lô-Pelzer et al. (2010) with wind speed equal to 0.1, 1, 5 and 20 m/s (indicated on the right of red lines).

2.5 Spatio-temporal dynamics of powdery mildew at the metapopulation scale

In Finland, A.-L. Laine leads a large-scale survey on the yearly distribution of the powdery mildew *Podosphaera plantaginis* which infects its host plant, *Plantago lanceolata*, in a metapopulation setting. The presence of this fungal pathogen has been recorded annually in approximately 4000 host populations, which are most often meadows, across the Åland archipelago, an area of 50×70km in southwest Finland; see Figure 2.11.

In Soubeyrand et al. (2009c), we combined mechanistic and statistical approaches to reconstruct the continuous-time infection dynamics of the powdery mildew based on discrete-time occurrence data. The model takes into account the main features of the dynamics of the pathogen in the Åland archipelago. Specifically, the model takes into account the strong seasonality of the dynamics (i.e. high extinction rate of local pathogen populations during winter, and epidemic expansion in summer). A Bayesian inference framework based on an MCMC algorithm was proposed to infer latent variables (i.e. infection times of meadows) and model parameters (e.g. the dispersal parameters and the effects of covariates on pathogen survival and meadow infection).

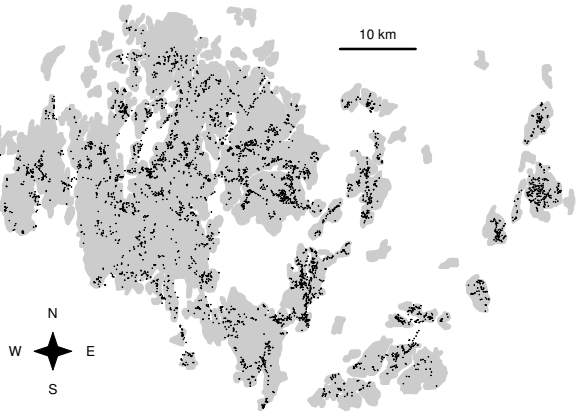


Fig. 2.11. Map of the Åland Islands and the populations of *Plantago lanceolata* (black dots).

2.5.1 Data

The powdery mildew *Podosphaera plantaginis* is an obligate pathogen of *Plantago lanceolata* in the Åland Islands. The host is a perennial herb. Within host populations, initial pathogen foci are established from resting spores or mycelium that have over-wintered in the dormant buds of the host plant. Alternatively, a spore may colonize the host population from surrounding populations. Some six to eight clonally produced generations follow one another in quick succession, often leading to a local epidemic with a substantial fraction of the host individuals being infected by late summer (Ovaskainen and Laine, 2006). Sexually produced resting spores appear towards the end of the growing season in August–September. During the dormancy from September to May, the pathogen population declines greatly as most host plants die back to rootstock.

Podosphaera plantaginis persists regionally as a metapopulation with frequent local extinctions and re-colonizations (Laine and Hanski, 2006). Infection prevalence has remained low, below 7%, yet there is evidence of on-going coevolution, with pathogen populations evolving to infect local co-occurring host plants more efficiently, in comparison with plants occurring elsewhere in the set of meadows (Laine, 2005, 2008).

There are altogether 4108 local populations of *P. lanceolata* in our data base, of which ca 3400 have been surveyed for the presence of the fungus in every year in 2001-06 (the populations that have not been surveyed are mostly very small or occur in peripheral areas). The survey takes place in early September, when the expansion of the local epidemic has terminated (Laine and Hanski, 2006). When the fungus is detected in the survey, a leaf

sample is collected for subsequent microscopic examination to confirm the identification. However, not all infected populations are likely to be detected. Based on thorough control surveys of a subset of the meadows that have been carried out every summer, we estimate that the probability of the field assistants missing an infection on an infected meadow is 6% (i.e. the diagnostic test sensitivity is 94%). This probability of misclassification is included in the model described in the next section.

Environmental variables characterizing each host population (habitat patch) were collected and were introduced in the model as explanatory variables.

2.5.2 Model

To infer the key processes of the host-pathogen interaction during one 'dormancy-growth season' cycle from patterns of occurrence, we constructed a mechanistic-statistical model involving a model of the dynamics and a model of the observation process. Input data to the model includes the covariates characterizing host populations and evoked in the previous section, and the spatial occupancy patterns of the pathogen observed in the beginning of dormancy (initial state) and at the end of the growing season (final state).

In the mechanistic part of the model, survival and extinction are modeled conditionally on the covariates and the health status of host populations in the beginning of dormancy. Colonizations are modeled conditionally on the covariates, on dispersal characteristics of the pathogen, and on the status of host populations in the beginning of the growing season. Colonizations are determined with an infection potential function, which varies in space and time according to infectiousness and spatial pattern of infected host populations.

The statistical part of the model makes the link between the annual dynamics and the occupancy patterns as determined in the beginning of dormancy and at the end of the growing season. This model component enabled us to handle the missed infections, the incomplete sampling and the transformation of pathogen abundance into a binary measure.

Notation

Consider a host-pathogen system comprising a metapopulation. Host populations occur in n distinct habitat patches with centroids x_i ($i \in \mathcal{I} = \{1, \ldots, n\}$). a_i denotes the area covered by host individuals in patch i.

We first focus on spatio-temporal dynamics of the pathogen during one year, which is assumed to consist of two successive periods: the *dormancy* period and the *growing season* period. Without loss of generality, we assume that dormancy occurs during the time interval $[-1, 0)$ while the growing season occurs during the interval $[0, 1)$. The initial time $t = -1$ is just after the end of the previous growing season, while time $t = 1$ corresponds to the beginning of the next season.

The binary variable Y_{it} denotes the *health status* of population i at time t: $Y_{it} = 0$ if i is susceptible and $Y_{it} = 1$ if it is infected. Healthy populations are immune during the dormancy and susceptible within the growing season, while infected populations are infectious only during the growing season. In addition, like in Section 2.1.2 the degrees of susceptibility and infectiousness depend on individual characteristics and time; see below.

The presence of the pathogen is assessed at the population level at times $t = -1$ and $t = 1$ as $Y_{i,-1}$ and Y_{i1}. Given that sampling is not complete (there are some populations whose health status is not observed) and that infections are not always detected, we introduce the observation variables $Y_{i,-1}^{obs}$ and Y_{i1}^{obs} with a value of zero if the meadow is observed as healthy, one if it is observed infected and `NA` if it is not sampled. A population that is observed to be healthy can be actually infected.

The *infection times* T_i ($i \in \mathcal{I}$) denote the times of initiation of local epidemics in the year under consideration. As a local epidemic can only occur during the growing season, $T_i \geq 0$. We assume that the pathogen survived in population i during the dormancy if and only if $T_i = 0$. In the case of local epidemics not due to survival of the pathogen in patch i the infection time is the colonization time. By convention, we set $T_i \geq 1$ if population i is still susceptible at time $t = 1$. Figure 2.12 gives an example of temporal evolution in the number of infected populations during the dormancy and the growing season.

The set of infection times is denoted by $\mathbf{T} = \{T_i : i \in \mathcal{I}\}$ and the sets of observed initial and final health statuses are denoted by $\mathbf{Y}_{-1}^{obs} = \{Y_{i,-1}^{obs} : i \in \mathcal{I}\}$ and $\mathbf{Y}_{1}^{obs} = \{Y_{i1}^{obs} : i \in \mathcal{I}\}$, respectively.

Mechanistic model

In the model the pathogen survives in patch i with probability $b_i s(Y_{i,-1}^{obs})$, which depends on individual characteristics encoded in $b_i \in [0, 1]$ and on $Y_{i,-1}^{obs} \in \{0, 1, \mathtt{NA}\}$. b_i gives the conditional probability of survival given that patch i was infected in the beginning of dormancy. The probability b_i was specified as a function of the observed covariates that are expected to be linked to pathogen survival: $b_i = \text{logit}^{-1}(B_i^T \beta)$ where B_i is a vector of covariates and β is a vector of parameters. Function s deals with misclassification and incompleteness of the observation process at time $t = -1$. It depends on observation–parameters which are supposed to be known (or assessed with additional data)[13].

The spread of the pathogen during the growing season was modeled as a spatio-temporal Poisson point process (Illian et al., 2008), as proposed in

[13] The function s is specified in an appendix of Soubeyrand et al. (2009c). It depends on the following observation–parameters: the probability that a population is observed as infected at time $t = -1$ if it is sampled; and the probability that a population is infected at time $t = -1$ given it is observed as non-infected.

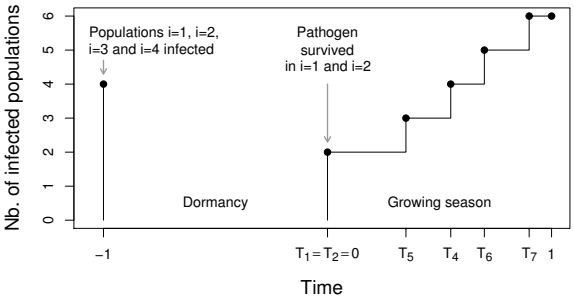


Fig. 2.12. Temporal evolution of the number of infected populations during the dormancy and the growing season corresponding to intervals [-1,0) and [0,1), respectively. Labels $i = 1, \ldots, 7$ are used to denote the populations and T_i is the infection time of population i. In this example, four populations ($i = 1, 2, 3, 4$) were infected in the beginning of dormancy but the pathogen survived only in populations 1 and 2. During the growing season, populations 5, 6 and 7 became colonized, population 4 was recolonized and population 3 remained susceptible.

Section 2.1.2 where the *Poisson specification* was introduced[14]. In this process, point (t, x) specifies a time and a location at which the numbers of dispersing incoming pathogen are large enough to potentially initiate a local epidemic in a susceptible population with a standard degree of susceptibility. Thus, each point stands for a potential colonization event.

The point process is governed by an intensity function λ quantifying the risk of infection at each space-time location. This intensity, called here *infection potential*, is generated by the already infected populations and it therefore varies in time and space with the number, the spatial locations and the infectiousness of these populations. The expression of λ at time t and location x is given by:

$$\lambda(t, x) = \sum_{j \in \mathcal{I}_t} c_j g_j(t - T_j) f(x, x_j), \tag{2.14}$$

where $\mathcal{I}_t = \{j \in \mathcal{I} : T_j < t\}$ is the set of populations infected before time t; c_j encodes characteristics of population j such as its physiological state and features of the surrounding habitat, which are expected to partly determine the infectiousness of j; g_j is a parametric disease progress function[15], which gives

[14] More rigourously, the point process incorporated in the model is a *piecewise* spatio-temporal Poisson point process where the pieces are defined as the periods of time between the colonization events.

[15] The function g_j was specified as follows: $g_j(t) = \min\{t^2, \omega a_j\}\mathbf{1}(t \geq 0)$, where ω is a positive parameter. The threshold ωa_j takes into account possible saturation effects in small populations.

the shape of the pathogen growth within population j; f is a dispersal function, which models pathogen dispersal as a function of the source location x_j and the location of the receiving population x. For f, we used the anisotropic exponential kernel (see Equation (2.7)) specified with von Mises anisotropy functions. The product $c_j g_j(t - T_j)$ specifies the degree of infectiousness of population j at time t. In the beginning of the growing season, just after time zero, the infection potential is generated only by those populations in which the pathogen survived during the dormancy.

An uninfected population i is colonized during the growing season if a point of the Poisson point process is deposited in i and it succeeds in initiating a local epidemic. The intensity of points deposited in i at time t is given by the product $a_i \lambda(t, x_i)$ of what is considered as the effective capture area by the instantaneous local infection potential. Any deposited point is assumed to initiate a local epidemic with probability d_i, which reflects the degree of susceptibility of i and encodes individual characteristics such as local climatic conditions.

Quantities c_j and d_i always appear in the model as the product $c_j d_i$. They can be jointly modeled as a function of observed covariates that are expected to be linked to infectiousness and susceptibility: $c_j d_i = \exp(C_j^T \gamma + D_i^T \delta)$, where C_j and D_i are vectors of covariates, and γ and δ are vectors of parameters.

Models of the observation processes

The observed final health status Y_{i1}^{obs} is assumed to follow a multinomial distribution in the set of values $\{0, 1, \texttt{NA}\}$. The probability that $Y_{i1}^{obs} = \texttt{NA}$ is the same for all populations and, therefore, is reduced to a single parameter. The probabilities that $Y_{i1}^{obs} = 0$ and that $Y_{i1}^{obs} = 1$ are assumed to depend on the actual health status Y_{i1} and on observation–parameters related to misclassification and incompleteness in the observation process[16].

2.5.3 Estimation

In the application, the input data are the observed health statuses $Y_{i,-1}^{obs}$ and Y_{i1}^{obs} (for years 2002 to 2006), the vectors of covariates B_i, C_i and D_i, and the observation–parameters. Based on these data and the model presented above, we inferred the infection times $\mathbf{T}$ and the vector of unknown parameters θ, which includes the parameters associated with the covariates, the parameter

[16] The conditional distribution of Y_{i1}^{obs} given Y_{i1} is specified in an appendix of Soubeyrand et al. (2009c). It depends on the following observation–parameters: the probability that a population is observed as infected at time $t = 1$ if it is sampled; and the probability that a population is infected at time $t = 1$ given it is observed as non-infected. These probabilities are supposed to be known (they were assessed with additional data).

of the growth functions g_j and the dispersal parameters. The inference was carried out in the Bayesian framework with an MCMC algorithm including Metropolis-Hastings updates.

2.5.4 Results

Survival probability

Survival during dormancy (i.e. over-wintering) cannot be observed directly, but our inference approach allows us to estimate this measure of the host-pathogen interaction. Indeed, by using the posterior distribution of the infection times, we can estimate the posterior probability that the pathogen survived during dormancy in any population i and year $k \in \{2002 \ldots, 2006\}$, and we can express its average, say $\hat{S}_i$, over time.

Figure 2.13 (left) shows $\hat{S}_i$ for all populations. It is apparent that there are regions of high survival probability especially near the coastline (where climate conditions are relatively mild). In other coastal regions, the survival probability has an intermediate value, whereas in the inland areas the survival probability is mostly low.

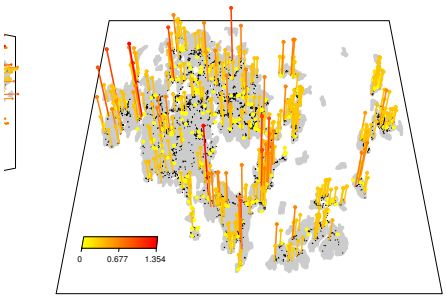


Fig. 2.13. Posterior estimates $\hat{S}_i$ and $\hat{E}_i$ of the survival probability during dormancy (left) and the encounter intensity during the growing season (right), respectively. Redder and taller the bar, higher the estimate. A black dot without any bar indicates a host population in which $\hat{S}_i = 0$ and $\hat{E}_i = 0$.

Encounter intensity

The encounter intensity is a quantitative measure of the host-pathogen interaction during the growing season. We defined it for population i and year k as the integral over time of the disease progress function:

$$E_{ik} = \int_0^1 g_{ik}(t - T_{ik})dt.$$

The motivation for this choice is that $g_{ik}(t - T_{ik})$ measures the spatial extent of the pathogen within population i at time t. This instantaneous encounter intensity is integrated over the growing season $[0, 1]$ to account for temporal variation in encounter. The encounter intensity E_i for population i during the entire study period is the sum of the annual encounter intensities: $E_i = \sum_{k=2002}^{2006} E_{ik}$.

Figure 2.13 (right) shows $\hat{E}_i$ for all populations. This map is rather similar to the one displaying the survival probabilities, but more populations have positive encounter intensities (more bars in the map), the spatial pattern is smoother, and the values are greater in the inland areas than in the case of survival probabilities. These features are obviously due to dispersal of the pathogen during the growing season.

Dispersal is quantified in the model by the dispersal function f. Figure 2.14 shows the posterior median of the dispersal function. The mean dispersal distance has the posterior median of 0.86 km and the 95% posterior interval of [0.64,1.18]. Furthermore, it is apparent that dispersal is oriented towards the East / South-East corresponding to the main wind direction in the Åland archipelago.

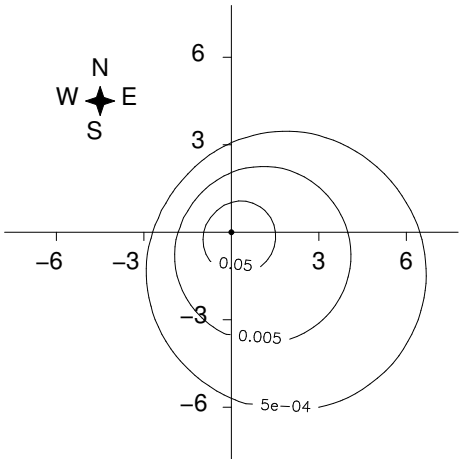


Fig. 2.14. Posterior median of the dispersal function h. Each contour line is a curve of constant density (0.05, 0.005 and 0.0005). The unit of the axes is the kilometer.

Effects of covariates

Table 2.3 shows the effects of covariates used to characterize survival, infectiousness and receptivity. Significant effects are those with stars and high mean squares. The probability of survival is particularly decreased for populations which are far from the shoreline and dry. The infectiousness is particularly

large for populations far from the shoreline, adjacent to roads and shady. The receptivity is increased for populations adjacent to roads and decreased for dry populations. It has to be noted that the estimated receptivity does not fluctuate as much as the estimated survival and infectiousness (mean squares for effects linked with receptivity are much smaller than those linked with survival and infectiousness).

Variable	Survival		Infectiousness		Receptivity	
P. lanceolata coverage at $t=-1$	$+^{***}$	(0.00)				
Distance to shoreline	$-^{***}$	(0.71)	$+^{***}$	(0.47)	$-$	(0.007)
Patch shadiness	$-$	(0.01)	$+^{***}$	(0.21)	$-$	(0.010)
Dry-hosts proportion at $t=-1$	$-^{*}$	(0.39)				
Dry-hosts proportion at $t=1$			$+$	(0.06)	$-^{***}$	(0.065)
Open-habitat proportion	$-^{**}$	(0.17)	$+$	(0.04)	$-^{**}$	(0.026)
Road presence	$-^{**}$	(0.12)	$+^{***}$	(0.30)	$+^{***}$	(0.070)

Table 2.3. Effects of variables used in vectors B_i, C_j and D_i corresponding to survival, infectiousness and receptivity, respectively. For each effect, the sign (+ or −) of the posterior median is given together with the level of the posterior probability that the effect has the sign of the posterior median (one star: less than 0.05, two stars: less than 0.005; three stars: less than 0.0005). The posterior median of the mean square of each effect is given between parentheses; it measures the variability due to each effect.

Regions of temporal stability

Based on an analysis not shown here (but presented in Soubeyrand et al., 2009c), there is a high turnover in the infection of meadows: every year, during the dormancy period, there is a significant shuffling of the pathogen populations. However, there may be regions in which the pathogen has a high chance of surviving from one year to another. Similarly, there may be regions in which the encounter intensity is high every year. We searched for such stability regions which are shown by Figure 2.13. As expected, the regions of stability are located in the coastal regions. Most areas of high stability for survival and encounter coincide, but the overlap is not complete.

Thus, we were able to identify the regions in the study area where over-wintering (which cannot be observed directly) has been most successful. These over-wintering sites represent foci that initiate local epidemics during the growing season. There is striking heterogeneity at the regional scale in both over-wintering success of the pathogen and in the encounter intensity between the host and the pathogen. Such heterogeneity has profound implications for the coevolutionary dynamics of the interaction. Specifically, the regions pointed out by Figure 2.13 could be hotspots for coevolution.

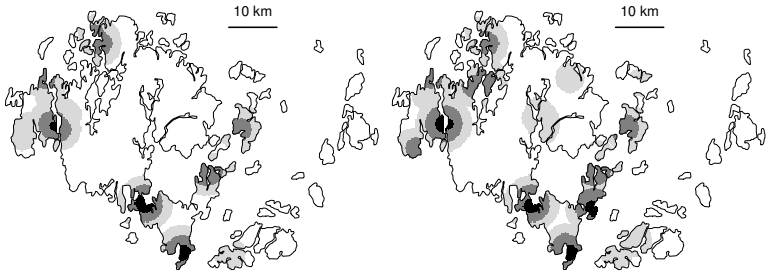


Fig. 2.15. Regions covered by stability disks (SD) of radius 1 km (black), 2 km (black and dark grey) and 3 km (black and grey). Left: stability for survival. Right: stability for encounter. A stability disk is a disk in which survival (or encounter) was high during the five-years study period. Here, survival (or encounter) is neither considered at the scale of the meadow nor at the scale of the metapopulation, but at an intermediate scale.

3

Genetic-space-time modeling and inference for epidemics

Author's references: Morelli et al. (2012), Mollentze et al. (2014), Soubeyrand (2016), Valdazo-González et al. (2015).

Viruses can cause human, animal and plant epidemics of high impact in developing and developed countries alike. For instance, hepatitis E caused 57,000 deaths in 2010 (Lozano et al., 2012). During the foot-and-mouth outbreak in Great Britain in 2001, some 6 million animals were culled (Anderson et al., 1996; Haydon et al., 2004). On the global scale and over three decades, the overall cost of sharka (infecting trees of the genus *Prunus*) was estimated to exceed 10 billion euros (Cambra et al., 2006).

In order to minimize these social, environmental and economic costs, we need to most effectively control infectious diseases and thus to better understand how pathogens spread within host populations, yet this is something we know remarkably little about. Identifying transmission links of an infectious disease through a host population is critical to understanding its epidemiology and informing measures for its control. Infected hosts close together in their locations and timing are often thought to be linked, but timing and locations alone are usually consistent with many different scenarios of *who infected whom*. To infer more reliably who-transmitted-to-whom over the course of disease outbreaks, pathogen genomic data have been combined with spatial and/or temporal data (Jombart et al., 2014; Hall et al., 2015; Lau et al., 2015; Mollentze et al., 2014; Morelli et al., 2012; Ypma et al., 2012, 2013). However, the manner in which these data have to be combined remains a modeling and statistical challenge today.

One of the approaches recently proposed is based on an extension of stochastic Susceptible-Exposed-Infectious-Removed (SEIR) models. It combines heterogeneous and multi-scale processes and data: it links the epidemiological scale —or host population(s) scale— and the micro-evolutionary scale —or pathogen genome scale. Section 3.1 presents the genetic-space-time model that we developed and that combines (i) an individual-based, spatial, semi-Markov SEIR model for the spatio-temporal dynamics of the pathogen, and (ii) a Markovian evolutionary model for the temporal evolution of genetic sequences of the pathogen. The resulting model is a state-space model including latent vectors of high dimension. Section 3.2 describes how model parameters

and latent variables were estimated in the Bayesian framework via approximate MCMC algorithms. Section 3.3 presents simulation studies assessing the performance of the approach and real case studies illustrating the application of the approach to foot-and-mouth outbreaks and a rabies endemic dynamic. A focus is given on the estimation of who infected whom.

3.1 Joint modeling of epidemiological and micro-evolutionary dynamics

The genetic-space-time SEIR model, presented below in Subsection 3.1.6, is a combination of a semi-Markov epidemic model and a Markovian evolutionary model. The following subsections show how these submodels and the resulting genetic-space-time SEIR model are constructed.

3.1.1 Discrete-state, continuous-time Markovian SEIR model

Here, we consider a classical SEIR model describing the temporal dynamics of numbers of susceptible, exposed, infectious and removed individuals in a population affected by a pathogen. Time is viewed as a continuous variable. Let $\mathbf{S}(t)$, $\mathbf{E}(t)$, $\mathbf{I}(t)$ and $\mathbf{R}(t)$ in $\mathbb{N}$ respectively denote the numbers of susceptible, exposed, infectious and removed individuals at time $t \geq 0$. The sum of these quantities is equal to the instantaneous total size of the population $\mathbf{T}(t) \in \mathbb{N}$, i.e. $\mathbf{S}(t) + \mathbf{E}(t) + \mathbf{I}(t) + \mathbf{R}(t) = \mathbf{T}(t)$ for any time $t \geq 0$.

In general, many different events can cause a change in the population pattern $(\mathbf{S}(t), \mathbf{E}(t), \mathbf{I}(t), \mathbf{R}(t))$, for instance the birth and death of susceptibles, the infection of susceptibles, the death of exposed individuals, the end of exposed stage (coinciding with the beginning of the infectious stage), the death in infectious individuals and the end of infectious stage (coinciding with the beginning of removed stage).

Here, we consider only three possible events, namely *infection*, *end of exposed stage* and *end of infectious stage*. Corresponding transition rates are provided in Table 3.1. In this model, the risk of infection is a combination of a basic risk, whose rate is $\alpha_0 \mathbf{S}(t)$, and an endogenous risk, whose rate $\alpha_1 \mathbf{S}(t)\mathbf{I}(t)/\mathbf{T}(t)$ is proportional to the number of infectious individuals in the population of interest. The basic risk may correspond to exogenous pathogen sources. For instance, in the case of zoonoses (i.e. diseases that can be transmitted from animals to humans) if the population of interest is the set of humans, the basic risk may correspond to animals infecting humans.

3.1.2 Spatial extension

Now, consider a space decomposed into $n \in \mathbb{N}^*$ districts. In each district $k \in \{1, \ldots, n\}$, the size $\mathbf{T}_k(t)$ of the resident population at time $t \geq 0$ is the

Table 3.1. Possible events and corresponding transition rates for the discrete-state, continuous-time Markovian SEIR model.

Description	Event	Rate
Infection	$\mathbf{S} \to \mathbf{S} - 1$ & $\mathbf{E} \to \mathbf{E} + 1$	$\alpha_0 \mathbf{S} + \alpha_1 \mathbf{SI}/\mathbf{T}$
End of exposed stage	$\mathbf{E} \to \mathbf{E} - 1$ & $\mathbf{I} \to \mathbf{I} + 1$	$\beta \mathbf{E}$
End of infectious stage	$\mathbf{I} \to \mathbf{I} - 1$ & $\mathbf{R} \to \mathbf{R} + 1$	$\delta \mathbf{I}$

sum of local numbers of susceptible, exposed, infectious and removed individuals denoted by $\mathbf{S}_k(t)$, $\mathbf{E}_k(t)$, $\mathbf{I}_k(t)$ and $\mathbf{R}_k(t)$, respectively. By assuming that contacts between individuals of different districts are possible, the local risk of infection is a combination of a basic risk whose rate is $\alpha_0 \mathbf{S}_k(t)$, a local endogenous risk whose rate is $\alpha_1 \mathbf{S}_k(t)\mathbf{I}_k(t)/\mathbf{T}_k(t)$, and a distant endogenous risk whose rate is $\alpha_1 \mathbf{S}_k(t) \sum_{j \neq k} w_{kj} \mathbf{I}_j(t)/\mathbf{T}_j(t)$. In the latter rate, the weight w_{kj} is a measure of the intensity of contacts between individuals of districts k and j. By convention, the intensity of contacts between individuals residing in the same district is equal to one. The spatial, discrete-state, continuous-time Markovian SEIR model considered in this section is defined by events and rates provided in Table 3.2. It has to be noted that under this model the sizes $\mathbf{T}_k$ of district subpopulations are constant across time.

Table 3.2. Possible events and corresponding transition rates for the spatial, discrete-state, continuous-time Markovian SEIR model.

Description	Event	Rate
Infection	$\mathbf{S}_k \to \mathbf{S}_k - 1$ & $\mathbf{E}_k \to \mathbf{E}_k + 1$	$\alpha_0 \mathbf{S}_k + \alpha_1 \mathbf{S}_k \mathbf{I}_k/\mathbf{T}_k$ $+\alpha_1 \mathbf{S}_k \sum_{j \neq k} w_{kj} \mathbf{I}_j/\mathbf{T}_j$
End of exposed stage	$\mathbf{E}_k \to \mathbf{E}_k - 1$ & $\mathbf{I}_k \to \mathbf{I}_k + 1$	$\beta \mathbf{E}_k$
End of infectious stage	$\mathbf{I}_k \to \mathbf{I}_k - 1$ & $\mathbf{R}_k \to \mathbf{R}_k + 1$	$\delta \mathbf{I}_k$

3.1.3 Particular case: individual-based version of the model

Now, let us consider a particular case of the previous model: assume that for all $k \in \{1, \ldots, n\}$, the size $\mathbf{T}_k(t) \equiv 1$. Thus, districts are replaced by single individuals, and the model becomes an individual-based model where the dynamics of the epidemics is modeled at the individual resolution. In addition, values of $\mathbf{S}_k(t)$, $\mathbf{E}_k(t)$, $\mathbf{I}_k(t)$ and $\mathbf{R}_k(t)$ are in $\{0, 1\}$ and their sum is equal to one whatever t. By assuming that each individual $k \in \{1, \ldots, n\}$ is located at x_k in the planar space $\mathbb{R}^2$, events and rates shown in Table 3.2 can be re-written as in Table 3.3. The location x_k can be viewed as the central or main location of k. We can see in Table 3.3 that the local endogenous risk disappeared from the expression of the rate of infection since a susceptible

individual cannot infects himself (another justification is that $\mathbf{S}_k(t)\mathbf{I}_k(t) = 0$, $\forall t \geq 0$). Moreover, the rate corresponding to the distant endogenous risk is now written $\alpha_1 \sum_{j \neq k} \mathbf{I}_j w(x_j - x_k)$ where w is a dispersal kernel whose value depends on the relative locations of individuals k and j. We specified w as a power-exponential kernel (already encountered in Chapter 2) parametrized by $\alpha_2 = (\alpha_{2,1}, \alpha_{2,2})$ and satisfying, for all $x \in \mathbb{R}^2$:

$$w(x) = \frac{\alpha_{2,2}}{2\pi(\alpha_{2,1})^2 \Gamma\left(\frac{2}{\alpha_{2,2}}\right)} \exp\left\{-\left(\frac{||x||}{\alpha_{2,1}}\right)^{\alpha_{2,2}}\right\}, \tag{3.1}$$

where $||x||$ is the Euclidean distance between the origin of the planar space and x. Thus, the measure of the intensity of contact between individuals k and j decreases with the distance separating the central locations of k and j.

Table 3.3. Possible events and corresponding transition rates for the individual-based, spatial, discrete-state, continuous-time Markovian SEIR model.

Description	Event	Rate
Infection	$\mathbf{S}_k : 1 \to 0$ & $\mathbf{E}_k : 0 \to 1$	$\alpha_0 + \alpha_1 \sum_{j \neq k} \mathbf{I}_j w(x_j - x_k)$
End of exposed stage	$\mathbf{E}_k : 1 \to 0$ & $\mathbf{I}_k : 0 \to 1$	β
End of infectious stage	$\mathbf{I}_k : 1 \to 0$ & $\mathbf{R}_k : 0 \to 1$	δ

3.1.4 Semi-Markov extension of the individual-based model

In the Markovian model presented in the previous subsection, the times spent by an individual in exposed and infectious stages are exponentially distributed. Depending on the context, this may be viewed as an unrealistic assumption. For instance, the exposed duration, that corresponds to a latency or incubation duration, is usually not exponentially distributed but has a distribution with a mode away from zero (e.g. see Hampson et al., 2009). Semi-Markov models (Barbu and Limnios, 2008) offer a framework to handle non-exponential durations in some of the possible states. Thus, in this subsection, we introduce a semi-Markov model where durations in the exposed and infectious stages are independently drawn under gamma distributions (see Table 3.4). The draws are also independent from the duration in the susceptible stage.

3.1.5 Markovian evolutionary model for a pathogen sequence

Now, suppose that the pathogen under consideration is an RNA virus that can evolve along time. More specifically, we suppose that mutations can occur at s

Table 3.4. Possible events and corresponding transition rates or distributions for the individual-based, spatial, discrete-state, continuous-time, semi-Markov SEIR model.

Description	Event	Rate	Distribution
Infection	$\mathbf{S}_k : 1 \to 0$ & $\mathbf{E}_k : 0 \to 1$	$\alpha_0 + \alpha_1 \sum_{j \neq k} \mathbf{I}_j w(x_j - x_k)$	
End of exposed stage	$\mathbf{E}_k : 1 \to 0$ & $\mathbf{I}_k : 0 \to 1$		$\Gamma(\beta_1, \beta_2)$
End of infectious stage	$\mathbf{I}_k : 1 \to 0$ & $\mathbf{R}_k : 0 \to 1$		$\Gamma(\delta_1, \delta_2)$

sites of the viral sequence between the four possible nucleobases that are adenine (A), cytosine (C), guanine (G) and uracil (U). We assume that mutations in different sites are independent but mutation rates vary as functions of the current nucleobases and the substituting nucleobases as in the 3-parameters Kimura substitution model (Kimura, 1981). Here, *mutation* and *substitution* are synonyms. Thus, at the s sites of the sequence under mutation, the mutation processes follow independent, discrete-state, continuous-time Markovian models that are defined by events and rates provided in Table 3.5.

Table 3.5. Possible events and corresponding substitution rates for the Markovian evolutionary model. Letters A, C, G and U denotes nucleobases adenine, cytosine, guanine and thymine, respectively.

Description	Event	Rate
Transition	A→G or G→A or C→U or U→C	μ_1
Transversion of type 1	A→U or U→A or C→G or G→C	μ_2
Transversion of type 2	A→C or C→A or G→U or U→G	μ_3

Under this setting, the expected proportions of transitions, type-1 transversions, type-2 transversions and unchanged nucleobases over an evolutionary time lag Δ separating two sequences are:

$$\begin{aligned}\rho &= (\rho_1, \rho_2, \rho_3, \rho_4) \\ &= \frac{1}{4}\big(1 - e_1 - e_2 + e_3,\ 1 - e_1 + e_2 - e_3,\ 1 + e_1 - e_2 - e_3,\ 1 + e_1 + e_2 + e_3\big),\end{aligned}$$

where $e_1 = \exp\{-2(\mu_1 + \mu_2)\Delta\}$, $e_2 = \exp\{-2(\mu_1 + \mu_3)\Delta\}$, $e_3 = \exp\{-2(\mu_2 + \mu_3)\Delta\}$, and μ_1, μ_2 and μ_3 are the genetic substitution rates per nucleotide per day, for transitions, type-1 transversions and type-2 transversions, respectively.

In addition, we make the following distributional assumption: the numbers of observed transitions, type-1 transversions, type-2 transversions and unchanged nucleobases over an evolutionary time lag Δ separating two sequences are distributed from a multinomial distribution, say $P_{\mu,s}(\cdot \mid \Delta)$, with size equal to the length s of the observed sequence fragment and with

the vector of probabilities ρ given above. Thus, for any nonnegative integers m_1, m_2, m_3, m_4 whose sum is s,

$$P_{\mu,s}\{(m_1, m_2, m_3, m_4) \mid \Delta\} = \frac{(s!) \times \rho_1^{m_1} \rho_2^{m_2} \rho_3^{m_3} \rho_4^{m_4}}{(m_1!) \times (m_2!) \times (m_3!) \times (m_4!)}.$$

3.1.6 Genetic-space-time SEIR model

The genetic-space-time SEIR model, whose structure is illustrated by Figure 3.1, is obtained by combining the semi-Markov SEIR model of Subsection 3.1.4 and the Markovian evolutionary model of Subsection 3.1.5. The two models are combined under the following list of assumptions.

- The disease reservoir (i.e. exogenous source) is assumed to simply consist of one virus sequence S_{exo} dated at time $t_{\text{exo}} \in \mathbb{R}$.
- We assume that, at any time, there is only one sequence of the virus (i.e. one viral variant) per infected individual[1]. The sequence in individual k at time t, for t such that $\mathbf{E}_k(t) = 1$ or $\mathbf{I}_k(t) = 1$, is denoted by $S_k(t)$ and is a vector of letters A, C, G and U.
- When a susceptible individual k is infected by an infectious individual j at time t, the sequence $S_j(t)$ is transmitted to k, i.e. $S_k(t) = S_j(t)$.
- For any individual, mutations of the virus sequence during the exposed and infectious stages are assumed to be independent from the epidemiological dynamics. When an individual is removed, the sequence is fixed. This is illustrated in Figure 3.2 where the length of the sequence under mutation is $s = 2$.
- Virus mutations in different infected individuals are assumed to be conditionally independent given the virus sequences at the infection times.

3.2 Estimation methods

3.2.1 Data structure

Data are assumed to be as follows:

- Data are collected in a spatio-temporal observation window included in the whole spatial domain and in the whole time frame covered by the disease dynamics (which is either an epidemic or an endemic dynamics);

[1] In Valdazo-González et al. (2015), the impact of within-host-unit genetic variation upon inferred transmission trees is assessed for the foot-and-mouth disease. In this case, this impact was moderate. However, for other contexts, using the within-host diversity of virus sequences can be informative. This is the topic of a project that I have recently proposed to the French research agency ANR.

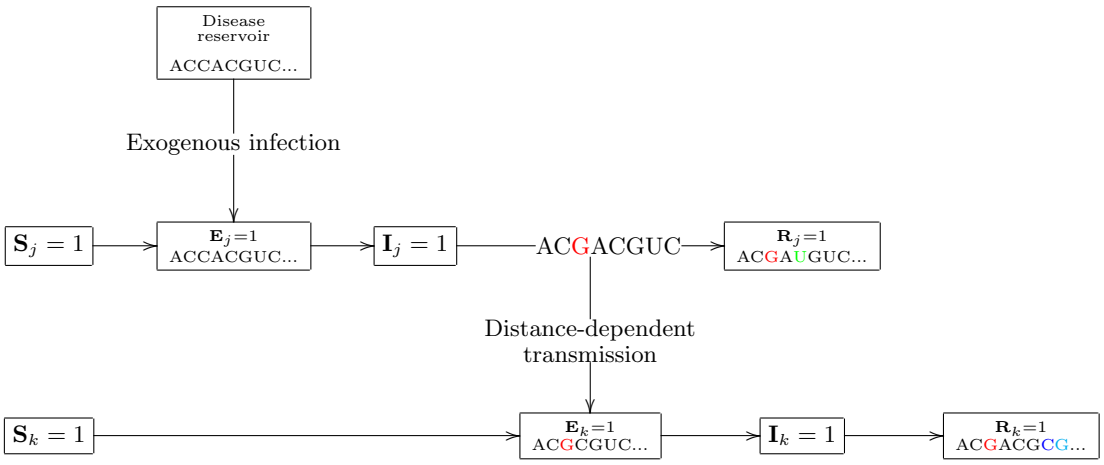


Fig. 3.1. Diagram illustrating the combination of the semi-Markov SEIR model and the Markovian evolutionary model. Individual j is infected by the disease reservoir with virus sequence "ACCACGUC...". Then, j becomes infectious and when j infects k, the sequence in j has evolved (C at the 3rd base mutated to G). Finally, the sequences in j and k independently evolve.

- Among all the infected cases in the spatio-temporal observation window, only a subset is observed (thus, sources of infection can be unobserved, and the unobserved sources can be inside or outside the observation window);
- The removed state corresponds to the death of the individual;
- Sampled individuals are observed when they die, that is to say at the transition from $\mathbf{I} = 1$ to $\mathbf{R} = 1$;
- Virus sequences that are available correspond to the states of the sequences at the death times (usually only a fragment of the sequence is available, the same fragment for all inividuals);
- The central locations $x_1, \ldots, x_n$ of sampled individuals are assumed to be the locations at the death;
- The sequence S_{exo} is assumed to be known (in real cases, S_{exo} can be reconstructed by a phylogenetic analysis of available genetic sequences: S_{exo} is typically the reconstructed sequence of the most recent common ancestor; see Mollentze et al., 2014).

Compared to the amount of variables in the model, data are particularly sparse. Indeed, looking at Figure 3.1, data consist of the sequence of the disease reservoir and locations, times and sequences collected when observed individuals are in state $\mathbf{R} = 1$. It has to be noted that, usually, only a fraction of infected individuals are observed. Now, consider Figure 3.2, infected individuals and pathogen sequences are observed in one of the state $\mathbf{R} = 1$ whereas they previously evolved in a high-dimension space of states (37 states

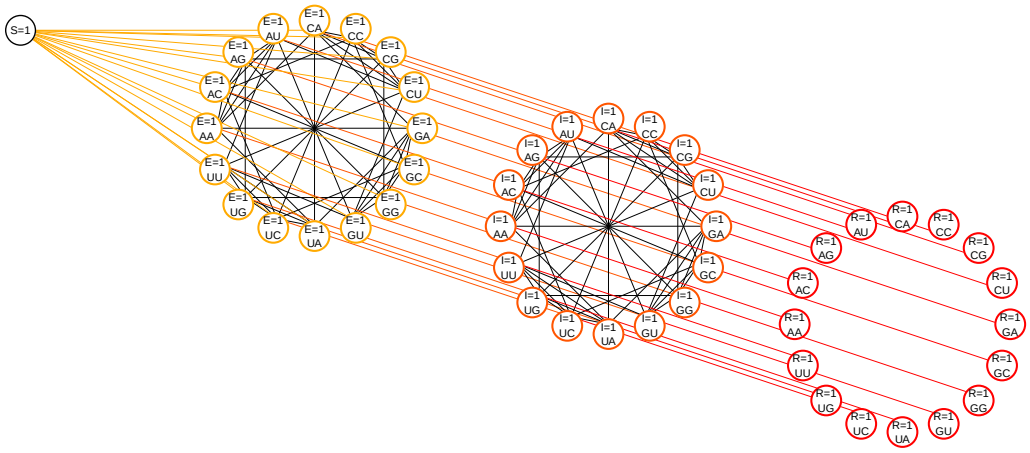


Fig. 3.2. Possible transitions for an individual in the genetic-space-time SEIR model, where the virus sequence is of length $s = 2$. Transitions from $\mathbf{S} = 1$ (susceptible) to $\mathbf{E} = 1$ (exposed), from $\mathbf{E} = 1$ to $\mathbf{I} = 1$ (infectious) and from $\mathbf{I} = 1$ to $\mathbf{R} = 1$ (removed) are irreversible, whereas transitions corresponding to substitutions in nucleobases are reversible.

in Figure 3.2; $1 + 2 \times 4^s$ states in general, where s is the length of the observed sequence fragment).

3.2.2 Posterior distribution, approximations and MCMC

Estimation of parameters and latent variables is based on the joint posterior distribution $p(J, T^{inf}, L, D, \theta \mid data)$ of the transmission tree J, infection times $T^{inf} = (T_1^{inf}, \ldots, T_n^{inf})$, exposed (or latency) durations $L = (L_1, \ldots, L_n)$, infectious durations $D = (D_1, \ldots, D_n)$, and parameters θ that contain infection and dispersal parameters $\alpha = (\alpha_0, \alpha_1, \alpha_{2,1}, \alpha_{2,2})$, latency parameters $\beta = (\beta_1, \beta_2)$, infectiousness parameters $\delta = (\delta_1, \delta_2)$, mutation parameters $\mu = (\mu_1, \mu_2, \mu_3)$ and the date t_{exo} of the exogenous sequence S_{exo}.

The transmission tree J is a function from $\{1, \ldots, n\}$ to $\{0, 1, \ldots, n\}$ that states who infected whom: an observed individual i is infected by a pathogen source $j = J(i)$ that is either another observed individual $j \in \{1, \ldots, n\}$, $j \neq i$, or the disease reservoir (exogenous source) denoted by 0.

Data are removal times $T^{end} = (T_1^{end}, \ldots, T_n^{end})$, central locations $X = (x_1, \ldots, x_n)$ and observed sequences $S^{end} = \{S_1(T_1^{end}), \ldots, S_n(T_n^{end})\}$. The posterior distribution is:

$$\begin{aligned} p(J,T^{inf},L,D,\theta \mid data) &= p(J,T^{inf},L,D,\theta \mid S^{end},T^{end},X,S_{\text{exo}}) \\ &\propto p(S^{end} \mid J,T^{inf},L,D,\theta,T^{end},X,S_{\text{exo}})p(J,T^{inf},L,D,\theta \mid T^{end},X,S_{\text{exo}}) \\ &= p(S^{end} \mid J,T^{inf},L,D,\theta,T^{end},X,S_{\text{exo}})p(J,T^{inf} \mid L,D,\theta,T^{end},X,S_{\text{exo}}) \\ &\quad \times p(L,D \mid \theta,T^{end},X,S_{\text{exo}})p(\theta), \end{aligned} \tag{3.2}$$

where $\propto$ means "proportional to" (the multiplicative constant does not depend on the unknowns (J,T^{inf},L,D,θ)), $p(S^{end} \mid J,T^{inf},L,D,\theta,T^{end},X,S_{\text{exo}})$ is called the genetic likelihood, $p(J,T^{inf} \mid L,D,\theta,T^{end},X,S_{\text{exo}})$ is called the transmission likelihood, $p(L,D \mid \theta,T^{end},X,S_{\text{exo}})$ is the distribution of latency and infectious durations and $p(\theta)$ is the prior distribution of parameters.

These distributions (except the prior) are based on assumptions made in Subsections 3.1.4, 3.1.5 and 3.1.6. Their expressions, which are provided in Mollentze et al. (2014) and Soubeyrand (2016), will not be detailed here. We only highlight that the genetic likelihood raised computation issues. Indeed, the genetic likelihood can be formally written as a function of the unobserved virus sequences at the infection times $S_k(T_k^{inf})$, $k = 1,\ldots,n$, i.e. the transmitted virus sequences. To avoid to handle latent vectors $S_k(T_k^{inf})$ (the dimension of this set of unknowns is $n \times s$), Morelli et al. (2012) and Mollentze et al. (2014) replaced the genetic likelihood by a pseudo-likelihood. Essentially, this pseudo-likelihood is based on the probability of evolution between two observed sequences given the time lag separating the two sequences. The time lag is not simply the difference between the times of observation of the two sequences but a duration that depends on the topology of the transmission tree J. To compute the time lag, one has to *come back* to the most recent common ancestor (MRCA) of the two sequences and to sum the durations separating each of the two sequences and the MRCA. Another proposal was made in Soubeyrand (2016) to handle computation issues: replacing the genetic likelihood by an approximate likelihood where the unobserved transmitted sequences are deterministically reconstructed conditionally on the transmission tree J, the observed sequences S^{end} and the exogenous sequence S_{exo}. The reconstruction is based on the parsimony principle commonly used in phylogenetics: the most parsimonious reconstruction of $\{S_i^{inf} : i = 1,\ldots,n, J(i) > 0\}$ is the one that requires the fewest evolutionary changes (i.e. the fewest substitutions of nucleobases; see Tuffley and Steel, 1997, for a formal definition).

To estimate parameters and latent variables, in particular the transmission tree J, Morelli et al. (2012), Mollentze et al. (2014) and Soubeyrand (2016) built MCMC algorithms that sample in the posterior distribution (3.2), where the genetic likelihood was replaced by the pseudo-likelihood or by the approximate likelihood introduced above. For 30 observed cases, one MCMC run takes about 2 hours (the computation code was developed with the R Statistical Software and embedded Fortran subroutines). For 200 cases, we ran parallel MCMCs (typically 20 chains) during 15 days to obtain satisfactory results.

In a partially sampled epidemic, any given infected host that was sampled might have been infected by: (i) another sampled host (through direct transmission), (ii) an unsampled host that had been infected directly or indirectly by a sampled infected host (termed *indirect transmission* here) or (iii) an unsampled host that has no ancestor within the sample (termed *exogenous transmission* here). The model of Morelli et al. (2012) allows for only a single virus introduction (i.e. a single *exogenous* transmission) followed by direct transmissions for the rest of the epidemic, and the MCMC algorithm was constrained accordingly. Mollentze et al. (2014) extended this model by allowing multiple unobserved cases to arise anywhere in both space and time within the set of inferred transmissions. Conceptually, indirect and exogenous transmissions involving unobserved ancestors can be modeled in the same way —as being external to the sampled dataset. Thus, to reduce complexity and computation time, Mollentze et al. (2014) distinguished only between direct and *unsampled* sources in a primary MCMC and proposed a post-processing algorithm to distinguish between indirect and true exogenous transmissions.

3.3 Applications

Because the estimation method proposed above is designed for complex models (with high-dimension unknowns and strong dependence structure) and is based on model approximations, we were not able to theoretically justify its efficiency. Instead, we justified it with simulation studies (Sections 3.3.1–3.3.3) before applying it to real cases (Sections 3.3.4 and 3.3.5).

As shown below, the estimation method applied to simulated data leads to satisfactory inferences despite the high-dimension of unknowns. This is due to a combination of (i) the dependence structure inherent to the model, especially the time–genetic dependence and the time–space dependence, and (ii) strongly informative prior distributions for a few parameters, for which prior knowledge were available. The sections below do not detail the priors that were used. However, note that vague or slightly informative priors were generally used for parameters except for mutation parameters (μ_1, μ_2, μ_3 in Table 3.5) in Sections 3.3.1 and 3.3.4, and except for latency and infectiousness parameters ($\beta_1, \beta_2, \delta_1, \delta_2$ in Table 3.4) in Sections 3.3.2, 3.3.3 and 3.3.5.

3.3.1 Simulated outbreaks with single introductions

Morelli et al. (2012) simulated 100 genetic-space-time data sets of an outbreak caused by a single introduction of the virus within a set of 20 hosts. The first infected host was located at the bottom-left corner of a rectangular study domain and the 19 remaining hosts were independently and uniformly drawn in the domain. Each of the 100 data sets was obtained by simulating the epidemiological and micro-evolutionary dynamics with fixed parameter values and by collecting genetic, spatial and temporal data similar to those described

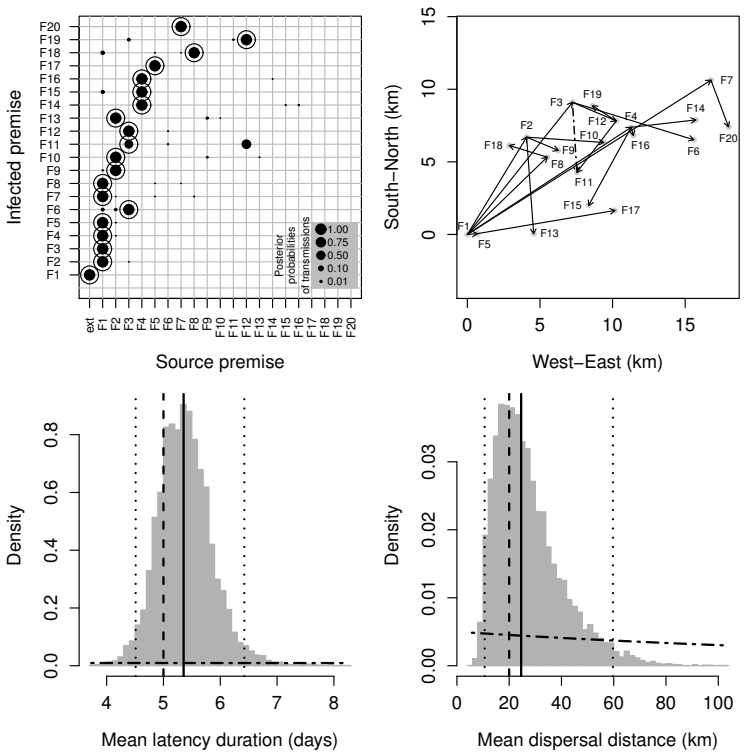


Fig. 3.3. Simulated outbreak infecting 20 hosts and estimation of transmission events, mean latency duration and mean transmission distance. Top left: true transmissions (circles) and posterior probabilities of transmissions (dot sizes are proportional to probabilities). Top right: tree with the highest posterior probability (solid arrows); Only transmission F12→F11 is not consistent with the true tree (the true transmission is F3→F11, dashed arrow). Bottom: posterior distributions (histograms) of mean latency duration (left) and mean dispersal distance (right); dashed lines: true values; dotted-dashed curves: prior distributions; solid lines: posterior medians; dotted lines: posterior quantiles 0.025 and 0.975.

in Section 3.2.1. The length of observed genetic sequences was $s = 8000$. Estimation was carried out with the MCMC algorithm based on the genetic pseudo-likelihood.

Figure 3.3 shows, for one simulated data set, the true transmission tree and its estimation as well as estimations of the mean latency duration and the mean transmission distance. Figure 3.4 shows, for the same simulated data set, the posterior distributions of infection times and latency durations. Only one true transmission (F3→F11) is not reconstructed accurately, the algorithm instead identifying F12→F11. However, the F3→F11 transmission has a high posterior probability and is included in the tree with the second

highest posterior probability (data not shown). Similarly, estimations of the mean latency duration, the mean dispersal distance, the infection times and the latency durations are satisfactory. Performance assessed from the 100 data sets are summarized in Table 3.6. Among the quantities that were analyzed, only the coverage rate of the standard deviation of the latency duration by its 95% posterior interval was rather low (0.43) certainly because this is a second order statistic.

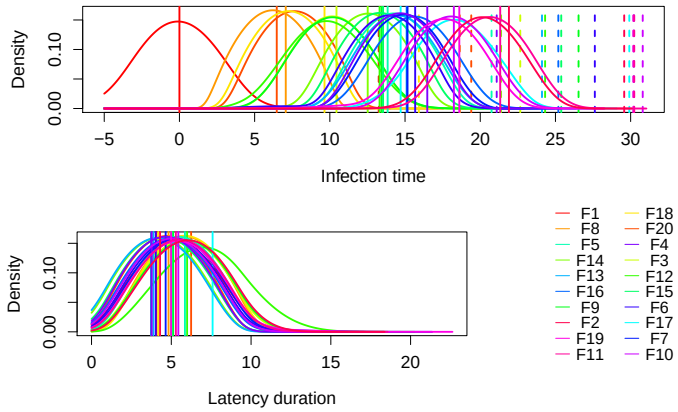


Fig. 3.4. Estimation of infection times and latency durations for the simulated outbreak infecting 20 hosts shown in Figure 3.3. Posterior distributions of infection times (top) and latency durations (bottom left). In both panels, vertical solid lines indicate the true values. In the top panel, vertical dashed lines indicate the virus observation times.

3.3.2 Simulated epidemics with multiple introductions – Case 1

Mollentze et al. (2014) assessed the accuracy of the method using 100 simulated datasets from each of six scenarios. Scenarios 1 to 4 were used to investigate overall accuracy and the effect of sampling rate on the reconstruction method, representing high (3/4 of all cases), moderate (2/3 of all cases), intermediate (1/2 of all cases) and low (1/4 of all cases) detection rates respectively. Scenarios 5 and 6 were used to test the sensitivity of the method to small and large misspecifications of epidemiological parameters. The misspecifications consisted of using strongly informative but biased prior distributions on the mean parameter values of the incubation and infectious durations (i.e. priors were centered on values different from those used in the simulations).

In this study, the simulation model contained a more realistic specification for the external source of infection than the inference model. While the infer-

Criterion	Value
Mean (Sd.) of posterior probabilities of true transmissions	0.85 (0.08)
Coverage rate of times of infection	0.78
Coverage rate of times of infectiousness (i.e. end of exposed stage)	0.93
Coverage rate of the source strength parameter α_1	0.97
Coverage rate of the dispersal parameter $\alpha_{2,1}$	0.89
Coverage rate of the mean latency	0.94
Coverage rate of the latency Sd.	0.43

Table 3.6. Performance of the estimation algorithm over a series of 100 simulated outbreaks infecting 20 hosts. The criteria used to assess the performance are the mean (and standard deviation noted Sd.) of the posterior probabilities of true pairwise transmissions, and the coverages by the 95% posterior intervals of the infection times, the times at which the hosts became infectious, the transmission parameters (source strength and dispersal parameter) and the latency mean and Sd.

ence model assumes a single external source with a constant infection strength (constant in both space and time), the simulation model allows for multiple sources of novel lineages, occurring both inside and outside the sampling region, with infection strengths that are localized in time and space. For each simulation, the epidemic was initiated at time zero with one infected host localized at the origin $(0, 0)$ and 119 susceptible hosts uniformly and randomly localized in the $[0.0, 0.3] \times [0.0, 0.1]$ rectangle. Genetic, spatial and temporal data were collected as described in Section 3.2.1 in a subregion of the rectangular domain (see Figure 3.5 displaying an example of simulated epidemic under Scenario 1). The length of observed genetic sequences was rather short: $s = 800$. Observing cases in a sub-region leads to consider an epidemic with multiple introductions. Estimation was carried out with an MCMC algorithm based on the genetic pseudo-likelihood followed by a post-processing analysis to distinguish between indirect and true exogenous transmissions (see Section 3.2.2).

Table 3.7 reveals that the reconstruction of direct transmissions remains fairly accurate regardless of sampling intensity (mean posterior probability of true transmission events > 0.73 for Scenarios 1–4). However, the reconstruction of transmission events is sensitive to the informative priors used for the incubation and infectious periods. This limits the suitability of the approach to diseases where the epidemiology is reasonably well known. Reconstruction of transmissions involving unobserved cases is moderately accurate at high sampling intensities, but becomes increasingly unreliable when 50% or less of the cases in the sampling region are sampled. At these sampling intensities the post-processing algorithm cannot accurately distinguish between indirect and exogenous connections.

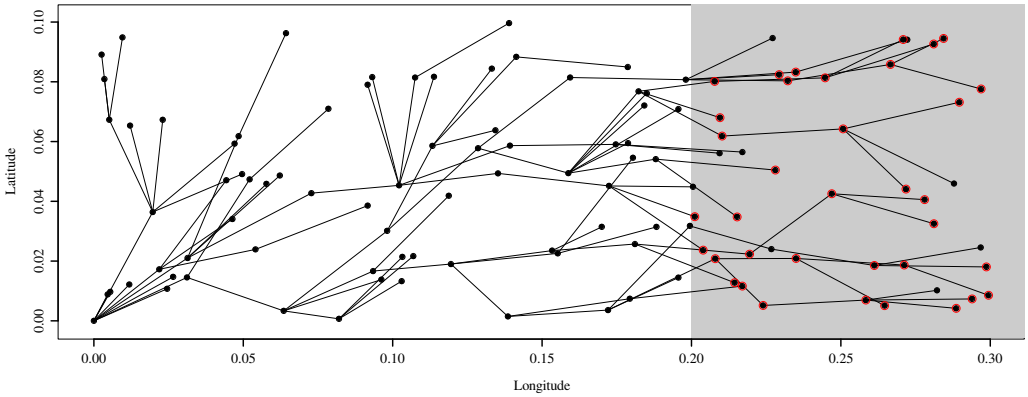


Fig. 3.5. Example of simulated epidemic over 120 hosts under Scenario 1. Black dots represent hosts, while black segments represent transmissions. Samples are taken from a sub-region, indicated in gray, and in this area, not all cases are detected or sampled. In simulations under Scenario 1, cases in the sampling area have a probability of 3/4 of being sampled. Sampled hosts are indicated in red.

Table 3.7. Performance of the estimation algorithm for the estimation of various transmission events. In each case the mean (and standard deviation) of the posterior probabilities of true transmission events is reported based on 100 simulations for each simulation scenario. Scenarios 1 to 4 have varying sampling rates in the study region, namely 3/4, 2/3, 1/2 and 1/4, respectively. Scenarios 5 and 6 correspond to small and large misspecifications of epidemiological parameters, respectively.

Transmission type	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Direct infection between observed cases	0.73 (0.12)	0.73 (0.10)	0.78 (0.13)	0.82 (0.12)	0.60 (0.15)	0.03 (0.03)
Infection of observed cases by unobserved[1] sources	0.64 (0.19)	0.62 (0.17)	0.54 (0.15)	0.57 (0.15)	0.73 (0.15)	0.76 (0.17)
Infection of observed cases by exogenous[2] sources	0.66 (0.22)	0.58 (0.20)	0.45 (0.18)	0.33 (0.14)	0.72 (0.18)	0.85 (0.17)
Direct or indirect infection between observed cases	0.63 (0.15)	0.62 (0.13)	0.57 (0.17)	0.40 (0.18)	0.54 (0.15)	0.08 (0.07)

[1] An unobserved source is a host that may have been infected, directly or indirectly, by an observed case.
[2] An exogenous source is a host that has not been infected, directly or indirectly, by an observed case.

3.3.3 Simulated epidemics with multiple introductions – Case 2

Soubeyrand (2016) simulated 50 data sets similar to those obtained under Scenario 1 in Section 3.3.2 (Scenario 1) and displayed in Figure 3.5. For each data set, four different estimation algorithms were applied:

- the first one used the genetic pseudo-likelihood and all genetic data;
- the second one used the genetic approximate likelihood and all genetic data;

- the third one used the genetic approximate likelihood and 50% of genetic data (50% of the available sequences were randomly selected and used for the inference, the other sequences were ignored);
- the fourth one used the genetic approximate likelihood and 25% of genetic data (25% of the available sequences were randomly selected and used for the inference, the other sequences were ignored).

In the two last cases, the number of sampled individuals remains unchanged, but genetic information that are exploited are reduced.

Table 3.8 shows that using the genetic approximate likelihood instead of the pseudo-likelihood allowed significant improvement of the identification of true endogenous sources when all genetic data are used (mean posterior probability of direct transmissions was increased from 0.75 to 0.82; p-value of pairwise t-test: 0.018). When the genetic sampling effort was decreased from 100% to 50% and 25%, the correct identification of both endogenous and unobserved sources was more uncertain but remained relatively high. Indeed, on average, there were 31 observed hosts in the simulated data sets and consequently, for each infected host there were 30 possible endogenous sources and 1 possible unobserved source. Therefore, obtaining a mean *posterior probability of true transmission* equal to 0.36 or 0.72 indicates that, in average, the true source is identified with a relatively high probability.

Table 3.8. Performance of the estimation algorithms for the estimation of various transmission events. Means (and standard deviations) of posterior probabilities of true transmission events are reported based on 50 simulations. Four estimation algorithms were applied, one based on the genetic pseudo-likelihood and three based on the genetic approximate likelihood with three different levels of genetic sampling effort. The genetic sampling effort is the percentage of observed cases for which the virus was sequenced and used in the inference.

Substitute of genetic likelihood	Pseudo-likelihood	Approximate likelihood		
Genetic sampling effort (%)	100	100	50	25
Direct infection between observed cases	0.75 (0.20)	0.82 (0.20)	0.48 (0.27)	0.36 (0.31)
Infection of observed cases by unobserved[1] sources	0.80 (0.18)	0.80 (0.18)	0.72 (0.21)	0.72 (0.21)

[1] An unobserved source is a host that may have been infected, directly or indirectly, by an observed case.

3.3.4 The 2007 outbreak of FMDV in the UK

The algorithm designed for outbreaks with single introductions was applied to a dataset collected during the 2007 outbreak of Foot-and-Mouth Disease

Virus (FMDV) in the UK, which infected 8 premises in Surrey and Berkshire (Cottam et al., 2008). The full virus sequences with $s = 8176$ nucleobases were used for the inference.

The reconstructed scenario with maximum posterior probability (Figure 3.6, top right) comprises two phases: IP1b was infected by an external source, and transmitted the virus to the neighboring premise IP2b and to IP5 further away; the virus remained contained and undetected on IP5 until it spread to a closeby premise IP4b; finally the virus spread from IP4b to the other premises. While the link made by IP5 between the two phases is highly supported, the estimation of the other transmissions was more uncertain: within the two clusters (IP1b, IP2b, IP5) and (IP5, IP4b, IP3b, IP3c, IP6b, IP7, IP8) several other transmission scenarios have non-negligible posterior probabilities (Fig. 3.6, top left). The mean estimated latency duration has a posterior median of 14 days and a 95%-credible interval of (6, 49) as shown in Fig. 3.6, bottom left; the long delay between the infection of IP5 and the subsequent transmissions is responsible for this result. The long distance between IP5 and its source (IP5 is 18.2 km away from IP1b) explains the large mean transmission distance (Fig. 3.6, bottom right), whose posterior median is 17 km and 95%-posterior interval is (5,58).

3.3.5 The endemic rabies dynamics in KZN, South Africa

The algorithm designed for multiple introductions was applied to a set of 176 canid-associated rabies cases detected between 1 March 2010 and 8 June 2011 in KwaZulu Natall (KZN), a South-East province of South Africa (Figure 3.7, left). This dataset contained 153 rabies cases detected in domestic dogs, 1 case detected in a jackal and 22 cases detected in domestic livestock. Livestock typically do not transmit rabies, and these cases were explicitly treated as dead-ends for transmission in the model. A sequence fragment of $s = 760$ nucleobases was used for the inference.

The majority of cases could not be linked through direct transmissions – 69 (95% posterior interval [PI]: 60-79) direct transmissions were identified, while unsampled sources were the most likely link for the remaining 107 (95%-PI: 97-117) cases (Figure 3.7, right, and 3.8). When considering only direct transmissions, there were several independent chains of transmission and many transmissions inferred to have taken place over long distances. The mean distance between the most probable directly connected cases was rather hight, namely 14.9 km (0.025- and 0.975-quantiles: 0.0 and 56.1 km; Figure 3.9, left). Occasional long-distance transmissions, particularly along the major highways that follow the KZN coast, have been identified before in the study area (based on phylogenetic patterns) and have been ascribed to motorized transportation of dogs (Coetzee and Nel, 2007). Road distances have also been shown to be a better predictor of rabies dissemination than absolute distances in northern Africa, again suggesting that humans may be responsible for long distance transmission of rabies (Talbi et al., 2010). The long distances and

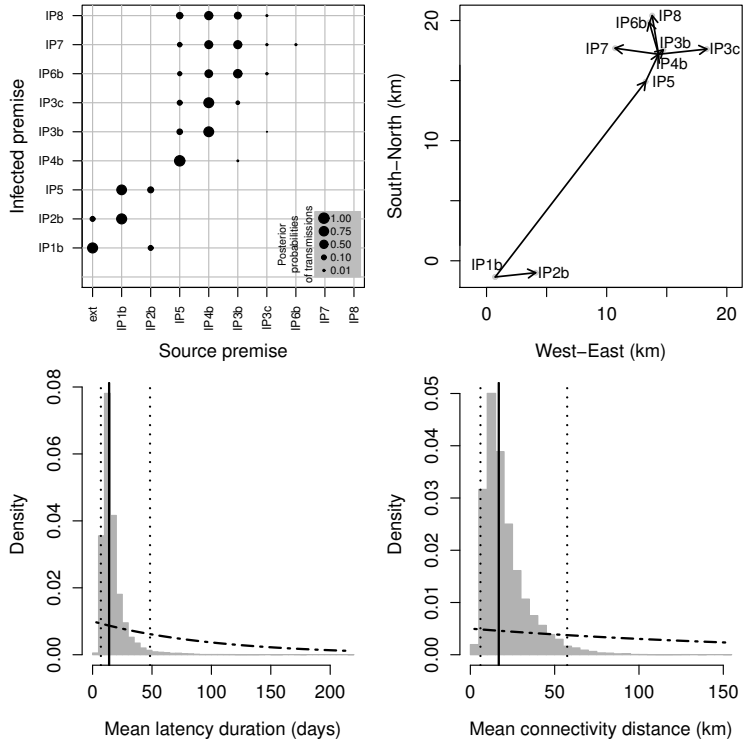


Fig. 3.6. Estimation output for the 2007 FMDV outbreak in the UK. Top left: posterior probabilities of transmissions (dot sizes proportional to probabilities). Top right: tree with the highest posterior probability mapped in space (black arrows). Bottom: posterior distributions (histograms) of mean latency duration (left) and mean transmission distance (right); dotted-dashed curves: prior distributions; solid lines: posterior medians; dotted lines: posterior quantiles 0.025 and 0.975.

short time-periods between cases in the transmission tree provide further evidence for motorized transportation of infected dogs, but such transmissions were not restricted to any one area and instead appear to be a common feature of the epidemiology of rabies in this area. This might be due to the high prevalence of circular human migration and migrant labour in many parts of KZN, with migrants visiting their rural households (and, it would seem, taking their dogs with them) on a regular basis (Posel and Marx, 2013). The analysis of indirect links using the post-processing algorithm mentioned in Section 3.2.2 is not presented here but can be found in Mollentze et al. (2014). Figure 3.9, right, shows the posterior distribution of the number of nucleotide substitutions between the pathogen sequences collected in pairs of directly connected cases. The large probability of zeros (no difference between sequences) partly explains the high uncertainty in the inferred transmission

links (see Figure 3.8) and highlights the necessity of combining heterogenous data, namely spatial, temporal and genetic data, to infer transmission trees and associated epidemiological parameters.

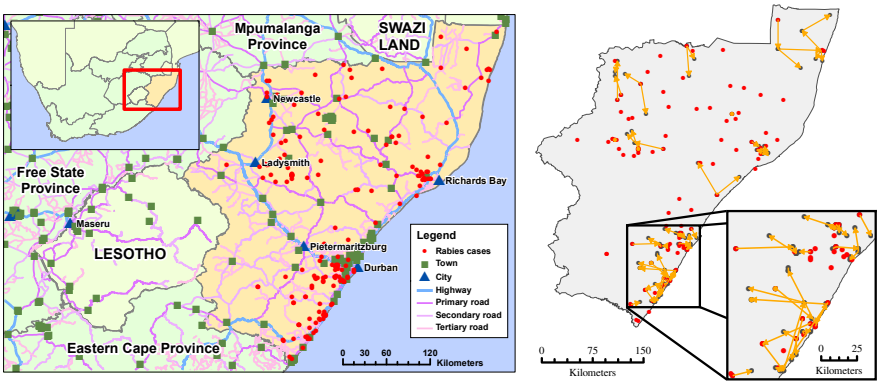


Fig. 3.7. Left: Detailed map of KwaZulu Natal (KZN) showing the cases detected between 1 March 2010 and 8 June 2011 in the context of major roads, towns and cities. Right: Transmission trees showing the direct pairwise transmissions with highest posterior probabilities. Transmission links between cases are represented by orange arrows (and dots when a transmission links cases at the same location), while red dots represent cases for which no direct ancestor was detected. The inset shows an enlarged view of connections in the southern coast of KZN, were the majority of cases where detected.

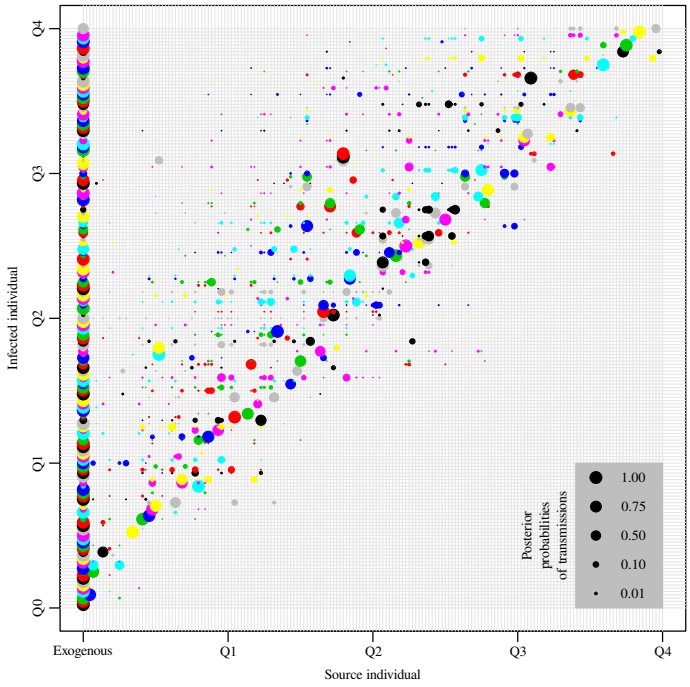


Fig. 3.8. Graphical representation of the posterior distribution of inferred direct sources. Individuals are arranged by observation date on both axes, with each infected individual (horizontal rows) indicated in an alternating color for clarity. Q0 indicates the start of the sampling period, while Q1–Q4 indicate the ends of quarters of the sampling period. "Exogenous" indicates infection from an external source, encompassing indirect transmissions and introductions from outside the dataset.

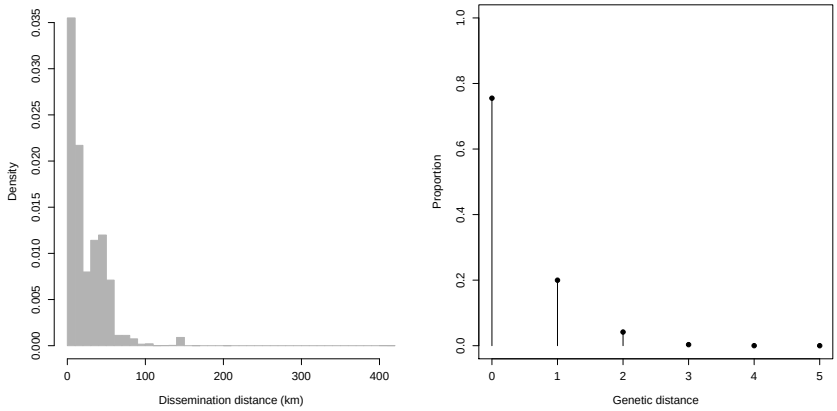


Fig. 3.9. Left: Posterior distribution of the transmission distances between directly connected cases. Right: Posterior distribution of the number of genetic differences between directly connected cases.

4

PDE-based mechanistic-statistical modeling

Author's references: Roques et al. (2011), Roques et al. (2014), Roques et al. (2016), Soubeyrand and Roques (2014).

Mathematical models are widely used in ecology, evolutionary biology and epidemiology to test hypotheses, simulate scenarios and make predictions. Countless mechanistic models, including analytical and agent-based models, have been developed, and their behaviors have been studied by determining stable equilibria and e.g. performing sensitivity analyses. Statistical models have been used to infer the values of model parameters. Each type of model has been criticized. For statistically orientated scientists, analytical models are typically over-simplified and agent-based models do not control for error propagation. For other scientists, statistical models are poorly and non-mechanistically linked to biological processes. Fortunately, with advances in statistical and computer sciences, it is now becoming possible to reconcile the different modeling approaches, and various kinds of hybrid mechanistic-statistical models can be constructed and fitted rigorously to data. These models, combining a *process model* and a *data model*, are often built as state-space models where the hidden layer is mechanistically constructed. They have been called physical-statistical models (Berliner, 2003) or mechanistic-statistical models (Soubeyrand et al., 2009c,d). Other examples of such models are given by Buckland et al. (2004), Rivot et al. (2004), Wikle (2003b) and Roques et al. (2011, see Section 4.2) in ecology, and by Campbell (2004), Wikle (2003a) and Roques et al. (2014, see Section 4.3) in environmental science.

Most of the mechanistic-statistical models that I have built contain a *process model* that is stochastic. For instance, in Section 2.5, we have seen a rather complex mechanistic-statistical model allowing for the analysis of the multi-year dynamics of a metapopulation. This model contains a *process model*, which is essentially based on an inhomogeneous spatio-temporal point process. Thus, the mechanistic part of the model is stochastic. In Soubeyrand et al. (2009d), we proposed another mechanistic-statistical model for analyzing the dynamics of a pest, namely the pine sawfly, across Finland and over three decades. Here also, the mechanistic component of the model is stochastic because it is based on a spatio-temporal gamma process including random jumps.

Such models based on stochastic *process models* generally contain numerous latent variables, which imply the use of particularly computer-intensive estimation algorithms. For the analysis of ecological and epidemiological dynamics, an alternative approach to building the *process model* is to use reaction-diffusion equations corresponding to a particular type of parabolic partial differential equations (PDE). This approach results from a long tradition: the family of PDEs adapted to population dynamics has been enriched for more than 60 years (Skellam, 1951; Holmes et al., 1994). This long-term research has produced numerous PDEs in good agreement with dispersal and reproduction properties of populations observed in nature as well as experimental systems (e.g. see Murray, 2002; Okubo and Levin, 2002; Shigesada and Kawasaki, 1997). The reason why reaction-diffusion for population dynamics deserves to be used in the mechanistic-statistical approach lies in their conciseness for taking into account relatively complex spatio-temporal dependencies: the spatio-temporal evolution of the population density, say $u(t,x)$ (where t denotes a time and x denotes a location), is governed via an equation where derivatives with respect to time (generally $\partial u/\partial t$) and derivatives with respect to space (generally $\partial^2 u/\partial x^2$) interplay. However, this formalism leads to equations that are generally viewed as limit models for large populations or as models for the expectation of the population dynamics —or higher-order moments. Therefore, reaction-diffusion equations have not to be considered as the panacea for modeling all population dynamics[1], but they can be advantageously utilized in a mechanistic-statistical approach as a way to mimic the main trend of the population dynamics.

In Wikle (2003b), a PDE-based mechanistic-statistical model is fitted to data collected during the spread of a bird population in North America. In this work, Wikle incorporated several noise processes into the reaction-diffusion equation, in addition to the randomness incorporated in the *data model*. His motivation was to obtain a more realistic process model than the pure reaction-diffusion equation. Generally, incorporating noise into the reaction-diffusion equation negatively impacts the computation effort to fit the model to data. Thus, there is a trade-off between the *freedom degree of a PDE-based model* and the *computation cost of the associated estimation algorithm*.

In this chapter, we present studies where reaction-diffusion equations (without noise or including latent processes) are used to build the *process model*. Section 4.1 proposes the construction of such models to study biological invasions. Section 4.2 illustrates the proposed approach with a real-life example, namely the expansion of the pine processionary moth in northern France. Section 4.3 illustrates the proposed approach in the context of long-term climatic dynamics. We present the latter example despite the large dif-

[1] Obviously, reaction-diffusion equations are not adapted, for example, to describe individual-to-individual transmissions of a virus studied in Chapter 3.

ference between climatic issues and epidemiological issues because this case allows us to show that the *data model* can be quite sophisticated[2].

4.1 Parameter estimation for reaction-diffusion models of biological invasions

Consider a reaction-diffusion equation which models the spread of an invasive species in a domain Ω included in $\mathbb{R}^d$, with $d = 1$ or $d = 2$:

$$\frac{\partial u}{\partial t} = D\,\Delta\, u + u(r - \gamma u), \tag{4.1}$$

where $u = u(t, x)$ is the population density at time $t > 0$ and space location $x \in \Omega$; $D > 0$ measures the dispersion rate; the operator $\Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$ stands for the spatial dispersion operator; the coefficient $r \in \mathbb{R}$ corresponds to the intrinsic growth rate of the species (that is, the growth rate in the absence of competition); and $\gamma > 0$ measures the effect of competition. Given some conditions on the boundary of Ω and an initial condition $u_0(x) = u(0, x)$, the equation (4.1) is well-posed in the sense that it admits a unique solution $u(t, x)$, for all $t > 0$ and $x \in \Omega$.

The parameter vector that we want to infer, say θ, can consist of several constants or functions incorporated in the model, e.g. D, r, γ and u_0. Ideally, if observations are noise-free (i.e. one observes exactly the solution of the equation, at least in some points in time and space), then one can use results from the inverse problem approach to estimate the model parameters. However, observations are generally noisy. In this case, one can adopt the state-space approach (or mechanistic-statistical approach) based on hierarchical modeling and statistical inference.

Broadly speaking, for biological invasions, inverse problems mainly have a theoretical interest: one investigates the existence and the uniqueness of parameters. However, the inverse problem approach has also a practical interest: it can be shown that very sparse (but noise-free) information can be sufficient to determine parameter values. The implication of such a result in a real-life study is that one can expect, at least for moderately noisy observations, to estimate parameters even with sparse data. Nevertheless, in a real-life study, one has to take into account the noise in the observations by building a model that connects the mechanistic vision of the studied phenomenon to a stochastic vision of the observation of this phenomenon. The observations are considered as stochastic because the quantities of interest (e.g. population

[2] An other example of a relatively sophisticated *data model* is given in Roques et al. (2016), where we adopted a PDE-based mechanistic-statistical approach to estimate a spatially heterogeneous rate of diffusion. In this work, instead of using data measuring the intensity of the population, like in Section 4.1, we use genetic data.

abundance) are often indirectly observed and because the observation process often implies a loss of information (e.g. spatial and temporal censorship, measurement uncertainty, binarized signal). The coupling of the mechanistic vision of the studied phenomenon with the stochastic vision of the observation can be made with the mechanistic-statistical approach.

In Soubeyrand and Roques (2014), we illustrated both the inverse problem approach and the mechanistic-statistical approach. The following sections provide a simulated example of the latter approach: we estimate the date and location of the introduction of an invading species using noisy observations which consist of impacts of the invading species towards the environment measured at discrete times and at discrete locations.

4.1.1 Model

Mechanistic model

The model for the dynamics of the species is the reaction-diffusion equation (4.1) in a two-dimensional domain $\Omega \subset \mathbb{R}^2$, which is depicted in Figure 4.1. The quantity u is assumed to obey Neumann conditions on the boundary $\partial\Omega$ of Ω: $\frac{\partial u}{\partial \nu} = 0$ almost everywhere on $\partial\Omega$, where ν is the outward unit normal to $\partial\Omega$. Besides, we assume that the equation is verified for all time $t > -t_0$, where $t + t_0 > 0$ corresponds to the time since the introduction of the species:

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\,\Delta u + u(r - \gamma u),\ t > -t_0,\ x \in \Omega, \\ \dfrac{\partial u}{\partial \nu}(t,x) = 0,\ t > -t_0,\ x \in \partial\Omega, \\ u(-t_0, x) = u_0(x),\ x \in \Omega. \end{cases} \tag{4.2}$$

All the coefficients D, r and γ are assumed to be constant in Ω. The function $u_0(x)$ corresponds to the initial density of the founding population at the date of introduction $-t_0$. The initial population density is assumed to be an exponential function with known shape but unknown location of introduction. More precisely, we assume that

$$u_0(x) = \exp\left(-20\,\|x - x_0\|\right),\ \text{for } x \in \Omega,$$

where $x_0 \in \Omega$ corresponds to the location of introduction and $\|\cdot\|$ is the Euclidean norm.

In this example, we aim to estimate the time t_0 and the location of introduction x_0 together with parameters D, r and γ; let $\theta = (D, r, \gamma, t_0, x_0)$.

Model of the observation process

Concerning the sampling scheme, the observations are carried out in $I = 6$ subdomains $\omega_1, \ldots, \omega_6 \subset \Omega$ (see their locations in Figure 4.1) and at $J = 10$

times $\tau_j = 0.1\,(j-1)$ for $j = 1, \ldots, 10$ in each subdomain. We assume that the impact $\bar{Y}_{ij}$ of the species in the domain ω_i and at the time τ_j is proportional to the number of individuals inside this subdomain at that time:

$$\bar{Y}_{ij} = \alpha \int_{\omega_i} u(\tau_j, x)\, dx,$$

for some known constant α which measures the mean impact per unit of population density. The measurements Y_{ij} of the impacts are assumed to follow independent Poisson distributions with mean values $\bar{Y}_{ij}$. Thus, the probability density function of $\mathbf{Y} = (Y_{ij})_{1 \le i \le I, 1 \le j \le J}$ is:

$$p(\mathbf{Y}; \theta) = \prod_{1 \le i \le I, 1 \le j \le J} \exp\left\{-\bar{Y}_{ij}\right\} \frac{\left\{\bar{Y}_{ij}\right\}^{Y_{ij}}}{Y_{ij}!}. \tag{4.3}$$

4.1.2 Estimation and results

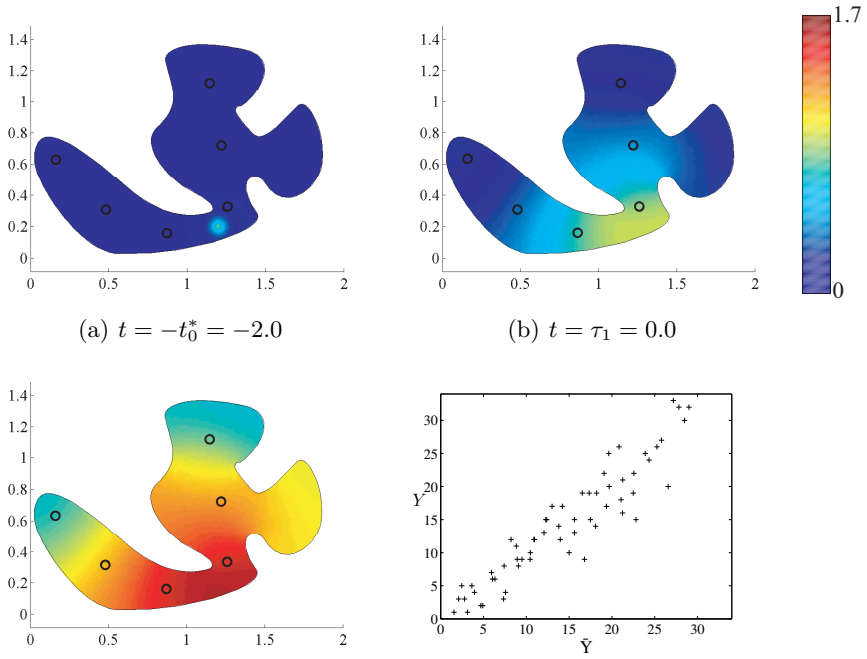
We simulated the model (4.2) in Ω during the time period $t \in (-t_0, 1)$ with the following values for the parameters: $D^* = 5 \cdot 10^{-2}$, $\gamma^* = 1$, $r^* = 2$, $t_0^* = 2$ (unit of time) and $x_0^* = (1.2, 0.2)$; let $\theta^* = (D^*, \gamma^*, r^*, t_0^*, x_0^*)$. Figure 4.1 shows the simulation of the population density at the origin $-t_0$ and at the sampling times $\tau_1 = 0$ and $\tau_{10} = 0.9$. It also shows the cloud of points $\{\bar{Y}_{ij} \times Y_{ij} \ : \ i = 1, \ldots, 6,\ j = 1, \ldots, 10\}$.

We chose a uniform prior distribution for the parameter vector θ:

$$\begin{aligned}\pi(\theta) = {} & \tfrac{1}{0.99 \times 9.9 \times 20 \times 10 \times |\Omega|} \mathbf{1}(10^{-2} < D < 1,\ 0.1 < \gamma < 10) \\ & \times \mathbf{1}(-10 < r < 10,\ 0 < t < 10,\ x_0 \in \Omega),\end{aligned}$$

where $\mathbf{1}$ is the indicator function, and we drawn a sample from the posterior distribution of θ with an MCMC algorithm including Metropolis-Hastings updates. The crucial advantage of defining the mechanistic model as a PDE lies in the following tricks: (i) the likelihood (and the acceptance probabilities in the Metropolis-Hastings updates) can be computed rapidly if the PDE can be solved rapidly, and (i) no latent variables has to be updated in the MCMC algorithm (because the values $\bar{Y}_{ij}$ are deterministic functions of θ).

The marginal posterior distributions of D, γ, r, t_0 and x_0 are presented in Figure 4.2. This toy example shows that the introduction date and location of an invading species satisfying a reaction-diffusion equation can be correctly estimated. In real situations, however, the inference might be less accurate, especially because the reaction-diffusion equation incorporates strong regularity assumptions about the population dynamics whose actual behavior might be not much regular.



(a) $t = -t_0^* = -2.0$ (b) $t = \tau_1 = 0.0$

(c) $t = \tau_{10} = 0.9$ (d) Real vs observed impacts

Fig. 4.1. Solution of the model (4.2) with parameter θ^* at times -2.0, 0.0 and 0.9 in panels (a), (b) and (c), respectively. The study domain Ω is delimited by the contour line, and the six sampling subdomains $\omega_1, \ldots, \omega_6$ are indicated by circles. Panel (d): Real impact $\bar{Y}_{ij}$ versus observed impact Y_{ij} for $i = 1, \ldots, 6$ and $j = 1, \ldots, 10$.

4.2 Application to the expansion of the pine processionary moth

Recent studies have reported a northward geographic range expansion of the pine processionary moth (*Thaumetopoea pityocampa*, Lepidoptera: Notodontidae, abbreviated as PPM below). In the Paris Basin, France, its range has shifted 87 km northward between 1972 and 2004, with a notable acceleration (55 km) during the last 10 years (Battisti et al., 2005; Robinet et al., 2007).

Because of its impact on forests, this expansion is likely to have important ecological consequences. It may also cause sanitary issues. The PPMs are entering semi-urban and urban areas; therefore, the insect has progressed from mere forest pest to urban medical threat. The threat arises from the way these organisms protect themselves against predation. When threatened, mature larvae release irritant hairs that cause allergic reactions in both man and warm-blooded domestic animals; reactions range from the cutaneous type to anaphylactic shock (Doutre, 2005).

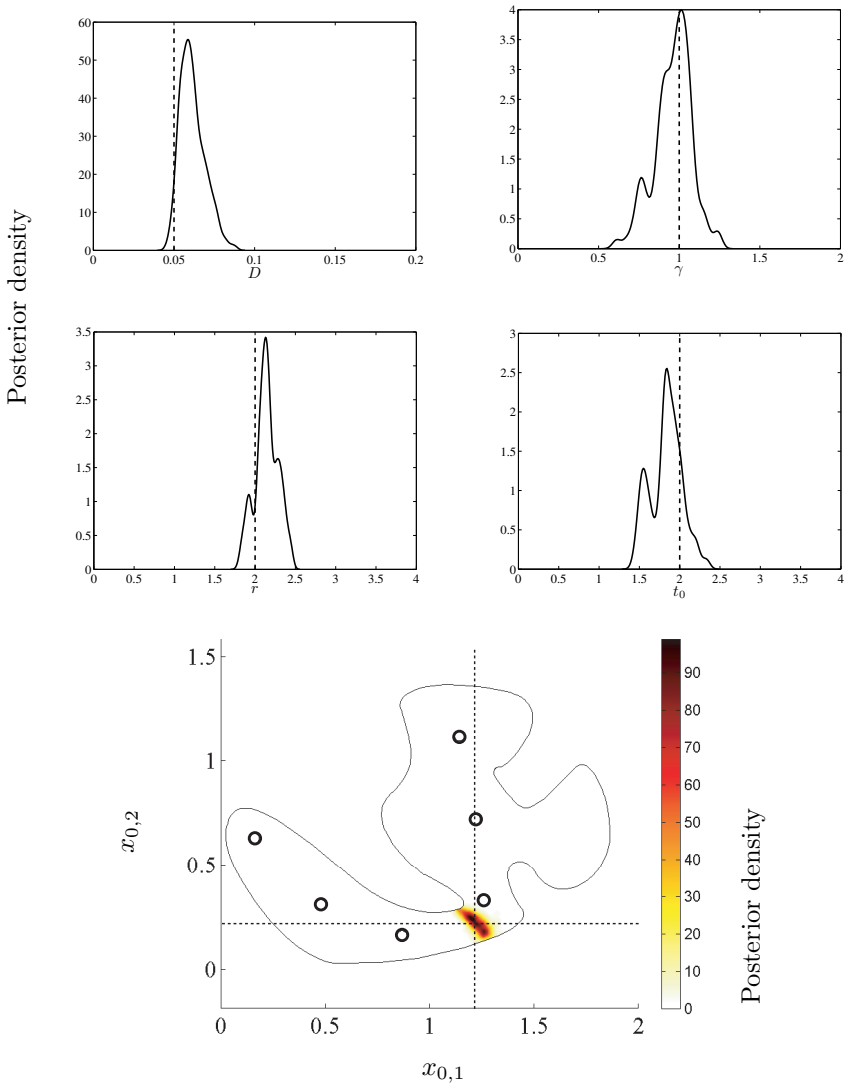


Fig. 4.2. Posterior distributions of D (top left; prior: Uniform($[0.01, 1]$)), γ (top right; prior: Uniform($[0.1, 10]$)), r (center left; prior: Uniform($[-10, 10]$)), t_0 (center right; prior: Uniform($[0, 10]$)) and $x_0 = (x_{0,1}, x_{0,2})$ (bottom; prior: Uniform(Ω)). In all panels, the dashed lines indicate the position of the true parameter values. In the bottom panel, Ω is delimited by the contour line, and the six sampling subdomains are indicated by circles.

Extensive measurements that have been carried out at different spatial scales in France show that the northern range of the PPM is not regular. This indicates that population expansion is faster in some regions. Determining these regions is of crucial importance for controlling and preventing PPM expansion.

In Roques et al. (2011), we investigated possible heterogeneity in the expansion of PPM in the Paris Basin. Our goal was to build a map that describes where the environment is favorable / unfavorable to PPM expansion. In this aim, we developed a mechanistic-statistical approach for analyzing the spatial variations in the range expansion of PPM by using binary measurements (i.e. presence / absence of PPM nests in sampling cells). The proposed method allowed us to infer the local effect of the environment on PPM population expansion. This effect is estimated at each position x using a parameter $F(x)$ that corresponds to the local fitness of PPMs.

4.2.1 Data

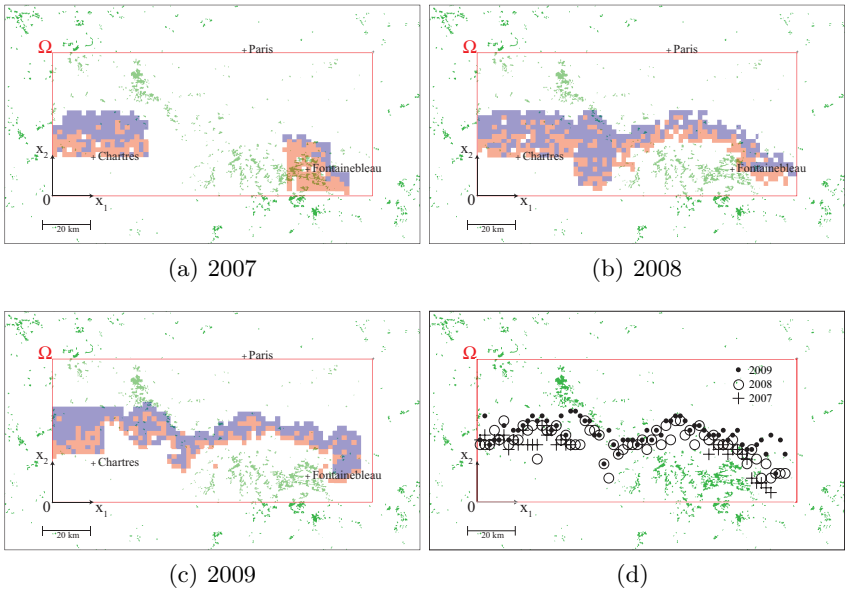
The life cycle of the PPM usually lasts for one year and can be divided into two main stages: (i) the adult stage and (ii) the larval stage. The adult stage starts at the beginning of the summer when adult moths emerge from the soil and begin taking flight. Next, mating and spreading occurs. Females lay 70-300 eggs, which are usually deposited simultaneously on pine trees. Larvae emerge from eggs during the second half of summer. Immediately after emergence, they build a common silk nest on the pine where eggs were deposited. At the beginning of spring, the larvae leave the nest and dig into the soil where they transform into pupae and remain for a few months until the next adult stage.

The clutch size, laying frequency and survival rates during the larval and adult stages may be influenced by environmental factors. Therefore, PPM fitness may depend on the spatial position of the individuals, and we aimed to estimate this local PPM fitness.

The study domain is a rectangular region Ω (134 km $\times$ 60 km) located in the Paris Basin; see Figure 4.3. The PPM range has been measured in 2007, 2008 and 2009 through direct observations of the presence of PPM nests in pines. The study region was mapped into a lattice made of $I = 2010$ square cells ω_i of the same size 2 km $\times$ 2 km. For each year n, $J_n < I$ cells were observed; the J_n sampled cells varied with time. Moreover, only binary data (presence or absence of PPM nests) have been recorded; see Figure 4.3 (a)-(c). These data indicate a northward range expansion of the PPM; see also Figure 4.3 (d).

4.2.2 Model

The mechanistic-statistical model that we developed for analyzing PPM expansion is quite complex. Below, we only present its skeleton.



(a) 2007 (b) 2008

(c) 2009 (d)

Fig. 4.3. Observed presence / absence of PPM nests in 2007 (a), 2008 (b) and 2009 (c). Blue squares in the study site Ω correspond to observed cells ω_i where PPM have not been detected. Red squares correspond to cells ω_i where PPM nests have been detected. (d): Position of the northernmost points where PPM nests have been detected during years 2007, 2008 and 2009. In each panel, green dots indicate sites with high densities of pine trees, but pine trees are present at least at low densities in most of the study domain.

Mechanistic model

The nest density evolves through a discrete-time process with a yearly time step. However, this evolution results from the dispersal and laying of adult PPMs, which is a continuous-time process. Our aim was to build a model which expresses nest density as a function of the adult density during the whole adult stage and of an environmental factor $F(x)$.

Let $v_n(t,x)$ be the density of adult PPMs at time t of year n and at location x and let $v_{n,0}(x) = v_n(t_{n,0},x)$ denote the density of adult PPM at the beginning $t_{n,0}$ of the adult stage of year n.

Let $w_n(t,x)$ be the cumulated density of adult PPMs at time t in the adult stage of year n. Besides, let $w_n^*(x) = w_n(t_n^*,x)$ where t_n^* is the end time of the adult stage of year n.

Let $u_n(x)$ denote the density of nests at location x and at the end of year n.

The mechanistic model establishes the links between the spatial and spatio-temporal functions defined above.

First, the initial density $v_{n,0}(x)$ (i.e. the density of emerging adults in year n) is obtained as a function of the density of nests $u_{n-1}(x)$ at the previous year (since larvae in nests leads to adults):

$$v_{n,0}(x) = r(u_{n-1}(x))u_{n-1}(x), \tag{4.4}$$

where $r(u)$ gives the number of emerging adults per nest unit when the local density of nests is u. Taking into account an Allee effect, the function r was defined by $r(u) = Ru/(1+u)$, where R is a positive constant.

Second, the density of adults $v_n(t,x)$ was assumed to satisfy a reaction-diffusion equation (during the adult stage) where the reaction is only composed of mortality: for any $t \in (t_{n,0}, t_n^*]$ during the adult stage of year n and for any $x \in \Omega$,

$$\frac{\partial v_n}{\partial t} = D\Delta v_n - \frac{v_n}{\nu}, \tag{4.5}$$

where D is the diffusion parameter and ν is the life expectancy (at each time unit a fraction $1/\nu$ of the individuals die). The initial condition for this PDE is given by $v_{n,0}$ defined in Equation (4.4) and no-flux conditions were assumed at the borders of a rectangular domain[3] including Ω.

Third, the cumulated population density at time t of year n and at position x, which is defined by:

$$w_n(t,x) = \int_{t_{n,0}}^{t} v_n(s,x)\, ds,$$

also satisfies a reaction-diffusion equation. Indeed, integrating Equation (4.5) between $t_{n,0}$ and t leads to:

$$\frac{\partial w_n}{\partial t} = D\Delta w_n - \frac{w_n}{\nu} + v_{n,0}, \quad t \in (t_{n,0}, t_n^*],\ x \in \Omega, \tag{4.6}$$

with $w_n(t_{n,0}, x) = 0$. The quantity $w_n^*(x)$ is obtained by solving Equation (4.6) at $t = t_n^*$.

Fourth, we assumed that the density of nests at year n results from the local cumulated density of adults during year n and satisfies:

$$u_n(x) = \min\left\{w_n^*(x)F(x), K(x)\right\}, \tag{4.7}$$

where $w_n^*(x)F(x)$ corresponds to the nest density which would be obtained at the end of year n for a non-constraining carrying capacity, and $K(x)$ corresponds to the spatially heterogeneous carrying capacity[4]. The so-called local fitness $F(x)$ is supposed to depend on local environmental factors but not on the carrying capacity which is already taken into account by Equation (4.7) via $K(x)$.

[3] This rectangular domain including Ω is ignored in the following to simplify the presentation of the model in this document. Details about this larger domain and its role in the model and in the inference are given by Roques et al. (2011).

[4] In this application, $K(x)$ was assessed by smoothing the map of pines shown in Figure 4.3 since nests are formed in pines.

Model of the observation process

As explained in Section 4.2.1, the study domain Ω was divided into I square cells ω_i with same area $\rho = 4$ km^2. Discrete time is indexed by $n = 0, \ldots, N$; the interval between n and $n+1$ corresponds to one year (=one cycle). We denote by $Y_n(i)$ the binary variable that takes the value 1 if PPM nests have been detected and 0 if no nest has been detected in the cell ω_i at year n.

If a cell ω_i has been observed during year n, the probability that $Y_n(i) = 1$ depends on the local nest density in the cell ω_i. We assumed that the detection variables were independently drawn from the following Bernoulli distributions:

$$Y_n(i)|u_n \underset{\text{indep.}}{\sim} \text{Bernoulli}\left\{1 - (1-p)^{\int_{\omega_i} u_n(x)dx}\right\}, \tag{4.8}$$

where $\int_{\omega_i} u_n(x)dx$ is the density of nests in the sampling cell ω_i, and p is the detection probability of one nest unit in one unit area. The probability p was determined via additional data assessing the quality of detection. The success probability of the Bernoulli distribution was obtained via the probability of not detecting $\int_{\omega_i} u_n(x)dx$ nest units by assuming that the observations of the $\int_{\omega_i} u_n(x)dx$ units are independent.

4.2.3 Estimation

We aim to estimate the local fitness $F(x)$ for $x \in \Omega$ and the diffusion parameter D of the mechanistic-statistical model presented above, the other parameters and functions in the model being given (including an initial value for u_n for a given year n before the first sampling year). Using Equation (4.8), the likelihood can be simply written as a function of F and D.

To handle the estimation of F we supposed that F is piecewise constant: Ω was discretized into $N = 35 \times 15 = 525$ rectangular subcells of the same size, and $F(x) = F_j$ for all x in the cell $j \in \{1, \ldots, N\}$. Thus estimating F becomes equivalent to estimating the values F_j.

In the absence of further information we assumed independent uniform prior distributions in $[0, F^{\max}]$ of the parameters F_j, and a uniform prior distribution (independent of F) in $[0, D^{\max}]$ of the parameter D:

$$F_j \sim \text{Uniform}(0, F^{\max}),\ j = 1, \ldots, N \text{ and } D \sim \text{Uniform}(0, D^{\max}).$$

From the definition of $F(x)$, and because each female can bear at most 300 eggs with a sex-ratio close to 1 : 1, we fixed $F^{\max} = 150$. The value of $D^{\max}$ was set to[5] 30km^2/day.

[5] When the diffusion coefficient is equal to D, the average dispersal distance of the individuals after τ days is $L = \sqrt{\pi\,\tau\,D}$ km. Whenever $\tau = 1$ (i.e. the life expectancy) we have for a low value of D $L(\tau = 1, D = 10^{-3}) = 0.06$ km and for $D = D^{\max}$ $L(\tau = 1, D = 30) = 9.7$ km.

We drawn a posterior sample of the parameters with a MCMC algorithm including Metropolis-Hastings updates. Like in Section 4.1.2, the likelihood (and the acceptance probabilities in the Metropolis-Hastings updates) can be computed rapidly because the PDE can be solved rapidly. However, here the number of parameters is much higher since we have to update 526 parameters (the F_js and D) in the MCMC algorithm.

4.2.4 Results

Posterior distribution of the fitness parameter F

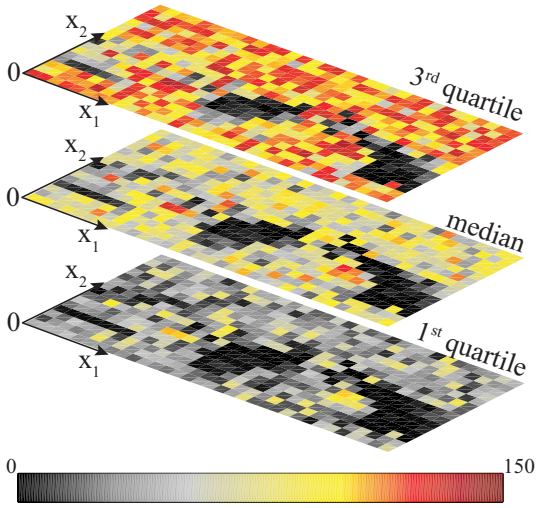


Fig. 4.4. Pointwise first, second (median) and third quartiles of the posterior distribution of the fitness function F in the domain Ω.

The posterior quartiles of F_j $(j = 1, \ldots, N)$ are shown in Figure 4.4. The distribution of F is clearly different from the prior distribution. This indicates that our binary observation data do carry information about the distribution of F. We can also observe that the distribution of the fitness F is spatially heterogeneous (i.e. the posterior distribution of F_i strongly depends on the position of the cell i) and spatially structured[6] (i.e. close regions tend to have close fitnesses). Thus, we notice several large unfavorable regions (black regions in Figure 4.4).

The posterior distribution of F is not strongly correlated with the host density shown in Figure 4.3 which was used to define the spatially heterogeneous carrying capacity K. This is a consequence of Equation (4.7) which

[6] A permutation test was built and applied to demonstrate the spatial structure.

incorporates both F and K such that F is defined for a non-constraining carrying capacity. Thus, F is not a simple function of K but the environmental factors determining F have to be disentangled.

Posterior distribution of the diffusion parameter D

The posterior distribution of the diffusion parameter D is shown in Figure 4.5. The posterior median of D is equal to 9.3 (mean 9.4 and standard error 0.4). This value $D = 9.3$ corresponds to an average dispersal distance equal to $\sqrt{\pi\tau D} = 5.4$ km when $\tau = 1$ (i.e. the life expectancy). This is higher than usually observed for Lepidoptera (see Kareiva (1983) and Shigesada and Kawasaki (1997), page 55) and may indicate that the dispersal is not purely diffusive[7].

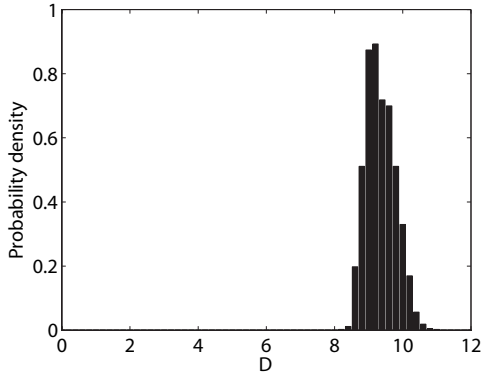


Fig. 4.5. Posterior distribution of the diffusion parameter D.

4.3 Side topic: Parameter estimation for climatic energy balance models with memory

In Roques et al. (2014), we studied parameter estimation for one-dimensional energy balance models with memory (EBMM) given localized temperature measurements. EBBMs are simple climatic models belonging to the class of nonlinear parabolic PDEs with delay terms. Adopting the inverse problem approach, we first shown that a space-dependent parameter can be determined

[7] A current work on the spatio-temporal dynamics of poplar rust led us to build a mechanistic-statistical model based on an integro-differential equation instead of a PDE. Thus, we are able to infer non-diffusive dynamics.

uniquely everywhere in the PDE's domain of definition, using only temperature information in a small subdomain. This result is valid only when the data correspond to exact measurements of the temperature. However, at the large temporal scales (e.g. several thousand of years), temperature measurements are noisy. For example, for temperatures reconstructed from ice cores or marine-sediment cores, data contain two sources of uncertainty: (i) in the value of the measured temperature; and (ii) in the accuracy of the dating, which tends to decrease as samples are derived from earlier time points (Salamatin et al., 1998). In this context, we proposed a mechanistic-statistical approach for estimating a space-dependent parameter of an EBBM. This approach is described below.

4.3.1 Model

Mechanistic model

We assume that the spatio-temporal dynamics of the temperature is governed by the following EBMM: for time $t > 0$ and location $x \in (0, 1)$ (corresponding for example to a latitude between the equator and the north pole of the Earth),

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + \alpha(x)(1 - a(T)) - g(T) - \left(\frac{1}{\tau}\int_{-\tau}^{0} T(t+s)\,\mathrm{d}s\right)^3. \tag{4.9}$$

Equation (4.9) includes

- a heat diffusion term $\frac{\partial^2 T}{\partial x^2}$;
- an incoming radiation term $\alpha(x)(1 - a(T))$ due to solar radiation; this term depends on a spatially varying *insolation* function α roughly measuring incident radiation at location x, and the *albedo* function[8] a which is assumed to be known;
- an outgoing radiation term $g(T)$ due to terrestrial radiation; this term coincides with the *greyness* function[9] g which is assumed to be known;
- an *history* term $\left(\frac{1}{\tau}\int_{-\tau}^{0} T(t+s)\,\mathrm{d}s\right)^3$ depending on a delay parameter $\tau > 0$.

By including the history term into the equation, the temperature values at a given time t depend on a weighted combination of past temperatures $T(t+s)$ over some range of $s < 0$. The dependence of temperature on the history term

[8] The albedo function is the ratio of reflected radiation from the surface to incident radiation upon it. On Earth, the albedo depend for instance on the vegetation cover and the presence of clouds. Here, the albedo function is simply modeled as a function of temperature like in Ghil and Childress (1987).

[9] The greyness function models the difference between black-body radiation equal to σT^4 (where σ is the Stefan-Boltzmann constant) and the radiation of the body of interest, for example the Earth.

corresponds to the delayed response of the incoming and outgoing radiation functions.

Because of the history term, the initial condition of the EBBM has to be of the form:

$$T(s,x) = T_0(s,x), \text{ for } s \in [-\tau, 0] \text{ and } x \in [0,1],$$

for some function T_0 defined on $[-\tau, 0] \times [0,1]$. Here we set $T_0 \equiv 10°\text{C}$. The boundary conditions are of Neumann's type:

$$\frac{\partial T}{\partial x}(t,0) = \frac{\partial T}{\partial x}(t,1) = 0 \text{ for } t \geq 0.$$

Figure 4.6 (top panels) shows the solution of Equation (4.9) for two values of the delay parameter, $\tau = 0.2$ ky and $\tau = 0.7$ ky, during the time interval $-\tau \leq t \leq t_{\max} = 5$ ky, where 1 ky = 1000 years. Figure 4.6 (bottom) shows the temporal evolution of the average temperature. The function α which was used for these simulations is plotted in Figure 4.8. When $\tau = 0.2$ ky, the solution of (4.9) exhibits small temporal variations that are quickly damped and a stable steady state is quickly reached. For $\tau = 0.7$, a stable periodic orbit is reached asymptotically, but the transient is considerably longer and larger amplitudes persist for quite a while.

Model of the observation process

We assume that data are obtained from ice cores collected at three sites $(x_1, x_2, x_3) = (0.5, 0.7, 0.9)$. At each location x_k $(k = 1, 2, 3)$, we denote by $t_{k1}, \ldots, t_{kI}$ the sequence of $I = 50$ decreasing epochs $(t_{\max} = 5 \geq t_{k1} > \ldots > t_{kI} \geq -\tau)$, at which the temperature $T(t_{ki}, x_k)$ is measured, based on laboratory sampling of the ice core extracted at location x_k. Let $Y_k(t_{ki})$ denote the measure of the temperature $T(t_{ki}, x_k)$.

The uncertainty in dating epochs implies that $Y_k(t)$ is actually a measure of the temperature $T(s(t), x_k)$, where s is a function deforming the time scale that can vary with k. This function is the result of errors in the chronostratigraphy, i.e. in the *age-depth plot*, of ice cores. Furthermore, the uncertainty in measuring the temperature value implies that $T(s(t), S_k)$ contains noise as well.

Our model for the observation process has to take into account these two sources of uncertainty. First, given $s(t_{ki})$, $i = 1, \ldots, I$, the observed variables $Y_k(t_{ki})$ are assumed to be conditionally drawn from independent normal distributions:

$$Y_k(t_{ki}) \mid s(t_{ki}) \underset{\text{indep.}}{\sim} \text{Normal}\left\{T(s(t_{ki}), x_k), \sigma^2\right\}, \tag{4.10}$$

where σ^2 is the variance of the noise in the temperature measurements.

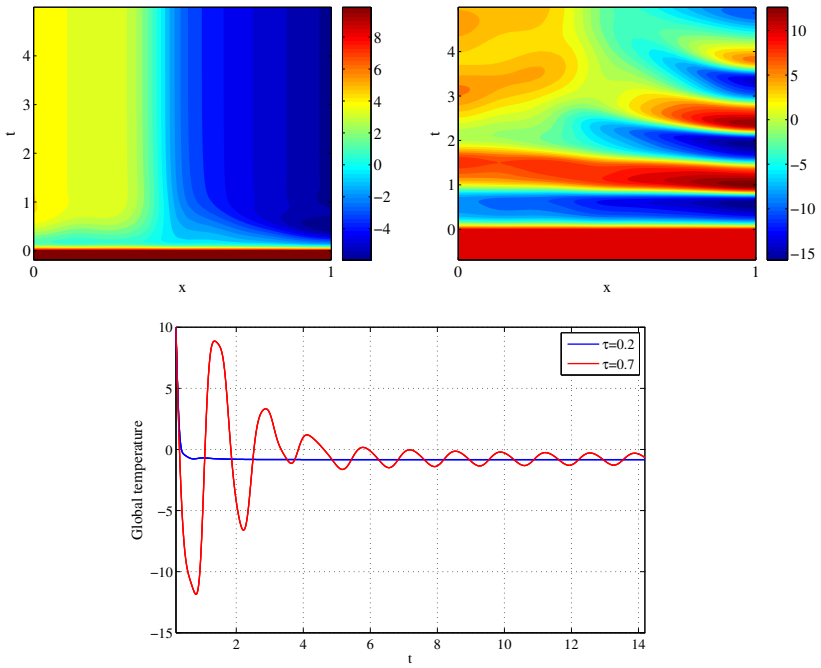


Fig. 4.6. Top: Solution $T(t,x)$ of the EBMM model (4.9) for $\tau = 0.2$ ky (left) and $\tau = 0.7$ ky (right). Bottom: Temporal variation in the global average temperature $\int_0^1 T(x,t)dx$ for $\tau = 0.2$ (blue line) and $\tau = 0.7$ (red line).

Second, the sampled times $s(t_{ki})$, $i = 1, \ldots, I$, are defined by the following random sequence:

$$s(t_{ki}) = t_{\max} - \sum_{j=1}^{i} \eta_{kj} \quad \text{with} \quad \eta_{kj} \underset{\text{indep.}}{\sim} \text{Gamma}\left(\frac{t_{k,j-1} - t_{kj}}{\kappa^2}, \kappa^2\right), \quad (4.11)$$

where κ^2 is a positive parameter that controls the shape of the distribution and $t_{k0} = t_{\max}$. Thus, the expectation of $s(t_{ki})$ in Equation (4.11) is t_{ki}, and its variance increases as t_{ki} moves further into the past. Another important feature of the model (4.11) is that it is order-preserving: if $t_{ki} > t_{ki'}$, then $s(t_{ki}) > s(t_{ki'})$. Therefore, there is no uncertainty on the order of times $t_{k1}, \ldots, t_{kI}$.

4.3.2 Estimation

Our aim was to estimate α, which is supposed to be a piecewise constant function with 30 jumps regularly located over [0,1], and the initial temperature T_0 during the time window $[-\tau, 0]$, the other parameters being fixed. We

performed this estimation using data obtained from the two dynamics shown in Figure 4.6. At each one of the locations $x_1 = 0.5$, $x_2 = 0.7$ and $x_3 = 0.9$, we randomly drew 50 epochs t_i in the interval $(-\tau, t_{\max})$ and recorded the temperatures $T(t_{ki}, x_k)$ at these epochs and locations. Using our observation model defined by Equations (4.10–4.11), we constructed noisy observations $Y_k(t_{ki})$ of these temperatures. The exact temperatures at the exact times are presented together with the measured values in Figure 4.7. Observed temperatures $Y_k(t_{ki})$ for negative times were used to determine the prior distribution of T_0; see Equation (4.12) below. Observed temperatures $Y_k(t_{ki})$ for positive times were used to make the posterior inference.

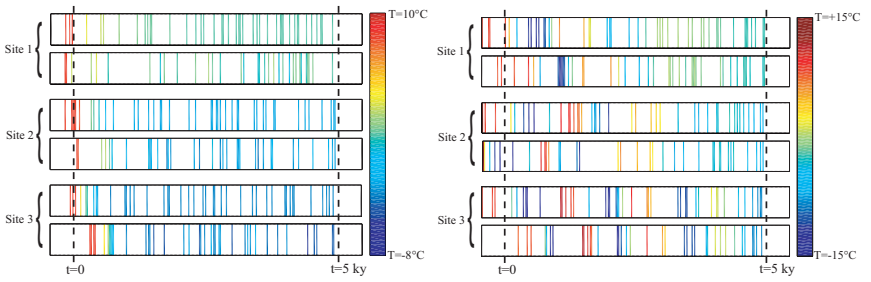


Fig. 4.7. Actual temperatures versus measured temperatures. At each site x_k, $k = 1, 2, 3$, the upper row corresponds to the actual temperatures at the actual times $s(t_{k1}), \ldots, s(t_{kI})$, while the lower row corresponds to the measured temperatures at targeted times $t_{k1}, \ldots, t_{kI}$. Left: $\tau = 0.2$ ky; Right: $\tau = 0.7$ ky.

The joint distribution of the observations $\{Y_k(t_{ki}),\ i = 1, \ldots, I,\ k = 1, 2, 3\}$, conditional on the *real* temperatures $\{T(s(t_{ki}), x_k),\ i = 1, \ldots, I,\ k = 1, 2, 3\}$ can be written as a multiple integral (due to variables η_{kj}) which was assessed using Monte Carlo integrations. Thus, the random variables η_{kj} were not explicitly inferred and a likelihood depending only on the unknowns α and T_0 can be written.

We assumed independent uniform prior distributions in a sufficiently large interval for the discretized values $\alpha_1, \ldots, \alpha_{31}$ of α:

$$\pi(\alpha_m) \underset{\text{indep.}}{\sim} \text{Uniform}(0, 1000).$$

For the sake of simplicity, we assumed that the prior distribution of T_0 was a Dirac distribution:

$$\pi(T_0) \sim \text{Dirac}(T_0^{obs}), \tag{4.12}$$

where T_0^{obs} is a constant obtained by averaging the observations $Y_k(t_{ki})$ over all negative times $t_{ki} < 0$ and $k = 1, 2, 3$. This means that T_0 is constant in space and time and takes the value T_0^{obs}.

We drawn a sample from the posterior distribution of the parameters using a MCMC algorithm with Metropolis-Hastings updates.

4.3.3 Results

Figure 4.8 shows the marginal posterior quantiles of $\alpha(x)$, when $\tau = 0.2$ ky and $\tau = 0.7$ ky. In both panels, the median of the posterior distribution is quite close to the true values of the coefficient $\alpha(x)$. Moreover, the true values do lie between the first and last deciles of the distribution, for all values of x. Remember that observations were only made at three sites x_k, and that all three sites lie in the right half of the model domain, $x_k \in [0.5, 1]$. This restriction only seems to affect somewhat the estimation error of $\alpha(x)$ in the left half of the model domain, $x \in [0, 0.5]$, for $\tau = 0.2$ ky.

Overall, the main difference between Figure 4.8 (left) and (right) is that the marginal distribution is more variable in the case $\tau = 0.2$ ky, for all x, meaning that the insolation coefficient $\alpha(x)$ is harder to estimate in this case. Such a result is somewhat surprising because the larger and faster variations in the case $\tau = 0.7$ ky lead to larger measurement errors; this can be easily seen in Figure 4.7. A possible explanation for this result is that the solution of (4.9) is much more sensitive to variations in the parameters when $\tau = 0.7$ ky.

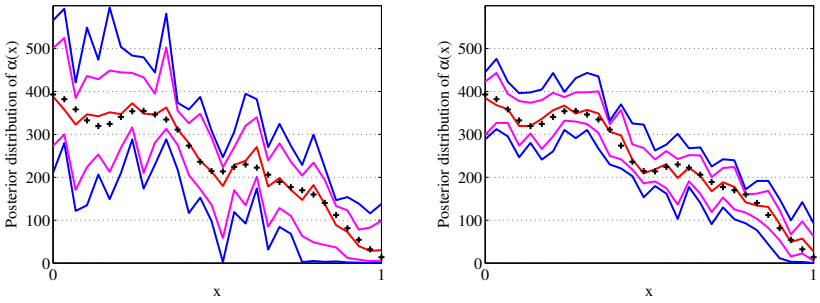


Fig. 4.8. Marginal posterior quantiles of the parameter $\alpha(x)$. The red curve is the posterior median of $\alpha(x)$, the magenta curves the first and last deciles of its distribution, and the blue curves the first and last percentiles. The true values of $\alpha(x)$ are given by the symbol $+$. Left: $\tau = 0.2$ ky; Right: $\tau = 0.7$ ky.

5

Parameter estimation without likelihood

Author's references: Soubeyrand et al. (2009a), Soubeyrand et al. (2013), Soubeyrand and Haon-Lasportes (2015).

Fitting realistic, spatio-temporal, epidemiological models to data is often a difficult task because one generally has to handle, for example, latent processes, spatial dependencies and heterogeneity in data. In the examples provided in the previous chapters, we have seen solutions for handling such complexities. Most of these solutions are based on the likelihood. However, parameter estimation can be carried out without likelihood (or without the use of the explicit form of the likelihood). These alternative approaches have recently seen a renewal with the development of approximate Bayesian computation (ABC; Marin et al., 2012).

In both the frequentist and the Bayesian frameworks, the likelihood function is one of the major components for statistical inference with a parametric model. Its use, however, has drawbacks in specific situations. First, it may be impossible to write down the likelihood in a numerically tractable form: see the cases of Boolean models (Van Lieshout and Van Zwet, 2001), Markov point processes (Møller and Waagepetersen, 2003), Markov spatial processes (Guyon, 1985) and spatial generalized linear mixed models (spatial GLMM; Diggle et al., 1998), where multiple integrals cannot be reduced due to spatial dependencies. Second, the likelihood may not be completely appropriate because of the associated assumptions. For instance, the likelihood is built under distributional assumptions, which may be tricky to specify in the case of insufficient information, as in classical geostatistics (Chilès and Delfiner, 1999); see also McCullagh and Nelder (1989, chap. 9) in regression analysis. In the same vein, all data are assumed to have the same weights in the likelihood, but the influence of outliers may be too large according to the analyst (Markatou, 2000).

The difficulties encountered with the likelihood can be circumvented with numerous frequentist and Bayesian algorithms and tools (e.g. MCEM, MCMC, pseudo-likelihood maximization, quasi-likelihood maximization, weighted likelihood maximization, generalized least squares estimation, method of moments, approximate Bayesian computation (ABC)). If my general practice is to adapt these approaches to the models and data I deal with, I have also

been occasionally interested in their methodological development. Thus, this chapter presents three topics concerning parameter estimation without likelihood. Section 5.1 presents the consequences of replacing the likelihood in the Bayesian formula of the posterior distribution by a function of a contrast. Section 5.2 presents an algorithm for optimizing the distance between functional summary statistics in ABC. Finally, Section 5.3 presents a study of the weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and its implications in ABC.

5.1 Contrast-based posterior distribution

A contrast is a function of the model parameters and the observed data which is minimized to estimate the parameters (Dacunha-Castelle and Duflo, 1982). The minimum contrast approach is a generic estimation method, which was developed in the frequentist perspective. The maximum likelihood estimation as well as the maximum pseudo-, weighted- or quasi-likelihood estimation, the diverse least squares methods, the method of moments and the M-estimation can be formulated as minimum contrast estimation problems.

In Soubeyrand et al. (2009a), we replaced the likelihood appearing in the Bayesian formula of the posterior distribution by a function of a contrast. This procedure provides a contrast-based (CB) posterior distribution that does not coincide, in the general case, with the classical posterior distribution. Thus, we investigated what are the posterior distribution and the MAP (maximum a posteriori) estimator based on a contrast.

Under mild conditions on the prior distribution, we show that the CB–MAP estimator inherits the asymptotic properties (consistency and asymptotic normality) of the minimum contrast estimator, as the classical MAP estimator inherits the asymptotic properties of the maximum likelihood estimator (Caillot and Martin, 1972). The limit variance matrix of the normalized estimator is $I_\theta^{-1}\Gamma_\theta I_\theta^{-1}$ where Γ_θ is the limit variance of the gradient of the contrast and I_θ is the limit Hessian matrix of the contrast.

Moreover, we show that the CB–posterior distribution is asymptotically equivalent to a normal distribution whose variance matrix is I_θ^{-1}. Therefore, when building the contrast, particular attention must be paid to satisfy, if possible, $I_\theta^{-1}\Gamma_\theta I_\theta^{-1} = I_\theta^{-1}$. Indeed, with such a contrast, inference can be made without computing matrices Γ_θ and I_θ: the posterior distribution can either be used as a limit distribution from a frequentist viewpoint or be used to make inference in the Bayesian way. When building a contrast satisfying $I_\theta^{-1}\Gamma_\theta I_\theta^{-1} = I_\theta^{-1}$ is not possible, the CB–posterior distribution can nevertheless be used to estimate I_θ^{-1}. Thus, the computation of the limit Hessian matrix of the contrast is avoided.

5.1.1 Incorporating a contrast in the Bayesian formula

Consider a set of parametric models described by the corresponding set of distributions $\{P_\alpha : \alpha \in \Theta\}$. Consider samples of increasing sizes $t \in T \subset \mathbb{N}$, drawn from the distribution P_θ with the true parameter θ. A contrast for θ is a random function $\alpha \mapsto U_t(\alpha)$ defined over Θ, depending on a sample of size t, and such that $\{U_t(\alpha) : t \in T\}$ converges in probability, as $t \to \infty$, to a function $\alpha \mapsto K(\alpha, \theta)$ which has a strict minimum at $\alpha = \theta$. The minimum contrast estimator is:

$$\hat{\theta}_t = \text{argmin}\{U_t(\alpha), \alpha \in \Theta\}.$$

Let $\mathcal{X}_t = \{X_i : i \leq t\}$ be a sample of size t drawn from a distribution in $\{P_\alpha : \alpha \in \Theta\}$. Then, the posterior distribution of α is:

$$p(\alpha \mid \mathcal{X}_t) = \frac{P_\alpha(\mathcal{X}_t)\pi(\alpha)}{\int_\Theta P_\beta(\mathcal{X}_t)\pi(\beta)d\beta},$$

where $P_\alpha(\mathcal{X}_t)$ denotes the likelihood and $\pi(\cdot)$ is a prior distribution defined over Θ. The contrast corresponding to the likelihood being $U_t^{lik}(\alpha) = -\frac{1}{t}\log P_\alpha(\mathcal{X}_t)$ (Dacunha-Castelle and Duflo, 1982), the posterior distribution can be written by replacing $P_\alpha(\mathcal{X}_t)$ by $\exp(-tU_t^{lik}(\alpha))$ in the previous equation.

Here, we propose to substitute the contrast associated with the likelihood in the Bayesian formula with any contrast $U_t(\alpha)$. This leads to a contrast-based (CB) posterior distribution denoted by $p_t(\alpha)$:

$$p_t(\alpha) = \frac{\exp(-tU_t(\alpha))\pi(\alpha)}{\int_\Theta \exp(-tU_t(\beta))\pi(\beta)d\beta}. \tag{5.1}$$

The CB–MAP estimator obtained by maximizing $p_t(\cdot)$ is denoted by:

$$\tilde{\theta}_t = \text{argmax}\{p_t(\alpha), \alpha \in \Theta\}.$$

$\tilde{\theta}_t$ is at the minimum of $\alpha \mapsto U_t(\alpha) - (1/t)\log\pi(\alpha)$ and, in general, does not coincide with the classical minimum contrast estimator $\hat{\theta}_t = \text{argmin}\{U_t(\alpha), \alpha \in \Theta\}$.

Below, we briefly present the asymptotic properties of the CB–MAP estimator and the CB–posterior distribution.

5.1.2 Consistency and asymptotic normality of the CB–MAP estimator

We noted above that the CB–MAP estimator $\tilde{\theta}_t$ is at the minimum of $\alpha \mapsto U_t(\alpha) - (1/t)\log\pi(\alpha)$. This function satisfies the definition of a contrast. Consequently, convergence properties of $\tilde{\theta}_t$ can be easily obtained by

using the contrast theory (Dacunha-Castelle and Duflo, 1982). Assume that the hypotheses required for the convergence of the classical minimum contrast estimator are satisfied. Let us assume in addition that the prior distribution $\pi(\cdot)$ is proper, differentiable and strictly positive over Θ. It follows that, as $t \to \infty$,

- $\tilde{\theta}_t$ converges in probability to θ and
- $\sqrt{t}(\tilde{\theta}_t - \theta)$ converges in law to the Gaussian distribution $\mathcal{N}\left(0, I_\theta^{-1} \Gamma_\theta I_\theta^{-1}\right)$,

where I_θ and Γ_θ are matrices satisfying:

$$\begin{aligned} \mathbf{H}U_t(\theta) &\to I_\theta \quad \text{in probability as } t \to \infty \\ \sqrt{t}\mathbf{grad}U_t(\theta) &\to \mathcal{N}(0, \Gamma_\theta) \quad \text{in law,} \end{aligned}$$

where $\mathbf{H}$ and $\mathbf{grad}$ are the Hessian and gradient operators, respectively.

The convergence results given above can also be obtained by noting that the asymptotic deviation between the classical minimum contrast estimator $\hat{\theta}_t$ and the CB–MAP estimator $\tilde{\theta}_t$ is of order $1/t$. More exactly, we have shown that:

$$\tilde{\theta}_t - \hat{\theta}_t = \frac{1 + o_{\text{proba}}(1)}{t\pi(\theta)} I_\theta^{-1} \mathbf{grad}\pi(\theta). \tag{5.2}$$

5.1.3 Convergence of the CB–posterior distribution

Under the assumption made above, the CB–posterior distribution $p_t(\cdot)$ is asymptotically equivalent to the density function of the Gaussian distribution $\mathcal{N}\left(\tilde{\theta}_t, (tI_\theta)^{-1}\right)$:

$$p_t(\alpha) \underset{t\to\infty}{\sim} \frac{1}{(2\pi)^{p/2}|(tI_\theta)^{-1}|^{1/2}} \exp\left(-\frac{1}{2}(\alpha - \tilde{\theta}_t)'(tI_\theta)(\alpha - \tilde{\theta}_t)\right). \tag{5.3}$$

This result allows us to figure out what the CB–posterior distribution is and how it can be used to make inference in the frequentist and Bayesian ways.

In the contrast theory, the distribution $\mathcal{N}\left(\tilde{\theta}_t, (tI_\theta)^{-1}\Gamma_\theta I_\theta^{-1}\right)$ is used to make frequentist inference about θ: the point estimator is $\tilde{\theta}_t$, and confidence zones are provided based on this normal distribution. Consequently, if the contrast is such that $I_\theta^{-1}\Gamma_\theta I_\theta^{-1} = I_\theta^{-1}$, then the CB–posterior distribution $p_t(\cdot)$, which approximates the density of $\mathcal{N}\left(\tilde{\theta}_t, (tI_\theta)^{-1}\right)$, can be directly used to make frequentist inference about θ: the mode of $p_t(\cdot)$ is the point estimator, and confidence zones can be directly determined from $p_t(\cdot)$. This case is particularly interesting since the calculation of the limit matrices $I_\theta = \lim_{t\to\infty} \mathbf{H}U_t(\theta)$ and $\Gamma_\theta = \lim_{t\to\infty} V_\theta(\sqrt{t}\mathbf{grad}U_t(\theta))$ is no more required.

Moreover, when the contrast satisfies $I_\theta^{-1}\Gamma_\theta I_\theta^{-1} = I_\theta^{-1}$, then the CB–posterior distribution $p_t(\cdot)$ can be used to make inference in the Bayesian way, i.e. to use $p_t(\cdot)$ as a real posterior density. The motivation is based on

the following analogy: when the contrast corresponding to the likelihood is employed (in this case, $I_\theta^{-1}\Gamma_\theta I_\theta^{-1} = I_\theta^{-1}$), then $p_t(\cdot)$ can be used (i) to make frequentist inference since it is an approximation of the limit distribution of the estimator (see above) and (ii) to make Bayesian inference since it is the classical posterior density. It has to be noted that, in general, the CB–posterior density $p_t(\cdot)$ does not coincide with the classical posterior density. It is a posterior density based on the information brought by the contrast under consideration.

If the contrast does not satisfy $I_\theta^{-1}\Gamma_\theta I_\theta^{-1} = I_\theta^{-1}$, then the CB–posterior distribution $p_t(\cdot)$ cannot be used to approximate the limit distribution of $\tilde{\theta}_t$ or to make Bayesian inference. However, $p_t(\cdot)$ can be used to estimate the matrix I_θ, so avoiding the calculation of the second derivatives of the contrast. Indeed, one can see from (5.3) that an estimate of I_θ is the matrix Ω^{-1}/t where Ω is the variance matrix of the normal density function centered around $\tilde{\theta}_t$ and fitted to $p_t(\cdot)$ (using a least square technique for example). If θ is real, I_θ can be more simply estimated by $2\pi p_t(\tilde{\theta}_t)^2/t$ since Equation (5.3) yields $p_t(\tilde{\theta}_t) \underset{t\to\infty}{\sim} (tI_\theta/2\pi)^{1/2}$. We have not found an equivalent way to easily estimate Γ_θ. Thus, this matrix must be assessed with analytical calculation of the second derivatives or with simulations.

5.1.4 Application to a Markovian spatial model

The simulation study presented here illustrates the application of the method for estimating a bivariate parameter of a spatial model. Here, the CB–posterior distribution is different from the limit distribution of the estimator; it cannot be directly used to make inference but can be used for estimating I_θ.

We built a data set by simulating a spatial Markov field X with two states, 0 and 1. The model is defined by the conditional probability of X_i given X_j, $j \in V(i)$ ($V(i)$ is the set of the four nearest neighbors of i) satisfying (Guyon, 1985):

$$\begin{aligned} P_\theta(X_i \mid X_j, j \neq i) =& P_\theta(X_i \mid X_j, j \in V(i)) \\ =& \frac{\exp\left(\theta_1 X_i + \theta_2 \sum_{j\in V(i)} X_i X_j\right)}{\left\{1 + \exp\left(\theta_1 + \theta_2 \sum_{j\in V(i)} X_j\right)\right\}}. \end{aligned}$$

The field was simulated on a $n \times n$ square grid $\mathcal{I}$ (here, $t = n^2 = 20^2$); see Figure 5.1 (left).

The classical likelihood cannot be analytically calculated for this model. Therefore, a pseudo-likelihood was proposed to make inference (Guyon, 1985). The pseudo-likelihood is the product of the conditional probabilities $\prod_{i\in\mathcal{I}} P_\theta(X_i \mid X_j, j \neq i)$. To estimate θ_1 and θ_2, we applied the estimation method proposed above by using a uniform prior density over $[-1.5, 1.5]^2$ and the contrast corresponding to the pseudo-likelihood:

$$U_{n^2}(\alpha) = -\frac{1}{n^2}\sum_{i\in\mathcal{I}} \log P_\alpha(X_i \mid X_j, j \in V(i)). \tag{5.4}$$

The CB–posterior density is shown in Figure 5.1 (center). The MAP estimate is $\tilde{\theta}_t = (-0.21, 0.38)$.

To give the limit distribution $\mathcal{N}(\tilde{\theta}_t, I_\theta^{-1}\Gamma_\theta I_\theta^{-1}/n^2)$ of the estimator, matrices Γ_θ and I_θ must be estimated. We computed the gradient and the Hessian of the contrast for $N = 1000$ Markov fields simulated under $\tilde{\theta}_t$, and we used the sample variance of the gradients for estimating Γ_θ and the sample mean of the Hessians for estimating I_θ. Thus, the estimate of the limit variance matrix $I_\theta^{-1}\Gamma_\theta I_\theta^{-1}/n^2$ was:

$$\begin{pmatrix} 0.14 & -0.055 \\ -0.055 & 0.022 \end{pmatrix}.$$

Figure 5.1 (right) shows the limit density function of the estimator together with the 95%-confidence zone. We can see that the true parameter belongs to this zone. Moreover, Figure 5.1 shows that the limit density is quite close to the posterior density. The pseudo-likelihood, which takes into account short-distance interactions, brings in this case almost the same information than the likelihood. It has however to be noted that this would not be the case if long-distance interactions had been introduced in the spatial Markov model.

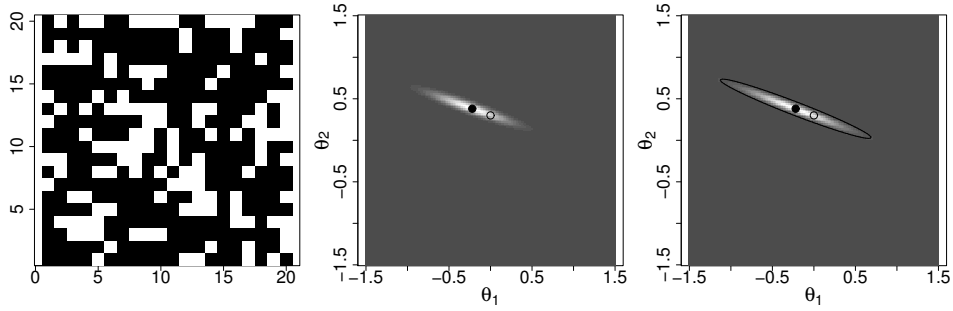


Fig. 5.1. Left: realization of a Markovian spatial process with two states over a 20×20 grid. Center: contrast-based posterior density. Right: limit density $\mathcal{N}(\tilde{\theta}_t, I_\theta^{-1}\Gamma_\theta I_\theta^{-1}/n^2)$. On the center and right panels, the MAP estimate and the true parameter are drawn with a black dot and a circle, respectively. On the right panel, the continuous line circumscribes the 95%-confidence zone.

5.2 Approximate Bayesian computation with functional statistics

In spatial statistics, one often relies on functional statistics to characterize spatial patterns[1]. Similarly, in population genetics, spatial genetic structures are often analyzed with functional statistics measuring the genetic differentiation with respect to the geographic distance[2].

Such functional statistics also enable the estimation of parameters of spatially explicit (and genetic) models. Recently, Approximate Bayesian Computation (ABC) has been proposed to estimate model parameters from functional statistics. However, applying ABC with functional statistics may be cumbersome because of the high dimension of the set of statistics and the dependencies among them. To tackle this difficulty, we proposed in Soubeyrand et al. (2013) an ABC procedure relying on an optimized weighted distance between observed and simulated functional statistics. We applied this procedure to a dispersal model characterized by a functional statistic linking genetic differentiation to geographic distance.

5.2.1 Background: the ABC–rejection procedure

Consider observed data $\mathcal{D} \in \mathbb{D}$ which are assumed to be generated under the stochastic model $\mathcal{M}_\theta$ parametrized by $\theta \in \Theta$ with prior density π. The data space $\mathbb{D}$ and the parameter space Θ are both included in multidimensional sets of real vectors.

The posterior distribution $p(\theta \mid \mathcal{D})$ can be estimated using the following ABC–rejection algorithm (Rubin, 1984):

A1. Carry out the next two steps, independently for i in $\{1, \ldots, I\}$,
1. Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$.
2. Accept θ_i if $\mathcal{D}_i = \mathcal{D}$, reject it otherwise.

The set of accepted θ_i forms a sample from the posterior distribution

$$p(\theta \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \theta)\pi(\theta)}{\int_\Theta f(\mathcal{D} \mid \alpha)\pi(\alpha)d\alpha},$$

where $f(\mathcal{D} \mid \theta)$ is the conditional probability distribution function of $\mathcal{D}$ given θ, i.e. the (intractable or unknown) likelihood of the model $\mathcal{M}_\theta$.

Algorithm **A1** is rarely usable because the probability of generating $\mathcal{D}_i$ equal to $\mathcal{D}$ is very low when the dimensionality of the data space $\mathbb{D}$ is large and this probability is even zero for continuous data. To circumvent this difficulty,

[1] E.g. the Ripley's K-function, the empty space function J, the nearest-neighbour distance distribution function G (Illian et al., 2008).

[2] E.g. the F_{ST} function (Rousset, 1997) and the Φ_{FT} function (Austerlitz and Smouse, 2002).

two ideas have been applied: the introduction of a tolerance threshold[3] and the replacement of the raw data by summary statistics (Pritchard et al., 1999). This leads to the following ABC–rejection algorithm:

A2. Carry out the next three steps, independently for i in $\{1, \ldots, I\}$,

1. Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$.
2. Compute the statistic $S_i = s(\mathcal{D}_i)$, where s is a function from $\mathbb{D}$ to the space $\mathbb{S}$ of statistics.
3. Accept θ_i if $d(S_i, S) \leq \epsilon(\tau)$, where d is a distance over $\mathbb{S}$ and $\epsilon(\tau) \in \mathbb{R}_+$ is a tolerance threshold for the distance between the observed statistic $S = s(\mathcal{D})$ and the simulated ones. The threshold $\epsilon(\tau)$ depends on the proportion τ of accepted θ_i among the I simulated parameters; $\epsilon(\tau)$ is the empirical quantile of order τ. Thereafter, τ is called the acceptance rate.

The set of accepted parameters, say $\Theta_{\tau,I} = \{\theta_i : d(S_i, S) \leq \epsilon(\tau), i = 1, \ldots, I\}$, forms a sample from the posterior distribution

$$p_{\epsilon(\tau)}(\theta \mid S) = \frac{\left(\int_{B(S,\epsilon(\tau))} \tilde{f}(z \mid \theta) dz\right) \pi(\theta)}{\int_\Theta \left(\int_{B(S,\epsilon(\tau))} \tilde{f}(z \mid \alpha) dz\right) \pi(\alpha) d\alpha}, \tag{5.5}$$

where $\tilde{f}(S \mid \theta)$ is the conditional probability distribution function of S given θ and $B(S, \epsilon(\tau))$ is the ball with center S and radius $\epsilon(\tau)$ in the space $\mathbb{S}$ with distance d.

When $\epsilon(\tau)$ tends to zero, $p_{\epsilon(\tau)}(\theta \mid S)$ may be a *good* approximation of the posterior distribution conditional on the statistic[4], i.e.

$$p(\theta \mid S) = \frac{\tilde{f}(S \mid \theta)\pi(\theta)}{\int_\Theta \tilde{f}(S \mid \alpha)\pi(\alpha) d\alpha}, \tag{5.6}$$

and the sample $\Theta_{\tau,I}$ of accepted parameters is approximately distributed under this posterior distribution. If, in addition, the statistics are sufficient, then $\tilde{f}(S \mid \theta) = f(\mathcal{D} \mid \theta)$ and $\Theta_{\tau,I}$ is approximately a sample from the classical posterior distribution $p(\theta \mid \mathcal{D})$ conditional on the data.

5.2.2 Selecting a weight function for functional statistics

Suppose now that S is a functional statistic in the space $\mathbb{S}$ of statistics, which is included in the space of real-valued and square-integrable functions defined over $\mathbb{R}$:

[3] We can guess that Rubin (1984) already suggested the application of a tolerance threshold since he used the inaccurate expressions "look just like" and "match" to compare the simulated and observed data.

[4] This can be shown under regularity assumptions about the conditional probability distribution function $S \mapsto \tilde{f}(S \mid \theta)$ of S given θ. However, such assumptions cannot be checked in usual applications of ABC where $\tilde{f}$ is generally analytically intractable.

$$\mathbb{S} \subset \left\{ g : \mathbb{R} \to \mathbb{R}, \int_{\mathbb{R}} g^2 < \infty \right\}.$$

Besides, we assume that the distance $d : \mathbb{S}^2 \to \mathbb{R}^+$, used in algorithm **A2** to compare observed and simulated statistics, is parametrized by a non-negative weight function $w : \mathbb{R} \to \mathbb{R}_+$ and satisfies:

$$d(S_i, S; w) = \int_{\mathbb{R}} w(r)\{S_i(r) - S(r)\}^2 dr. \tag{5.7}$$

The weight function is expected to modulate the squared difference between $S_i(r)$ and $S(r)$ with respect to the information about the parameters brought by the statistics at r. In the applications that we tackled, w is in sets of positive piecewise constant functions with finite number of jumps, with known jump locations and with integral over $\mathbb{R}$ equal to one. Thus, in the optimization of w, which is proposed below, we have to select a finite number of jump levels.

The weight function that we proposed is the optimized function w_{opt} obtained by minimizing a mean square error (MSE) of a point estimate of θ (Rohatgi, 2003, chap. 4). The MSE that we used is a Bayesian MSE (BMSE): the square error is integrated over Θ with respect to the prior distribution π. This approach, detailed below in algorithm **A3**, is analogous to minimizing the mean square error of prediction where θ is the random variable to be predicted (McCulloch and Searle, 2001, chap. 9).

The optimized weight function w_{opt} as well as an optimized acceptance rate τ_{opt} are determined within the following ABC–rejection algorithm:

A3. Carry out the next four steps,

1. For i in $\{1, \ldots, I\}$, independently generate θ_i from π, simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$ and compute the functional statistic $S_i = s(\mathcal{D}_i)$;
2. For j in $\{1, \ldots, J\}$, independently generate θ'_j from π, simulate $\mathcal{D}'_j$ from $\mathcal{M}_{\theta'_j}$ and compute the functional statistic $S'_j = s(\mathcal{D}'_j)$;
   (θ'_j, S'_j), $j = 1, \ldots, J$, *will be used as pseudo-observed data sets (PODS) for optimizing the weights and the acceptance rate;*
3. Select the weight function and the acceptance rate which minimize the following BMSE criterion:

$$\mathrm{BMSE}_J(w, \tau) = \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(\hat{\theta}'_{jk}(w, \tau) - \theta'_{jk})^2}{V(\theta'_{jk})}. \tag{5.8}$$

In (5.8), θ'_{jk}, $k = 1, \ldots, K$, are the K components of θ'_j ($\Theta \subset \mathbb{R}^K$, $K \geq 1$); $V(\theta'_{jk})$ is the prior variance of θ'_{jk} depending only on π and allows the scaling of the parameter components; the point estimates $\hat{\theta}'_{jk}(w, \tau)$ are the marginal posterior medians of θ'_{jk}:

$$\hat{\theta}'_{jk}(w, \tau) = \mathrm{Median}\{\theta_{ik} : d(S_i, S'_j; w) \leq \epsilon(\tau), i = 1, \ldots, I\},$$

obtained by applying the last step of algorithm **A2** with S'_j for the observed statistic, S_i for the simulated statistics, $d(\cdot, \cdot; w)$ for the distance and τ for the acceptance rate. The BMSE is minimized over the space function $\mathbb{W} = \{w : \mathbb{R} \to \mathbb{R}_+, \int_{\mathbb{R}} w = 1\}$ and the interval $(0, 1]$:

$$(w_{opt}, \tau_{opt}) = \text{argmin}_{w,\tau \in \mathbb{W} \times (0,1]} \text{BMSE}_J(w, \tau). \tag{5.9}$$

4. For i in $\{1, \ldots, I\}$, accept θ_i if $d(S_i, S; w_{opt}) \leq \epsilon(\tau_{opt})$.

The set of accepted parameters $\Theta_{opt} = \{\theta_i : d(S_i, S; w_{opt}) \leq \epsilon(\tau_{opt}), i = 1, \ldots, I\}$ forms a sample from the posterior distribution (5.5) with $\epsilon(\tau) = \epsilon(\tau_{opt})$ and with $B(S, \epsilon(\tau))$ equal to the ball with center S and radius $\epsilon(\tau_{opt})$ in the space $\mathbb{S}$ with distance $d(\cdot, \cdot; w_{opt})$. Thus, weighting the distance modifies the posterior under which the accepted parameters are drawn. However, when $\epsilon(\tau_{opt})$ tends to zero, the new posterior distribution (like the one given in Equation (5.5)) may be a good approximation of $p(\theta \mid S)$ given in Equation (5.6).

Note that the BMSE in Equation (5.8) is the Monte-Carlo approximation of the exact BMSE equal to $\sum_{k=1}^{K} E\{(\hat{\theta}'_{jk}(w, \tau) - \theta'_{jk})^2\}/V(\theta'_{jk})$. Besides, other criteria than the BMSE may be used to select w and τ, e.g. mean square errors or mean absolute errors based on the posterior mode, the posterior mean or posterior quantiles.

In applications, w is a positive piecewise constant functions with a finite number of jumps, with known jump locations and with integral over $\mathbb{R}$ equal to one. Therefore, the optimization program (5.9) consists in minimizing the BMSE with respect to a finite number of jump levels and the acceptance rate τ. This optimization was carried with the Nelder-Mead algorithm (Nelder and Mead, 1965) modified to take into account the linear constraints over the jump levels and τ.

5.2.3 Using a pilot ABC run

After a first (pilot) run of algorithm **A3**, which yields a pilot posterior sample, namely Θ_{pilot}, one may proceed to a second selection of the weight function and the acceptance rate by restricting the computation of the MSE to simulations close to Θ_{pilot}. This approach, detailed in Appendix, is based on the minimization of a partial mean square error (PMSE), which depends on Θ_{pilot}, and can be implemented without supplementary simulation. In what follows, the algorithm including the pilot ABC run is denoted by **A4**.

5.2.4 Application to a dispersal model

Algorithm **A3** and **A4** were applied to simulated data (namely, a simple step-model and a modified Thomas process). The performance of the optimized weight function was compared with the performance of the constant weight

function and the weight function obtained by equalizing the variances of the statistics, which are common weight functions in ABC. We also compared our approach with algorithm **A2** including a prior transformation of summary statistics via the PLS method (Wegmann et al., 2009), the minimum entropy method and the two-stage method (Nunes and Balding, 2010). Essentially, these approaches reduce the dimension of the summary statistics to avoid the negative impact of uninformative or correlated statistics on inference accuracy. Overall, Algorithms **A3** and **A4** were better than the other approaches for most of the tested performance criteria, and **A4** was better than **A3**. This advantage was certainly obtained because **A3** and **A4** take into account dependencies between values of the functional statistic along the support of the function, and are based on a quantitative weighting of statistics.

ABC is known to be a computer intensive approach. In this regard, we showed that, with the optimized weights (without pilot study), we can run ten times fewer simulations and reach an estimation accuracy equal to the one obtained with the constant weights. This result is particularly useful when simulations are very time consuming[5]. The use of a pilot ABC allows to reduce the number of simulations required to reach a given level of inference accuracy, however, the risk with such an iterative procedure is that the pilot study results in an overly narrow region in the space of parameters.

We applied our approach to fit a dispersal model of pollen. In this application, raw data consisted of genetic information from molecular markers collected from a set of wild-service trees (*Sorbus torminalis*) sampled in a delimited area. The functional statistics that we used in the implementation of ABC was the observed Φ_{FT} function, which provides a measure of pairwise genetic differentiation with respect to pairwise geographic distance. Details of the model and data are not described here[6]. Figure 5.2 shows the observed functional statistics, the optimized weight function and the posterior sample obtained from Algorithm **A4** for the dispersal parameters. In this application, the gain in using the optimized weight function exists but is moderate, because the functional statistics is particularly noisy (see Figure 5.2, top). For more regular functional statistics[7], for which dependencies along the support of the function are higher, the gain can be larger.

[5] In our applications, for 10^5 simulations, the optimizations took a few hours or less with a desktop computer

[6] See Soubeyrand et al. (2013) for detailed information about the model and data.

[7] An example of more regular functional statistics is the pair correlation function used in Soubeyrand et al. (2013) for inferring parameters of the modified Thomas process.

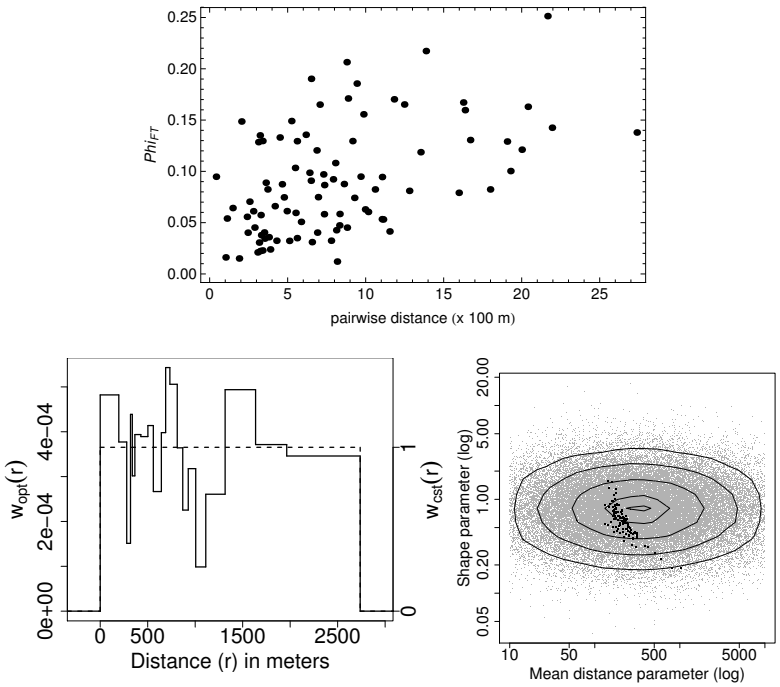


Fig. 5.2. Top: Observed Φ_{FT} function, which provides a measure of pairwise genetic differentiation with respect to pairwise geographic distance; this functional statistic is used in Algorithm **A4** to estimate the parameters of a pollen dispersal model. Bottom left: optimized weight function (solid line) to be compared with the constant weight function (dashed line). Bottom right: joint prior distribution (contour lines and grey dots) and joint posterior distribution (black dots) of the model parameters, namely the mean distance parameter and the shape parameter of a power-exponential dispersal kernel.

5.3 A Bernstein-von Mises theorem for Approximate Bayesian computation

In ABC, the choice of summary statistics and the distance between the summary statistics is crucial in regards to inference accuracy. The section above deals with the choice of the distance between summary statistics. In this new section, we are interested in the choice of summary statistics themselves. More specifically, we are interested in the case where (some of) the summary statistics are point estimates of parameters (PEP), as in Drovandi et al. (2011), Fearnhead and Prangle (2012), Gleim and Pigorsch (2013) and Mengersen et al. (2013).

In the classical Bayesian framework, posteriors conditional on PEP can be viewed as specific cases of posteriors conditional on partial information (Doksum and Lo, 1990; Soubeyrand et al., 2009a). In Soubeyrand and Haon-

Lasportes (2015), we provide new results of weak convergence when PEP are either maximum likelihood estimates (MLE) or pseudo-maximum likelihood estimates (MPLE). The case where PEP are MPLE is of specific interest because, in ABC, it may be possible to compute MPLE via simplifications of the dependence structure in the model and to use MPLE as summary statistics.

The results of weak convergence that we provided can be viewed as new extensions of the Bernstein – von Mises (BvM) theorem. For parametric models, from which independent observations are made, the BvM theorem (i) states conditions under which the posterior distribution is asymptotically normal and (ii) subsequently leads to the efficiency of Bayesian point estimators and to the convergence of Bayesian confidence sets towards frequentist limit confidence sets (Walker, 1969; Freedman, 1999). Thus, the BvM theorem can be viewed as a frequentist justification of posterior distributions for the estimation of parameters. Numerous extensions of the BvM theorem have been proposed, for instance, when the model is semiparametric or nonparametric (Bickel and Kleijn, 2012; Bontemps, 2011; Castillo and Nickl, 2013; Rivoirard and Rousseau, 2012), when observations are dependent (Borwanker et al., 1971; Tamaki, 2008), when the model is misspecified (Kleijn and van der Vaart, 2012) and when the model is nonregular (Bochkina and Green, 2014).

In Soubeyrand and Haon-Lasportes (2015), we extended the BvM theorem (i) when raw observations are replaced by the MLE or an MPLE and (ii) when the posterior conditional on an MPLE is approximated via ABC. The BvM extensions obtained in the classical Bayesian framework (Point (i)) are stepping stones that lead to the BvM extension obtained in the ABC framework (Point (ii)). Advancing theory in ABC has generally no direct practical implications because assumptions that may be required to prove theorems cannot be checked for a real-life implicit stochastic model whose distribution theory is intractable. However, showing an analytic result for a large class of theoretically tractable models may lead to conjecture that the result holds for some stochastic implicit models. Specifically, the work presented below allows us to conjecture that (i) an ABC–posterior distribution conditional on an MPLE is asymptotically normal and centered around the MPLE and (ii) resulting point estimates and confidence sets converge towards their frequentist analogues.

5.3.1 Notation

We use notation introduced in Section 5.2 and simply remind here that observed data are denoted by $\mathcal{D} \in \mathbb{D}$. Let $p(\mathcal{D} \mid \theta)$ denote the likelihood of the model and $p(\theta \mid \mathcal{D}) = p(\mathcal{D} \mid \theta)\pi(\theta)/p(\mathcal{D})$ the full sample posterior of the parameter vector $\theta \in \Theta$. The vector $\hat{\theta}_{ML} \in \Theta$ is the maximum likelihood estimate (MLE) of θ: $\hat{\theta}_{ML} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\mathcal{D} \mid \theta)$. The posterior of parameters conditional on the MLE is $p(\theta \mid \hat{\theta}_{ML}) = p(\hat{\theta}_{ML} \mid \theta)\pi(\theta)/p(\hat{\theta}_{ML})$, where $p(\hat{\theta}_{ML} \mid \theta)$ is the p.d.f. of the MLE given θ. Besides, we are interested in mod-

els whose likelihoods are not tractable because of the dependence structure in the data, but for which we can build tractable pseudo-likelihoods, say $\tilde{p}(\mathcal{D} \mid \theta)$. A pseudo-likelihood is generally built by ignoring some of the dependencies in the data (Gaetan and Guyon, 2008; Gourieroux et al., 1983). Let the vector $\hat{\theta}_{MPL} \in \Theta$ denotes the maximum pseudo-likelihood estimate (MPLE) of θ: $\hat{\theta}_{MPL} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \tilde{p}(\mathcal{D} \mid \theta)$. The posterior of parameters conditional on the MPLE is $p(\theta \mid \hat{\theta}_{MPL}) = p(\hat{\theta}_{MPL} \mid \theta)\pi(\theta)/p(\hat{\theta}_{MPL})$, where $p(\hat{\theta}_{MPL} \mid \theta)$ is the p.d.f. of the MPLE given θ.

5.3.2 Posterior conditional on the MLE

The full sample posterior $p(\theta \mid \mathcal{D})$ and the posterior conditional on the MLE $p(\theta \mid \hat{\theta}_{ML})$ exactly coincide in specific cases (e.g. when the MLE are sufficient statistics), but do not coincide in general. In this section, we provide an asymptotically equivalent distribution for $p(\theta \mid \hat{\theta}_{ML})$.

Following Walker (1969) and Lindley (1965, p. 130), we consider a set $\mathcal{D} = (D_1, \ldots, D_n)$ of n i.i.d. variables drawn from a parametric distribution with density $f(\cdot \mid \theta)$ with respect to a σ-finite measure on the real line, where θ is in $\Theta \subset \mathbb{R}^q$. Under this setting and additional regularity conditions, the BvM theorem establishes the asymptotic normality of the full sample posterior (Walker, 1969, Theorem 2 and conclusion): the full sample posterior density of θ is, for large n, equivalent to the normal density with mean vector equal to the MLE $\hat{\theta}_{ML}$ and covariance matrix equal to $\Omega_n(\hat{\theta}_{ML})^{-1}$:

$$p(\theta \mid \mathcal{D}) \underset{n\to\infty}{\sim} \phi_{\hat{\theta}_{ML}, \Omega_n(\hat{\theta}_{ML})^{-1}}(\theta),$$

where $\phi_{\mu,\Sigma}$ denotes the density of the normal distribution with mean vector μ and covariance matrix Σ, and $\Omega_n(\alpha)$ is the $q \times q$ matrix with element (i, j) equal to $\left(-\partial^2 \log p(\mathcal{D} \mid \theta)/\partial\theta_i \partial\theta_j\right)_{\theta=\alpha}$.

To provide an asymptotically equivalent distribution for $p(\theta \mid \hat{\theta}_{ML})$ as in BvM theorems, we assume in Lemma 5.1 (see below) that the MLE is asymptotically normal and consistent. For example, consider the same statistical model than above and assume that assumptions made in Lehmann and Casella (1998, Theorem 5.1 of the MLE asymptotic normality, p. 463) are satisfied. In particular, assume that data were generated with parameter vector θ. Then, the density of $\hat{\theta}_{ML}$ is, for large n and given θ, equivalent to the normal density with mean vector equal to the true parameter vector θ and covariance matrix equal to $n^{-1} I(\theta)^{-1}$:

$$p(\hat{\theta}_{ML} \mid \theta) \underset{n\to\infty}{\sim} \phi_{\theta, n^{-1} I(\theta)^{-1}}(\hat{\theta}_{ML}). \tag{5.10}$$

where $I(\theta)$ denotes the $q \times q$ Fisher information matrix.

Lemma 5.1 (Asymptotic normality of the posterior conditional on the MLE). *Suppose that the MLE satisfies Equation* (5.10) *with non-singular*

matrix $I(\theta)$. Under regularity assumptions[8]*, when $n \to \infty$, the posterior density $p(\theta \mid \hat{\theta}_{ML})$ conditional on the MLE is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{ML}$ and covariance matrix $n^{-1}I(\hat{\theta}_{ML})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density is asymptotically one in probability:*

$$p(\theta \mid \hat{\theta}_{ML}) \underset{n\to\infty}{\sim} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta), \qquad \forall \theta \in B_n$$

$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta) d\theta \underset{P}{=} 1.$$

Thus, over the subset B_n, which asymptotically contains all the mass of the normal density $\phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\cdot)$, the posterior conditional on the MLE is asymptotically equivalent to this normal distribution.

From a frequentist point of view, the BvM theorem, which concerns the full sample posterior $p(\theta \mid \mathcal{D})$, is a justification of the Bayesian approach for parameter estimation since the Bayesian confidence sets asymptotically coincide with the frequentist limit confidence sets (Freedman, 1999). Lemma 5.1 shows a similar result for the posterior conditional on the MLE $p(\theta \mid \hat{\theta}_{ML})$. Thus, Lemma 5.1 can also be viewed as a justification of the use of the posterior conditional on asymptotically normal MLE for parameter estimation. Note that results similar to the one provided by Lemma 5.1 have already been obtained for the estimation of an univariate location parameter; see Doksum and Lo (1990) and references therein.

5.3.3 Posterior conditional on an MPLE

We then obtained the following lemma which is analogous to Lemma 5.1 but which concerns an MPLE.

Lemma 5.2 (Asymptotic normality of the posterior conditional on an MPLE). *Assume that, given the vector θ under which the data $\mathcal{D}$ were generated, the p.d.f. of the MPLE $\hat{\theta}_{MPL}$ is equivalent to the normal density with mean vector θ and covariance matrix $g(n)^{-1}J(\theta)^{-1}$:*

$$p(\hat{\theta}_{MPL} \mid \theta) \underset{n\to\infty}{\sim} \phi_{\theta, g(n)^{-1}J(\theta)^{-1}}(\hat{\theta}_{MPL}),$$

where g is a positive increasing function such that $g(n) \to \infty$ and $J(\theta)$ is a positive-definite matrix. Under regularity assumptions[9]*, when $n \to \infty$, the posterior density $p(\theta \mid \hat{\theta}_{MPL})$ conditional on the MPLE is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{MPL}$ and covariance matrix $g(n)^{-1}J(\hat{\theta}_{MPL})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density is asymptoticallyone:*

[8] See Soubeyrand and Haon-Lasportes (2015) for details.
[9] See Soubeyrand and Haon-Lasportes (2015) for details.

$$p(\theta \mid \hat{\theta}_{MPL}) \underset{n\to\infty}{\sim} \phi_{\hat{\theta}_{MPL}, g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}}(\theta), \qquad \forall \theta \in B_n$$

$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{MPL}, g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}}(\theta) d\theta = 1.$$

Lemma 5.2 justifies the use of the posterior conditional on the MPLE for parameter estimation because the Bayesian confidence sets that are provided by this posterior asymptotically coincide with the frequentist limit confidence sets obtained by maximizing the pseudo-likelihood.

The asymptotic normality of the MPLE required in Lemma 5.2 has been obtained for various models, especially random Markov fields and spatial point processes; see Gaetan and Guyon (2008, chap. 5), Gourieroux et al. (1983), Møller and Waagepetersen (2003, chap. 9) and references therein. It has to be noted that information is lost when MPLE are used rather than MLE and, consequently, that estimation accuracy is decreased (e.g. this has been shown for simple Markovian models using asymptotic estimation variances (Gaetan and Guyon, 2008, chap. 5)).

5.3.4 Approximate posterior conditional on an MPLE

Here, we derive implications of Lemma 5.2 in the framework of ABC when (some of) the summary statistics are MPLE. We consider the (simple) ABC–rejection algorithm **A2** described in Section 5.2. The set of accepted parameters, say $\Theta_{\epsilon,I} = \{\theta_i : d(S_i, S) \leq \epsilon, i = 1, \ldots, I\}$, forms a sample from the posterior $p_\epsilon(\theta \mid S)$, where ϵ is the tolerance threshold and $S = s(\mathcal{D})$ is the set of summary statistics.

Theorem 5.3 (Asymptotic normality of the ABC–posterior conditional on an MPLE). *Consider the ABC–rejection algorithm that samples in the posterior $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ of θ conditional on the vector of summary statistics $S = \hat{\theta}_{MPL}$. Assume that when $\epsilon \to 0$, $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ converges pointwise to $p(\theta \mid \hat{\theta}_{MPL})$. Then, under assumptions of Lemma 5.2, when $n \to \infty$ and $\epsilon \to 0$, the posterior $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{MPL}$ and covariance matrix $g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density goes to one in probability and that does not depend on ϵ:*

$$p_\epsilon(\theta \mid \hat{\theta}_{MPL}) \underset{n\to\infty, \epsilon\to 0}{\sim} \phi_{\hat{\theta}_{MPL}, g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}}(\theta), \qquad \forall \theta \in B_n$$

$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{MPL}, g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}}(\theta) d\theta \underset{P}{=} 1.$$

As explained in the introduction of this section, this result leads us to conjecture that, for some stochastic implicit models, (i) the ABC–posterior distribution conditional on an MPLE is asymptotically normal and centered around the MPLE and (ii) resulting point estimates and confidence sets converge towards their frequentist analogues. We also provided in Soubeyrand

and Haon-Lasportes (2015) an analogous result when the MPLE is used in conjunction with supplementary statistics.

5.3.5 Application to a toy example

The simplified example presented here illustrates the application of ABC conditional on an MPLE and a supplementary statistic. The model $\mathcal{M}_\theta$ under consideration is the following bivariate normal distribution:

$$\mathcal{N}\left(\binom{\mu}{\mu}, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right),$$

parameterized by the mean μ and the correlation ρ; we set $\theta = (\mu, \rho)$. Observed data $\mathcal{D} = \{(D_k^{(1)}, D_k^{(2)}) : k = 1, \ldots, n\}$ are $n = 100$ vectors independently drawn under this normal distribution with $\mu = 0$ and $\rho = 0.5$. We use a uniform prior distribution π over the rectangular domain $(-3, 3) \times (-1, 1)$. The maximum likelihood estimates of μ and ρ are the empirical mean of $(D_k^{(1)} + D_k^{(2)})/2$ and the empirical correlation of $(D_k^{(1)}, D_k^{(2)})$, $k = 1, \ldots, n$. Here, we applied ABC with the two following statistics:

$$S = s(\mathrm{d}) = \begin{pmatrix} \hat{\mu}_{MPL} \\ S_0 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} D_k^{(1)} \\ \mathbf{1}\{\mathrm{sign}(D_k^{(1)}) = \mathrm{sign}(D_k^{(2)})\} \end{pmatrix},$$

where $\hat{\mu}_{MPL}$ is an MPLE of μ that uses only partial information contained in the sample (i.e. only the first component of sampled vectors), and S_0 is a supplementary statistic that gives the mean number of vectors in the sample whose components $D_k^{(1)}$ and $D_k^{(2)}$ have the same sign ($\mathbf{1}\{\cdot\}$ is the indicator function).

To assess the convergence of ABC when ϵ tends to zero, we applied ABC with varying ϵ, with $I = 10^5$ simulations, and with the distance $d(S_i, S) = (\hat{\mu}_{MPL,i} - \hat{\mu}_{MPL})^2 + (S_{0,i} - S_0)^2$, where $S_i = (\hat{\mu}_{MPL,i}, S_{0,i})$ is the vector of statistics computed for the simulation i. As usual in ABC–rejection, instead of fixing ϵ, we fixed the sample size τ of the posterior sample (i.e. the number of accepted parameter vectors); note that ϵ decreases when τ decreases. The sample size τ was fixed at values ranging from 10 to 5000. For each value of τ, we computed the local posterior probability[10] (LPP) around the true parameter vector $\theta = (0, 0.5)$. We expect that this LPP increases with the efficiency of the inference procedure. The LPP was computed for 50000 datasets and Figure 5.3 shows its mean and standard deviation when τ varies. The mean LPP around the true parameters increases when the sample size τ (and ϵ) tends to zero; meanwhile, the dispersion of the LPP increases. This is the signature of the classical bias–variance trade-off.

[10] The LPP around the true parameter vector is defined as the proportion of accepted parameter vectors in the small rectangle $[-0.015, 0.015] \times [-0.005, 0.005]$ whose center is $\theta = (0, 0.5)$ and whose sides are 200 times smaller than the sides of the parameter space $(-3, 3) \times (-1, 1)$.

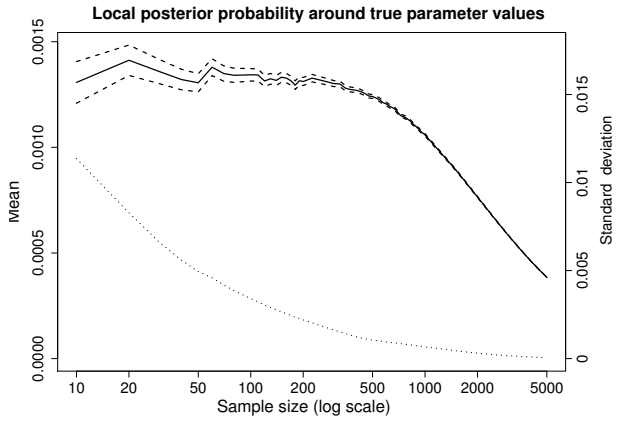


Fig. 5.3. Mean (solid line), pointwise 95%-confidence envelopes (dashed lines) and standard deviation (dotted line) of the local posterior probability around the true parameter vector $\theta = (0, 0.5)$ as a function of sample size τ.

To automatically select the sample size τ and optimize the distance between summary statistics, we applied the procedure presented in Section 5.2[11]. Thus, the distance is weighted: $d(S_i, S; w_1, w_2) = w_1(\hat{\mu}_{MPL,i} - \hat{\mu}_{MPL})^2 + w_2(S_{0,i} - S_0)^2$, and the triplet (τ, w_1, w_2) is optimized under constraints using the BMSE. For one of the 50000 datasets simulated above, Figure 5.4 shows ABC–posterior samples obtained when d is not weighted and τ is fixed at 5000, 1000 and 200, and when d is weighted and (τ, w_1, w_2) is optimized. The red contour line shows the smallest 95%-posterior area obtained with the classical Bayesian computation. The first three panels illustrate the bias–variance trade-off when τ tends to zero. The fourth panel illustrates the difference between the classical Bayesian inference conditional on all data and the ABC inference conditional on partial information and with optimized τ (here $\tau = 585$). The relevancy of the optimized sample size $\tau = 585$ can be seen by projecting this value on Figure 5.3: (i) the expected LPP around the true parameters is comparable to expected LPP obtained with lower τ, and (ii) the standard deviation of the LPP is strongly decreased compared with standard deviations obtained with lower τ.

5.3.6 ABC, MPLE and real-life studies

The asymptotic results presented above were obtained for a large but limited class of models satisfying regularity assumptions. In real-life studies where ABC is applied, these assumptions cannot be checked and, consequently, our

[11] The procedure presented in Section 5.2 was developed for functional statistics but can obviously be adapted to vectors of statistics.

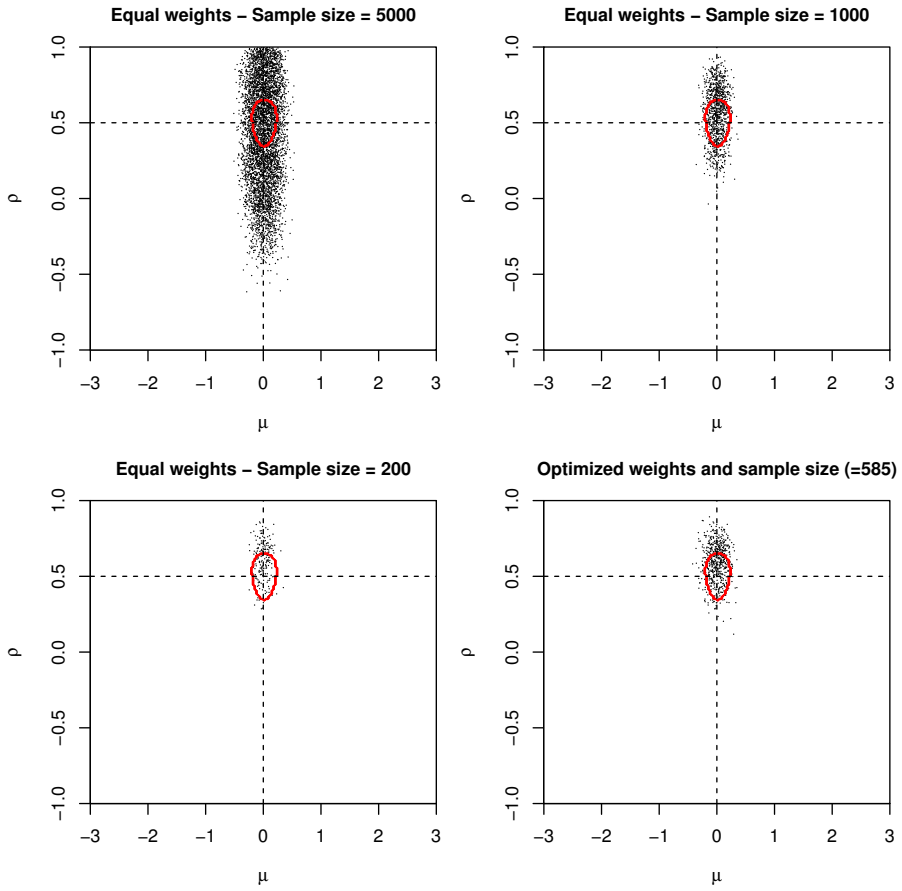


Fig. 5.4. ABC–posterior samples (dots) obtained when d is not weighted and τ is fixed at 5000 (top left), 1000 (top right) and 200 (bottom left), and when d is weighted and (τ, w_1, w_2) is optimized (bottom right). In each panel, dashed lines are intersecting at the true value $(0, 0.5)$ of the parameter vector $\theta = (\mu, \rho)$ and the red contour line gives the smallest 95%-posterior area obtained with the classical Bayesian computation.

asymptotic results may not hold. Therefore, it is crucial (i) to combine MPLE and supplementary summary statistics to tend to a set of sufficient summary statistics (Joyce and Marjoram, 2008) and (ii) to apply a method for selecting, weighting or transforming the summary statistics to avoid to take into account non-relevant statistics (in the toy example, we applied the approach presented in Section 5.2).

The strong implication of using MPLE as summary statistics in ABC is that an analytic work has to be made: the dependence structure of the model has to be simplified to write a tractable pseudo-likelihood and, eventually, to find an analytic expression for the maximizer. This additional work is however expected to yield relevant summary statistics directly informing (a subset of) the parameters. To illustrate the interest of combining ABC and MPLE in real-life examples, I started to study how to derive MPLE for parameters arising in dispersal models used to analyze metapopulation data. Indeed, it should be possible to obtain rapidly-computable MPLE for these models because they can be transformed into simple GLM by ignoring some dependencies induced by the dispersal kernel.

6

Miscellaneous

6.1 Snapshot of other contributions

6.1.1 Statistical tests

Mrkvička et al. (2012) proposed a testing procedure for marked spatial point processes with non-stationary marks. The objective of the test is to assess the goodness-of-fit of the mark distribution when this distribution depends on an unknown parameter vector that is spatially heterogeneous. The proposed procedure was applied, for instance, to a marked spatio-temporal point process related to the Classic Maya collapse.

Soubeyrand et al. (2014a) provides statistical tools to analyze fragmented time directionality in time series and the spatial distribution of this directionality. These tools are available in the `R`-package `FeedbackTS` available on CRAN (`https://cran.r-project.org/web/packages/FeedbackTS/`) and were used to elucidate feedbacks processes in historical[1] daily rainfall data collected in Australia and the USA; see also Bigg et al. (2015) and the website `http://w3.avignon.inra.fr/rainfallfeedback/`.

Soubeyrand et al. (2014c) proposed a method to rank pathogen strains in regard to their contribution to natural epidemics, and to assess the statistical significance of the ranking. This approach, which links genetic data (i.e. the genetic diversity of the pathogen) and epidemiological data (i.e. the spatio-temporal spread of the disease), was applied to assess the epidemiological performance of several strains of powdery mildew collected from its host plant *Plantago lanceolata*. It is implemented in the `R`-package `StrainRanking` (`https://cran.r-project.org/web/packages/StrainRanking/`).

Kretzschmar et al. (2010) proposed a permutation-based testing procedure to study the spatial structure of a population observed in non-Euclidean space, such as a tree. This procedure was applied to investigate the aggregation patterns of aphids on *Citrus* trees.

[1] These data are daily rainfall measurements collected over approximately 100 years.

6.1.2 Spatio-temporal modeling

Penczykowski et al. (2015) investigated the link between winter conditions and the metapopulation dynamics of powdery mildew on its host plant *Plantago lanceolata* in Åland archipelago. A particular focus was put on the effect of winter conditions on the spatial synchrony of the pathogen. In this study, we developed a stochastic patch occupancy model allowing us to explore the effect of warmer winters on the synchrony pattern.

Crété et al. (2013) proposed a continuous time-and-state epidemic model and an associated estimation method that allowed the fitting of the model to ordinal categorical data observed at discrete times. This approach was applied to analyze the spatio-temporal dynamics of apple scab.

6.1.3 Temporal modeling

Dussaubat et al. (2013) studied the effect of the parasite *Nosema ceranae* on honey bee workers. In this study, we developed a stochastic, mechanistic temporal model of bee activity fitted to time series providing the daily number of exits from the hive. To avoid making distributional assumptions, the model was fitted with a quasi-likelihood approach and demonstrated the negative effect of the parasite on the duration of bee activity.

Soubeyrand et al. (2007a) proposed a temporal survival model with time-varying covariates for analyzing the survival of mosquitofish. The model was fitted to group and time-period censored data and allowed for the characterization of the evolution of daily survival probability of mosquitofish during their growth.

6.1.4 Residual analysis

During my PhD I worked on the use of residuals to specify the distributional properties of latent processes included in hierarchical models. Soubeyrand et al. (2006) developed such a method for mixed models and frailty models with group random effects (i.e. random effects that are constant within groups of observations). Soubeyrand and Chadœuf (2007) developed the residual-based specification of hidden random fields included in hierarchical spatial models.

6.1.5 `R` packages

Recently, I have implemented some methods into `R` packages. These packages are briefly described in this section.

The `StrainRanking` package (available on `R CRAN` repository) utilizes demographic and genetic data collected during epidemics to rank pathogen strains with respect to their contribution to the epidemics (Soubeyrand et al., 2014c).

The `FeedbackTS` package (available on `R CRAN` repository) provides exploratory tools for the analysis of feedback (i.e. fragmented time directionality) in a single time series and in a set of time series collected across a spatial domain (Soubeyrand et al., 2014a).

The `CloNcaSe` package proposes a method to estimate the effective size and sex rate of a partially clonal population sampled at two different times (Ali et al., 2016).

The `GMCPIC` package (Generalized Monte Carlo plug-in test with calibration) proposes a computer intensive procedure to test the equality of two unknown vectors of probabilities p_1 and p_2. The `GMCPIC` package was specifically developed to test differences between pathogen compositions with small samples and sparse data.

6.2 Supervision

I have supervised the internships of 5 Master students and 2 Licence students. Most of these students followed programs in applied statistics. In addition, I have (co-)supervised 3 postdoctoral fellows and 2 PhD students presented in the following table.

Supervision of PhD students and postdoctoral fellows

Period	Description
2015	Advisor of E. Walker (Postdoc) Construction and R-packaging of a spatial exposure-hazard model
2014	Advisor of M. Leclerc (Postdoc) Estimation of the contribution of genetically modified maize to the large-scale mortality of non-target Lepidoptera
2012–2015	Co-advisor of L. Rimbaud (PhD) Model-based design and assessment of management strategies for epidemics in a heterogeneous landscape
2012–2013	Advisor of V. Garetta (Postdoc) Statistical analysis of re-emergence of plant pathogens
2007–2010	Co-advisor of V. Georgescu (PhD) Model-based clustering for multivariate and mixed-mode data: Application to multi-species spatial ecological data

L. Rimbaud wrote two articles: Rimbaud et al. (2015b) provided a review about sharka epidemiology with a focus on the optimization of control strategies, and Rimbaud et al. (2015a) proposed a method to estimate the mismatch between incubation and latency periods and applies this method to the sharka virus. L. Rimbaud is also preparing two other articles applying sensitivity analysis methods to investigate the efficiency of management strategies for sharka.

V. Georgescu wrote two articles and one technical report: Georgescu et al. (2009) developed a statistical approach based on model-based classification and tools of spatial statistics to explore assemblages of species abundances; Georgescu et al. (2014) and Georgescu et al. (2015) proposed automated MCEM algorithms for estimating the parameters of hierarchical models with multivariate and multitype response variables.

6.3 Teaching

My main activity is research. However, I have been involved in a few teaching programs which are listed in the following table.

Teaching

Period	Description
2011–2013	Participation in the construction and supervision of the BIOBAYES training school (Initiation to Bayesian statistics for researchers in biology) INRA; the one-week training school was organized once in 2011 and once in 2013
2008–2011	Participation in a lecture about Modeling in Life Science Master 2 at Centrale School of Marseille, 18h
2003–2005	Training lesson in Probability and Statistics Licence 2 at the University of Avignon, 68h
2002–2003	Training projects in Statistics Licence 3 at ENSAI, Rennes, 10h

The training school BIOBAYES, which is cited in this table, led us to write a textbook about Bayesian statistics under the name *Collectif BIOBAYES*. The complete reference is:

> Collectif BIOBAYES (2015). *Initiation à la Statistique Bayésienne – Bases Théoriques et Applications en Alimentation, Environnement, Epidémiologie et Génétique*. Editions Ellipses.

The *Collectif BIOBAYES* is formed by Albert I., Ancelet S., David O., Denis J.-B., Makowski D., Parent E., Rau A. and myself.

The lecture about Modeling in Life Science, which was mainly given by L. Roques and in which I was involved to illustrate how to estimate parameters of PDE-based models, led L. Roques to write a textbook. In this textbook, I participated in writing one chapter, which reproduces what I taught to students and deals with topics presented in Chapter 4 of the present document. The reference of this chapter is:

> Soubeyrand S. and Roques L. (2013). Problèmes inverses et estimations de paramètres. PDF file. In: Roques L. (Author). *Modèles de Réaction-Diffusion pour l'Ecologie Spatiale*. Editions QUAE, Versailles.

I have also participated to the writing of a chapter in an exercise book for plant epidemiologists:

> Lannou C. and Soubeyrand S. (2015). Measure of life-cycle traits of a biotrophic pathogen (pp.149-152). In Stevenson K.L. and Jeger M.J. *Exercices in Plant Disease Epidemiology, 2nd edition*. The American Phytopathological Society, St. Paul, Minnesota.

6.4 Network and projects

Since 2011, I am the coordinator of a network of researchers called *ModStat-SAP* (Modeling and Statistics in Animal and Plant Health), whose website is `http://informatique-mia.inra.fr/reseau-modstatsap/`. This network is funded by 3 divisions of INRA, and gather about 100 participants in French research institutes and universities. The main activities of the network is the organization of annual meetings and workshops.

The following table provides the list of the main projects I have been involved in.

Projects

Period	Description
2013–2016	PEERLESS – funded by the French research agency ANR Predictive ecological engineering for landscape ecosystem services and sustainability
2010–2015	PLANTFOODSEC Project – funded by the European Commission (FP7) Plant and Food Biosecurity
2010–2012	Group dispersal project – funded by the SPE division of INRA Building a theoretical framework for group dispersal in plant epidemiology I was the coordinator of this project
2009–2013	EMILE project – funded by the French research agency ANR Inference methods and software for evolutionary studies I was the coordinator of a work-package in this project

6.5 Perspectives of research

Most of the works presented in this manuscript sketch the functioning of systems under study and provide elements to improve how these systems are understood. This approach will remain the core of my future research. However, I will also orientate my research towards statistics for predictive epidemiology. Thus, in the analysis of a data set, the question will not only be *What happened?* but also *What will happen?* The following list of points provides examples of new topics in my research, which will allow me to investigate both questions.

6.5.1 Dispersal graphs substituting dispersal kernels

The dispersal kernel is an important component of my studies in particular (as illustrated in Chapter 2) and of dispersal studies in general. So far, I have mostly used and built relatively simple parametric forms, and in models with multiple sources across space and time, the dispersal was generally assumed to be stationary (i.e. the dispersal kernel is constant across space and time).

Numerous approaches have been investigated to increase the realism of windborne particle dispersal by taking into account airflows, turbulences and their spatio-temporal inhomogeneities (Nathan et al., 2011). In this vein, the explicit simulation of particle trajectories based on computationally intensive fluid dynamics tools (e.g. Navier-Stokes equation) is an attractive approach which is, however, not directly adaptable to inference issues.

Another approach (beyond dispersal kernels) to represent propagation processes is based on network modeling, which has been of particular interest in human and animal epidemiology (Colizza et al., 2006; Beaunée et al., 2015). Network modeling is a flexible way of representing space and contact that generally focuses on prevailing spatial elements (nodes of the network) and links (edges of the network). Networks have been used in numerous applications, including the study and the management of human, animal and plant diseases (see the review by Moslonka-Lefebvre et al., 2011). Although the use of networks has been fruitful in analyzing the spread of plant diseases and assessing the risk of disease emergence (e.g. Brooks et al., 2008; Harwood et al., 2009; Jeger et al., 2007), they have been relatively rarely used in plant epidemiology, while they could efficiently challenge other types of spatially explicit epidemic models (advantages of network modeling are, for instance, the possibility to derive theoretical properties, their scale-free formalization, the heterogeneity of link weights, and their adaptability to temporal changes in link weights).

One of my research perspectives is the construction of *dispersal graphs* based on both (i) computationally intensive fluid dynamics tools and (ii) network modeling. These dispersal graphs should allow the representation of non-stationary and anisotropic dispersal processes at mesoscales (from the large region to the continent). In a project submitted to the French national research

agency, we plan to construct dispersal graphs by using the HYSPLIT model[2], which provides trajectories of air masses and associated meteorological data. This software will be used to assess the connectivity (i.e. the weights of the network edges) between different areas (i.e. the nodes of the network). Then, the weighted network will be exploited for characterizing dispersal probabilities and, subsequently, specifying disease propagation models. Such models, based on meteorological processes that are *forecastable*, are particularly relevant for developing predictive epidemiology.

6.5.2 Genetic-space-time models that handle high-throughput sequencing

Concerning the genetic-space-time approach for inferring transmission trees (see Chapter 3), numerous perspectives can be developed in line with recently published methodological advances (e.g. Lau et al., 2015; Hall et al., 2015). In the medium term, I will attempt to investigate some methodological issues within this set of perspectives. One of these issues is the use of high-throughput sequencing (HTS) data of viral genomes sampled from infected hosts to infer transmission links of infectious diseases.

HTS data enable the characterization of viral populations at the intra-host level and, today, can be collected during epidemics, such as during the 2014 Ebola outbreak in West Africa (Gire et al., 2014). By revealing the polymorphic nature of intra-host populations of pathogens, HTS data are expected to give more insights on transmission links than more basic sequencing data, e.g. consensus and majority sequences (Stack et al., 2012; Wright et al., 2011). Therefore, taking into account HTS data in the genetic-space-time models for inferring transmission trees should enable more robust and accurate inferences.

In the existing genetic-space-time SEIR model, genetic information is accounted for in the reconstruction of transmission links by calculating the probabilities that sequences collected from potential source hosts could be directly related to those from infected hosts. When only the consensus sequence is utilized, the probability of being directly related reduces to the probabilities of genetic evolution between sequences. To utilize HTS data, the probability of being directly related must be decomposed into: (i) the probability of observed genetic changes between sequences, (ii) the probability of sub-sampling when infections occur (infection bottlenecks), and (iii) the probability of sub-sampling when sequences are sampled (sampling bottlenecks)[3].

I proposed to investigate this topic in a project submitted to the French national research agency. The project brings together scientists with skills in

[2] HYSPLIT model (Hybrid Single Particle Lagrangian Integrated Trajectory Model): *http://ready.arl.noaa.gov/HYSPLIT.php*.

[3] One can also add the probability of sequencing errors, which might result in artifactual minority variants.

statistics, modeling, software development, epidemiology, virology and evolutionary biology. The methodology developed during the project will be applied to data sets collected from epidemics caused by the Ebola virus, the swine and equine influenza viruses, the endive necrotic mosaic virus and the watermelon mosaic virus.

6.5.3 Hamiltonian Monte-Carlo for dispersal models

In my work, I have extensively used MCMC with Metropolis-Hastings samplers to infer model parameters and latent processes of hierarchical dispersal models. However, for models with complex dependence structures, such as those presented in Sections 2.4 and 2.5, the large computation times required for obtaining long enough chains limit the possibility of testing numerous model specifications.

An alternative sampler, which generally allows the reduction of computation times is the Hamiltonian sampler that was first introduced in the statistical physics literature (Duane et al., 1987) and applied afterwards to statistical inference issues; see Neal (2011), Girolami and Calderhead (2011) and references therein. The Hamiltonian sampler can be viewed as a specific Metroplis-Hastings sampler, in which the proposal is based on two key components: (i) some auxiliary random variables and (ii) an Hamiltonian dynamics applied to the parameters/variables to be updated and to the auxiliary variables. The auxiliary random variables allow the updating process to be stochastic. The (deterministic) Hamiltonian dynamics allows large jumps that are accepted with high probability.

Recently, Girolami and Calderhead (2011) and Zhang and Sutton (2014) proposed versions of Hamiltonian Monte-Carlo (HMC; i.e. MCMC with Hamiltonian sampler) which can efficiently tackle estimation for spatial hierarchical models such as log-Gaussian Cox point processes (Illian et al., 2008, chap. 6) and, by extension, spatial generalized linear mixed models (spatial GLMM; Diggle et al., 1998). The trick is to tune the algorithm with auxiliary variables, whose distributional characteristics depend on the current values of model parameters and latent variables.

One of my research perspectives is to investigate new versions of HMC adapted to hierarchical dispersal models, such as the model presented in Section 2.4 (which can be viewed as an extension of a spatial GLMM) and other dispersal models not based on latent Gaussian processes (which cannot be tackled with integrated nested Laplace approximations (INLA); Rue et al., 2009). In this aim, E. Walker (postdoctoral fellow) and I initiated a working group on HMC and produced an introductory technical report (Walker and Soubeyrand, 2016).

6.5.4 Statistical predictive epidemiology

Since 2014, I have been involved in a working group that discusses plant health crises in general, and the way INRA can contribute to crisis response

in particular. The management of plant health crises is handled by State administrations and services, which may request the support of research institutes such as INRA. One of the potential INRA contributions, which is of particular interest for me, concerns advanced analysis of surveillance data and modeling of epidemics.

For instance, after the detection of the bacterium *Xylella fastidiosa* in Corsica in the summer of 2015, the *Plant Health and Environment* division of INRA and the State administration in charge of this crisis have established a contract including a *data analysis and modeling* work package. In this work package, we will (i) analyze surveillance and environmental data to figure out the potential future extent of the disease caused by *Xylella fastidiosa*, (ii) build risk maps based on propagation models, and (iii) propose risk-based sampling for improving surveillance.

Beyond the *Xylella fastidiosa* crisis, my aim is to develop research on statistical predictive epidemiology in the context of plant health crises caused by the (re-)emergence of pathogens and pests. A key feature of this context is that crises are often caused by pathogens and pests that are not studied in depth in research institutes. Thus, we generally do not have at our disposal an adequate propagation model (i.e. a model adapted to the pathogen/pest and the environment of interest) developed before the occurrence of the crisis. To circumvent this difficulty, a generic approach must be developed. The generic approach that I have in mind will be based on (i) a library of propagation models including spatio-temporal SEIR stochastic models and PDE-based invasion models, (ii) the mechanistic-statistical approach for linking propagation models and surveillance/environmental data, (iii) the model-averaging approach for combining predictions obtained with several models, (iv) the use of genomic data, which can lead to more accurate inferences on dispersal capacities, and (v) risk-based sampling approaches for designing surveillance as a function of propagation predictions.

Thus, in the next few years, I will promote the construction of an integrated methodological framework, which should contribute to developing predictive epidemiology and improving emergency response in the context of plant health crises. This construction should be facilitated in the next few years with the recruitment of one master student in 2016, one engineer in statistics in 2016–2017, and probably one PhD student in 2016–2019.

Appendix

ABC algorithm including a pilot ABC run

A4. Carry out the next three steps,

1. Select the set $\mathcal{J}$ with size $|\mathcal{J}| < J$ formed by the indices $j \in \{1, \ldots, J\}$ corresponding to the $|\mathcal{J}|$ smallest distances between θ'_j and Θ_{pilot}, this distance being defined by:

$$\min \left\{ \sum_{k=1}^{K} \frac{(\theta'_{jk} - \theta_{pilot,k})^2}{V(\theta'_{jk})} : \theta_{pilot} \in \Theta_{pilot} \right\},$$

   where $\theta_{pilot,k}$, $k = 1, \ldots, K$, are the K components of θ_{pilot}.

2. Select the weight function and the acceptance rate that minimize the following partial MSE (PMSE) criterion:

$$\text{PMSE}_{\mathcal{J}}(w, \tau) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \sum_{k=1}^{K} \frac{(\hat{\theta}'_{jk}(w, \tau) - \theta'_{jk})^2}{V(\theta'_{jk})}. \tag{.1}$$

   Terms in Equation (.1) are the same than those in Equation (5.8) except that the sum is restricted to $\mathcal{J}$. The PMSE is minimized over the space function $\mathbb{W}$ and the interval $(0, 1]$:

$$(w_{opt}^{pilot}, \tau_{opt}^{pilot}) = \text{argmin}_{w,\tau \in \mathbb{W} \times (0,1]} \text{PMSE}_{\mathcal{J}}(w, \tau). \tag{.2}$$

3. For i in $\{1, \ldots, I\}$, accept θ_i if $d(S_i, S; w_{opt}^{pilot}) \leq \epsilon(\tau_{opt}^{pilot})$.

The set of accepted parameters forms a sample from the posterior distribution (5.5) with $\epsilon(\tau) = \epsilon(\tau_{opt}^{pilot})$ and with $B(S, \epsilon(\tau))$ equal to the ball with center S and radius $\epsilon(\tau_{opt}^{pilot})$ in the space $\mathbb{S}$ with distance $d(\cdot, \cdot; w_{opt}^{pilot})$. Here also weighting the distance modifies the posterior under which the accepted parameters are drawn. However, when $\epsilon(\tau_{opt}^{pilot})$ tends to zero, the new posterior distribution (like the one given in Equation (5.5)) may be a good approximation of $p(\theta \mid S)$ given in Equation (5.6).

References

Ali, S., Soubeyrand, S., Gladieux, P., Giraud, T., Leconte, M., Gautier, A., Mboup, M., Chen, W., Vallavieille-Pope, C., and Enjalbert, J. (2016). CloNcaSe: Estimation of sex frequency and effective population size by clonemate re-sampling in partially clonal organisms. *Molecular ecology resources*.

Allard, D. and Soubeyrand, S. (2012). Skew-normality for climatic data and dispersal models for plant epidemiology: when application fields drive spatial statistics. *Spatial Statistics*, 1:50–64.

Anderson, R. M., Donnelly, C. A., Ferguson, N. M., Woolhouse, M. E., Watt, C., Udy, H., MaWhinney, S., Dunstan, S., Southwood, T., Wilesmith, J., et al. (1996). Transmission dynamics and epidemiology of BSE in British cattle. *Nature*, 382:779–788.

Austerlitz, F., Dick, C. W., Dutech, C., Klein, E. K., Oddou-Muratorio, S., Smouse, P. E., and Sork, V. L. (2004). Using genetic markers to estimate the pollen dispersal curve. *Molecular Ecology*, 13:937–954.

Austerlitz, F. and Smouse, P. E. (2002). Two-generation analysis of pollen flow across a landscape. iv. estimating the dispersal parameter. *Genetics*, 161:355.

Aylor, D. E. (1990). The role of intermittent wind in the dispersal of fungal pathogens. *Annual Review of Phytopathology*, 28:73–92.

Aylor, D. E. (1999). Biophysical scaling and the passive dispersal of fungus spores: relationship to integrated pest management strategies. *Agricultural and Forest Meteorology*, 97:275–292.

Aylor, D. E. and Flesch, T. K. (2001). Estimating spore release rates using a lagrangian stochastic simulation model. *Journal of Applied Meteorology*, 40:1196–1208.

Barbu, V. S. and Limnios, N. (2008). *Semi-Markov chains and hidden semi-Markov models toward applications — Their Use in Reliability and DNA Analysis*, volume 191 of *Lecture Notes in Statistics*. Springer.

Battisti, A., Stastny, M., Netherer, S., Robinet, C., Schopf, A., Roques, A., and Larsson, S. (2005). Expansion of geographic range in the pine processionary

moth caused by increased winter temperatures. *Ecological Applications*, 15:2084–2096.

Beaunée, G., Vergu, E., and Ezanno, P. (2015). Modelling of paratuberculosis spread between dairy cattle farms at a regional scale. *Veterinary research*, 46:1–13.

Berliner, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research*, 108:8776.

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40:206–237.

Bigg, E. K., Soubeyrand, S., and Morris, C. E. (2015). Persistent after-effects of heavy rain on concentrations of ice nuclei and rainfall suggest a biological cause. *Atmospheric Chemistry and Physics*, 15:2313–2326.

Bochkina, N. A. and Green, P. J. (2014). The bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42:1850–1878.

Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39:2557–2584.

Borwanker, J., Kallianpur, G., and Prakasa Rao, B. L. S. (1971). The Bernstein-von Mises theorem for Markov processes. *The Annals of Mathematical Statistics*, pages 1241–1253.

Bourgeois, A., Gaba, S., Munier-Jolain, N., Borgy, B., Monestiez, P., and Soubeyrand, S. (2012). Inferring weed spatial distribution from multi-type data. *Ecological Modelling*, 226:92–98.

Bousset, L., Jumel, S., Garreta, V., Picault, H., and Soubeyrand, S. (2015). Transmission of *Leptosphaeria maculans* from a cropping season to the following one. *Annals of Applied Biology*, 166:530–543.

Brooks, C. P., Antonovics, J., and Keitt, T. H. (2008). Spatial and temporal heterogeneity explain disease dynamics in a spatially explicit network model. *The American Naturalist*, 172:149–159.

Buckland, S. T., Newman, K. B., Thomas, L., and Koesters, N. B. (2004). State-space models for the dynamics of wild animal populations. *Ecological Modelling*, 171:157–175.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach.* Springer, New York, 2nd edition.

Caillot, P. and Martin, F. (1972). Le modèle bayésien. *Annales de l'IHP, section B*, 8:183–210.

Cambra, M., Capote, N., Myrta, A., and Llácer, G. (2006). Plum pox virus and the estimated costs associated with sharka disease. *EPPO Bulletin*, 36:202–204.

Campbell, E. P. (2004). An introduction to physical-statistical modelling using Bayesian methods. Technical Report 49, CSIRO Mathematical & Information Sciences, Australia.

Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41:1999–2028.

Cheng, D. and Xiao, Y. (2016). Excursion probability of Gaussian random fields on sphere. *Bernoulli*, 22:1113–1130.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics. Modeling Spatial Uncertainty*. Wiley, New York.

Coetzee, P. and Nel, L. H. (2007). Emerging epidemic dog rabies in coastal south africa: a molecular epidemiological analysis. *Virus research*, 126:186–195.

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS*, 103:2015–2020.

Cottam, E. M., Wadsworth, J., Shaw, A. E., Rowlands, R. J., Goatley, L., Maan, S., Maan, N. S., Mertens, P. P. C., Ebert, K., Li, Y., Ryan, E. D., Juleff, N., Ferris, N. P., Wilesmith, J. W., Haydon, D. T., King, D. P., Paton, D. J., and Knowles, N. J. (2008). Transmission pathways of foot-and-mouth disease virus in the united kingdom in 2007. *PLoS Pathogen*, 4:e1000050.

Crété, R., Pumo, B., Soubeyrand, S., Didelot, F., and Caffier, V. (2013). A continuous time-and-state epidemic model fitted to ordinal categorical data observed on a lattice at discrete times. *Journal of Agricultural, Biological, and Environmental Statistics*, 18:538–555.

Cunniffe, N. J., Koskella, B., Metcalf, C. J. E., Parnell, S., Gottwald, T. R., and Gilligan, C. A. (2015). Thirteen challenges in modelling plant diseases. *Epidemics*, 10:6–10.

Dacunha-Castelle, D. and Duflo, M. (1982). *Probabilités et Statistiques. Problèmes à Temps Fixe*, volume 1. Masson, Paris.

Dargatz, C., Georgescu, V., and Held, L. (2005). Stochastic modelling of the spatial spread of influenza in Germany. Technical report, Ludwig-Maximilians Universität, Munich, Germany.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society C*, 47:299–350.

Djurle, A. and Yuen, J. E. (1991). A simulation model for *Septoria nodorum* in winter wheat. *Agricultural Systems*, 37:193–218.

Doksum, K. A. and Lo, A. Y. (1990). Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, 18:443–453.

Doutre, M.-S. (2005). Occupational contact urticaria and protein contact dermatitis. *European Journal of Dermatology*, 15:419–424.

Drovandi, C. C., Pettitt, A. N., and Faddy, M. J. (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society C*, 60:317–337.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195:216–222.

Dussaubat, C., Maisonnasse, A., Crauser, D., Beslay, D., Costagliola, G., Soubeyrand, S., Kretzchmar, A., and Le Conte, Y. (2013). Flight behavior and pheromone changes associated to *Nosema ceranae* infection of honey bee workers (*Apis mellifera*) in field conditions. *Journal of Invertebrate Pathology*, 113:42–51.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. Roy. Stat. Soc. B*, 74:419–474.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge University Press.

Frantzen, J. (2007). *Epidemiology and Plant Ecology: principles and applications*. World Scientific, New Jersey.

Freedman, D. (1999). On the Bernstein-Von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27:1119–1140.

Gaetan, C. and Guyon, X. (2008). *Modélisation et statistique spatiale*. Springer-Verlag, Berlin.

Georgescu, V., Desassis, N., Soubeyrand, S., Kretzschmar, A., and Senoussi, R. (2014). An automated MCEM algorithm for hierarchical models with multivariate and multitype response variables. *Communications in Statistics – Theory and Methods*, 43:3698–3719.

Georgescu, V., Desassis, N., Soubeyrand, S., Kretzschmar, A., and Senoussi, R. (2015). Model-based classification of multivariate and mixed-mode data with an automated Monte Carlo EM algorithm. Technical report, INRA, Biostatistics and Spatial Processes.

Georgescu, V., Soubeyrand, S., Kretzschmar, A., and Laine, A.-L. (2009). Exploring spatial and multitype assemblages of species abundances. *Biometrical Journal*, 51:979–995.

Ghil, M. and Childress, S. (1987). *Topics in Geophysical Fluid Dynamics, Dynamo Theory, and Climate Dynamics*. Springer, New York.

Gibson, G. J. (1997). Makov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *J. R. Statist. Soc. C*, 46:215–233.

Gilligan, C. A. (2008). Sustainable agriculture and plant diseases: an epidemiological perspective. *Philosophical Transactions of the Royal Society of London B*, 363:741–759.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345:1369–1372.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society B*, 73:123–214.

Gleim, A. and Pigorsch, C. (2013). Approximate Bayesian computation with indirect summary statistics. *preprint*.

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19:1327–1349.

Gneiting, T. and Raftery, A. E. (2005). Strictly proper scoring rules, prediction, and estimation. Technical Report 463R, Department of Statistics, University of Washington.

Gourieroux, C., Monfort, A., and Trognon, A. (1983). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52:681–700.

Gregory, P. H. (1945). The dispersion of air-borne spores. *Transactions of the British Mycological Society*, 28:26–72.

Gregory, P. H. (1968). Interpreting plant disease dispersal gradients. *Annual Review of Phytopathology*, 6:189–212.

Guyon, X. (1985). Estimation d'un champ par pseudo-vraisemblance conditionnelle : étude asymptotique et application au cas markovien. In des Facultés Universitaires de St Louis, E., editor, *Actes de la 6ème rencontre Franco-Belge de Statitisticiens*.

Hall, M., Woolhouse, M., and Rambaut, A. (2015). Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLoS Computational Biology*, 11:e1004613.

Hampson, K., Dushoff, J., Cleaveland, S., Haydon, D. T., Kaare, M., Packer, C., and Dobson, A. (2009). Transmission dynamics and prospects for the elimination of canine rabies. *PLoS Biology*, 7:e1000053.

Hanski, I. and Gaggiotti, O. E., editors (2004). *Ecology, Genetics, and Evolution of Metapopulations*. Elsevier Academic Press, Amsterdam.

Harrell, F. E. (2013). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media.

Harwood, T. D., Xu, X., Pautasso, M., Jeger, M. J., and Shaw, M. W. (2009). Epidemiological risk assessment using linked network and grid based modelling: *Phytophthora ramorum* and *Phytophthora kernoviae* in the UK. *Ecological Modelling*, 220:3353–3361.

Haydon, D. T., Kao, R. R., and Kitching, R. P. (2004). The UK foot-and-mouth disease outbreak — the aftermath. *Nature Reviews Microbiology*, 2:675–681.

Herrmann, I., Herrmann, T., and Wagner, S. (2011). Improvements in anisotropic models of single tree effects in cartesian coordinates. *Ecological Modelling*, 222:1333 – 1336.

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

Huet, S. (2004). *Statistical tools for nonlinear regression: a practical guide with S-PLUS and R examples*. Springer Science & Business Media.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *PNAS*, 101:15124–15129.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.

Ingold, C. T. (1971). *Fungal spores. Their liberation and dispersal*. Oxford University Press.

Jeger, M. J. (2000). Theory and plant epidemiology. *Plant Pathology*, 49:651–658.

Jeger, M. J., Pautasso, M., Holdenrieder, O., and Shaw, M. W. (2007). Modelling disease spread and control in networks: implications for plant sciences. *New Phytologist*, 174:279–297.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology*, 10:e1003457.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7:1–26.

Kareiva, P. M. (1983). Local movement in herbivorous insects: Applying a passive diffusion model to mark-recapture field experiments. *Oecologia*, 57:322–327.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 115:700–721.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci.*, 78:454–458.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Klein, E. K., Lavigne, C., Foueillassar, X., Gouyon, P.-H., and Larédo, C. (2003). Corn pollen dispersal: quasi-mechanistic models and field experiments. *Ecological Monographs*, 73:131–150.

Koizumi, S. and Kato, H. (1991). Dynamic simulation of blast epidemics using a multiple canopy spore dispersal model. In Teng, P. S., editor, *Rice Blast Modeling and Forecasting*, pages 75–88, Manila. International Rice Research Institute.

Kretzschmar, A., Soubeyrand, S., and Desassis, N. (2010). Aggregation patterns in hierarchy/proximity spaces. *Ecological Complexity*, 7:21–31.

Laine, A.-L. (2005). Spatial scale of local adaptation in a plant pathogen population. *Journal of Evolutionary Biology*, 18:930–938.

Laine, A.-L. (2008). Temperature-mediated patterns of local adaptation in a natural plant-pathogen metapopulation. *Ecol. Lett.*, 11:327–333.

Laine, A.-L. and Hanski, I. (2006). Large-scale spatial dynamics of specialist plant pathogen. *Journal of Ecology*, 94:217–226.

Lannou, C., Soubeyrand, S., Frezal, L., and Chadœuf, J. (2008). Autoinfection in wheat leaf rust epidemics. *New Phytologist*, 177:1001–1011.

Lau, M. S., Marion, G., Streftaris, G., and Gibson, G. (2015). A systematic Bayesian integration of epidemiological and genetic data. *PLoS Computational Biology*, 11:e1004633.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation, 2nd edition*, volume 31. Springer.

Lindley, D. V. (1965). *Introduction to probability and statistics from Bayesian viewpoint. Part 2: inference*. Cambridge University Press.

Lô-Pelzer, E., Bousset, L., Jeuffroy, M., Salam, M., Pinochet, X., Boillot, M., and Aubertot, J. (2010). SIPPOM-WOSR: a simulator for integrated pathogen population management of phoma stem canker on winter oilseed rape: I. Description of the model. *Field crops research*, 118:73–81.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380:2095–2128.

Marin, J. M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Stat. Comput.*, 22:1167–1180.

Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56:483–486.

McCartney, H. A. and Fitt, B. D. L. (2006). Dispersal of foliar fungal plant pathogens: mechanisms, gradients and spatial patterns. In Cooke, B. M., Jones, D. G., and Kaye, B., editors, *The Epidemiology of Plant Diseases, 2nd Ed.*, pages 159–192, Dordrecht. Springer.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd edition.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

McRoberts, N., Hughes, G., and Madden, L. (2003). The theoretical basis and practical application of relationships between different disease intensity measurements in plants. *Annals of applied Biology*, 142:191–211.

Mengersen, K. L., Pudlo, P., and Robert, C. P. (2013). Bayesian computation via empirical likelihood. *PNAS*, 110:1321–1326.

Molchanov, I. (1997). *Statistics of the Boolean Model for Practitioners and Mathematicians*. Wiley, Chichester.

Mollentze, N., Nel, L. H., Townsend, S., le Roux, K., Hampson, K., Haydon, D. T., and Soubeyrand, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B*, 281:20133251.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, Boca Raton.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *J. R. Statist. Soc. B*, 39:283–326.

Morelli, M. J., Thébaud, G., Chadœ uf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *Plos Computation Biology*, 8:e1002768.

Moslonka-Lefebvre, M., Finley, A., Dorigatti, I., Dehnen-Schmutz, K., Harwood, T., Jeger, M. J., Xu, X., Holdenrieder, O., and Pautasso, M. (2011). Networks in plant epidemiology: from genes to landscapes, countries, and continents. *Phytopathology*, 101:392–403.

Mrkvička, T. and Soubeyrand, S. (2015). On parameter estimation for doubly inhomogeneous cluster point processes. Technical report, INRA, Biostatistics and Spatial Processes.

Mrkvička, T., Soubeyrand, S., and Chadœuf, J. (2012). Goodness-of-fit test of the mark distribution in a point process with non-stationary marks. *Statistics and Computing*, 22:931–943.

Mrkvička, T., Muška, M., and Kubečka, J. (2014). Two step estimation for Neyman-Scott point process with inhomogeneous cluster centers. *Statistics and Computing*, 24:91–100.

Mrkvička, T., Myllymäki, M., and Hahn, U. (2015). Multiple Monte Carlo testing with applications in spatial point processes. arXiv:1506.01646 [stat.ME].

Murray, J. D. (2002). *Mathematical Biology*. Springer-Verlag, 3rd edition.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2015). Global envelope tests for spatial processes. arXiv:1307.0239 [stat.ME].

Nathan, R., Katul, G. G., Bohrer, G., Kuparinen, A., Soons, M. B., Thompson, S. E., Trakhtenbrot, A., and Horn, H. S. (2011). Mechanistic models of seed dispersal by wind. *Theoretical Ecology*, 4:113–132.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman and Hall – CRC Press, Boca Raton.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical Applications in Genetics and Molecular Biology*.

Okubo, A. and Levin, S. A. (2002). *Diffusion and ecological problems – modern perspectives*. Springer-Verlag, 2nd edition.

Ovaskainen, O. and Laine, A.-L. (2006). Inferring evolutionary signals form ecological data in a plant-pathogen metapopulation. *Ecology*, 87:880–891.

Papaïx, J., Adamczyk-Chauvat, K., Bouvier, A., Kiêu, K., Touzeau, S., Lannou, C., and Monod, H. (2014). Pathogen population dynamics in agricultural landscapes: The ddal modelling framework. *Infection, Genetics and Evolution*, 27:509–520.

Parham, P. E. and Ferguson, N. M. (2006). Space and contact networks: capturing the locality of disease transmission. *Journal of The Royal Society Interface*, 3:483–493.

Penczykowski, R. M., Walker, E., Soubeyrand, S., and Laine, A.-L. (2015). Linking winter conditions to regional disease dynamics in a wild plant–pathogen metapopulation. *New Phytologist*, 205:1142–1152.

Porcu, E., Bevilacqua, M., and Genton, M. G. (2015). Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *Journal of the American Statistical Association*, (in press).

Posel, D. and Marx, C. (2013). Circular migration: a view from destination households in two urban informal settlements in south africa. *The Journal of Development Studies*, 49:819–831.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth oh human y chromosomes: a study of y chromosome mibrosatellites. *Molecular Biology and Evolution*, 16:1791–1798.

Rapilly, F. (1991). *L'Epidémiologie en Pathologie Végétale*. INRA Editions, Paris.

Rieux, A., Soubeyrand, S., Bonnot, F., Klein, E., Ngando, J., Mehl, A., Ravigne, V., Carlier, J., and de Lapeyre de Bellaire, L. (2014). Long-distance wind-dispersal of spores in a fungal plant pathogen: estimation of anisotropic dispersal kernels from an extensive field experiment. *PloS one*, 9:e103225.

Rimbaud, L., Dallot, S., Delaunay, A., Borron, S., Soubeyrand, S., Thébaud, G., and Jacquot, E. (2015a). Assessing the mismatch between incubation and latency for vector-borne diseases: the case of sharka. *Phytopathology*, 115:1408–1416.

Rimbaud, L., Dallot, S., Gottwald, T. R., Decroocq, V., Soubeyrand, S., Jacquot, E., Labonne, J., and Thébaud, G. (2015b). Sharka epidemiology and worldwide management strategies: learning lessons to optimize disease control in perennial plants. *Annual Review of Phytopathology*, 53:357–378.

Rivoirard, V. and Rousseau, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40:1489–1523.

Rivot, E., Prévost, E., Parent, E., and Baglinière, J. L. (2004). A Bayesian state-space modelling framework for fitting a salmon stage-structured population dynamic model to multiple time series of field data. *Ecological Modelling*, 179:463–485.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

Robinet, C., Baier, P., Josef, P., Schopf, A., and Roques, A. (2007). Modelling the effects of climate change on the potential feeding activity of *Thaumetopoea pityocampa* (den. & schiff.) (lep., notodontidae) in France. *Global Ecology and Biogeography*, 16:460–471.

Rohani, P., Keeling, M. J., and Grenfell, B. T. (2002). The interplay between determinism and stochasticity in childhood diseases. *The American Naturalist*, 159:469–481.

Rohatgi, V. K. (2003). *Statistical Inference*. Dover Publication, Mineola.

Ronce, O. (2007). How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics*, 38:231–253.

Roques, L., Chekroun, M. D., Cristofol, M., Soubeyrand, S., and Ghil, M. (2014). Parameter estimation for energy balance models with memory. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 470:20140349.

Roques, L., Soubeyrand, S., and Rousselet, J. (2011). A statistical-reaction-diffusion approach for analyzing expansion processes. *Journal of Theoretical Biology*, 274:43–51.

Roques, L., Walker, E., Franck, P., Soubeyrand, S., and Klein, E. K. (2016). Using genetic data to estimate diffusion rates in heterogeneous landscapes. *Journal of Mathematical Biology*.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145:1219.

Rousset, F. (2012). Demographic consequences of the selective forces controlling density-dependent dispersal. In Clobert, J., Beguette, M., Benton, T. G., and Bullock, J. M., editors, *Dispersal Ecology and Evolution*, pages 266–279, Oxford. Oxford University Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12:1151–1172.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, 71:319–392.

Salamatin, A., Lipenkov, V., Barkov, N., Jouzel, J., Petit, J., and Raynaud, D. (1998). Ice core age dating and paleothermometer calibration based on isotope and temperature profiles from deep boreholes at Vostok Station (East Antarctica). *Journal of Geophysical Research*, 103:8963–8977.

Sasváry, Z. (1994). *Positive Definite and Definitizable Functions*. Akademie Verlag, Berlin.

Scherm, H., Ngugi, H., and Ojiambo, P. (2006). Trends in theoretical plant epidemiology. *European Journal of Plant Pathology*, 115:61–73.

Shigesada, N. and Kawasaki, K. (1997). *Biological Invasions: Theory and Practice*. Oxford University Press, Oxford.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika*, 38:196–218.

Snow, J. (1855). *On the Mode of Communication of Cholera*. Churchill, London, 2 edition.

Soubeyrand, S. (2016). Construction of semi-Markov genetic-space-time SEIR models and inference. *Journal de la Société Française de Statistique*, 157:129–152.

Soubeyrand, S., Beaudouin, R., Desassis, N., and Monod, G. (2007a). Model-based estimation of the link between the daily survival probability and a time-varying covariate, application to mosquitofish survival data. *Mathematical biosciences*, 210:508–522.

Soubeyrand, S., Carpentier, F., Desassis, N., and Chadœuf, J. (2009a). Inference with a contrast-based posterior distribution and application in spatial statistics. *Statistical Methodology*, 6:466–477.

Soubeyrand, S., Carpentier, F., Guiton, F., and Klein, E. K. (2013). Approximate Bayesian computation with functional statistics. *Statistical Applications in Genetics and Molecular Biology*, 12:17–37.

Soubeyrand, S. and Chadœuf, J. (2007). Residual-based specification of a hidden random field included in a hierarchical spatial model. *Computational Statistics & Data Analysis*, 51:6404–6422.

Soubeyrand, S., Chadœuf, J., Sache, I., and Lannou, C. (2007b). A frailty model to assess plant disease spread from individual count data. *Journal of Data Science*, 5:63–86.

Soubeyrand, S., Enjalbert, J., Kretzschmar, A., and Sache, I. (2009b). Building anisotropic sampling schemes for the estimation of anisotropic dispersal. *Annals of Applied Biology*, 154:399–411.

Soubeyrand, S., Enjalbert, J., and Sache, I. (2008a). Accounting for roughness of circular processes: Using gaussian random processes to model the anisotropic spread of airborne plant disease. *Theoretical Population Biology*, 73:92–103.

Soubeyrand, S., Enjalbert, J., Sanchez, A., and Sache, I. (2007c). Anisotropy, in density and in distance, of the dispersal of yellow rust of wheat: Experiments in large field plots and estimation. *Phytopathology*, 97:1315–1324.

Soubeyrand, S. and Haon-Lasportes, E. (2015). Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC. *Statistics and Probability Letters*, 107:84–92.

Soubeyrand, S., Held, L., Höhle, M., and Sache, I. (2008b). Modelling the spread in space and time of an airborne plant disease. *Journal of the Royal Statistical Society C*, 57:253–272.

Soubeyrand, S., Laine, A., Hanski, I., and Penttinen, A. (2009c). Spatiotemporal structure of host-pathogen interactions in a metapopulation. *The American Naturalist*, 174:308–320.

Soubeyrand, S., Morris, C. E., and Bigg, E. K. (2014a). Analysis of fragmented time directionality in time series to elucidate feedbacks in climate data. *Environmental Modelling & Software*, 61:78–86.

Soubeyrand, S., Mrkvička, T., and Penttinen, A. (2014b). A nonstationary cylinder–based model describing group dispersal in a fragmented habitat. *Stochastic Models*, 30:48–67.

Soubeyrand, S., Neuvonen, S., and Penttinen, A. (2009d). Mechanical-statistical modelling in ecology: from outbreak detections to pest dynamics. *Bulletin of Mathematical Biology*, 71:318–338.

Soubeyrand, S. and Roques, L. (2014). Parameter estimation for reaction-diffusion models of biological invasions. *Population ecology*, 56(2):427–434.

Soubeyrand, S., Roques, L., Coville, J., and Fayard, J. (2011). Patchy patterns due to group dispersal. *Journal of Theoretical Biology*, 271:87–99.

Soubeyrand, S., Sache, I., Hamelin, F., and Klein, E. K. (2015). Evolution of dispersal in asexual populations: to be independent, clumped or grouped? *Evolutionary Ecology*, 29:947–963.

Soubeyrand, S., Sache, I., Lannou, C., and Chadœuf, J. (2006). Residual-based specification of the random-effects distribution for cluster data. *Statistical Methodology*, 3:464–482.

Soubeyrand, S., Thébaud, G., and Chadœuf, J. (2007d). Accounting for biological variability and sampling scale: a multi-scale approach to building epidemic models. *Journal of the Royal Society Interface*, 4:985–997.

Soubeyrand, S., Tollenaere, C., Haon-Lasportes, E., and Laine, A.-L. (2014c). Regression-based ranking of pathogen strains with respect to their contribution to natural epidemics. *PloS one*, 9:e86591.

Stack, J. C., Murcia, P. R., Grenfell, B. T., Wood, J. L., and Holmes, E. C. (2012). Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proceedings of the Royal Society of London B*, page rspb20122173.

Starrfelt, J. and Kokko, H. (2012). The theory of dispersal under multiple. In Clobert, J., Beguette, M., Benton, T. G., and Bullock, J. M., editors, *Dispersal Ecology and Evolution*, pages 19–28. Oxford University Press.

Stockmarr, A. (2002). The distribution of particles in the plane dispersed by a simple 3-dimensional diffusion process. *Journal of Mathematical Biology*, 45:461–469.

Stoyan, D., Kendall, W. S., and Mecke, J. (1995). *Stochastic Geometry and its Applications, 2nd Ed.* Wiley, Chichester.

Strange, R. N. (2003). *Introduction to plant pathology.* John Wiley & Sons.

Talbi, C., Lemey, P., Suchard, M. A., Abdelatif, E., Elharrak, M., Nourlil, J., Faouzi, A., Echevarría, J. E., Vazquez Moron, S., Rambaut, A., et al. (2010). Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Pathog*, 6:e1001166.

Tamaki, K. (2008). The Bernstein-von Mises theorem for stationary processes. *J. Japan Statist. Soc*, 38:311–323.

Theophrastus (1916). *Enquiry into plants – English translation by Sir Arthur Hort.* Harvard University Press, London.

Travis, J. M. J. and Dytham, C. (2002). Dispersal evolution during invasions. *Evolutionary Ecology Research*, 4:1119–1129.

Tuffley, C. and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607.

Tufto, J., Engen, S., and Hindar, K. (1997). Stochastic dispersal processes in plant populations. *Theoretical Population Biology*, 52:16–26.

Valdazo-González, B., Kim, J. T., Soubeyrand, S., Wadsworth, J., Knowles, N. J., Haydon, D. T., and King, D. P. (2015). The impact of within-herd genetic variation upon inferred transmission trees for foot-and-mouth disease virus. *Infection, Genetics and Evolution*, 32:440–448.

Van Lieshout, M. N. M. and Van Zwet, E. W. (2001). Exact sampling from conditional boolean models with applications to maximum likelihood inference. *Advances in Applied Probability*, 33:339–353.

Wagner, S., Walder, K., Ribbens, E., and Zeibig, A. (2004). Directionality in fruit dispersal models for anemochorous forest trees. *Ecological modelling*, 179:487–498.

Wälder, K., Näther, W., and Wagner, S. (2009). Improving inverse model fitting in trees–anisotropy, multiplicative effects, and bayes estimation. *Ecological Modelling*, 220:1044–1053.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Stat. Soc. B*, 31:80–88.

Walker, E. and Soubeyrand, S. (2016). Hamiltonian Monte Carlo in practice. Technical report, INRA, Biostatistics and Spatial Processes.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient Approximate Bayesian Computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182:1207–1218.

Wei, G. C. G. and Tanner, M. A. (1990). Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. of the American Statistical Association*, 85:699–704.

Weinan, E. and Engquist, B. (2003). Multiscale modeling and computation. *Notices Am. Math. Soc.*, 50:10621070.

Weinan, E., Engquist, B., and Huang, Z. (2003). Heterogeneous multiscale method: a general methodology for multiscale modeling. *Phys. Rev. B*, 67:09210–1.

Wikle, C. K. (2003a). Hierarchical models in environmental science. *International Statistical Review*, 71:181–199.

Wikle, C. K. (2003b). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84:1382–1394.

Wilson, J. D. (2000). Trajectory models for heavy particles in atmospheric turbulence: comparison with observations. *Journal of Applied Meteorology*, 39:1894–1912.

Wright, C. F., Morelli, M. J., Thébaud, G., Knowles, N. J., Herzyk, P., Paton, D. J., Haydon, D. T., and King, D. P. (2011). Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of virology*, 85:2266–2275.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions*, volume 1. Springer, New York.

Ypma, R. J. F., Bataille, A. M. A., Stegeman, A., Koch, G., Wallinga, J., and van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B*, 279:444–450.

Ypma, R. J. F., Jonges, M., Bataille, A., Stegeman, A., Koch, G., van Boven, M., Koopmans, M., van Ballegooijen, W. M., and Wallinga, J. (2013). Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *Journal of Infectious Diseases*, 207:730–735.

Zhang, Y. and Sutton, C. (2014). Semi-separable Hamiltonian Monte Carlo for inference in Bayesian hierarchical models. In *Advances in Neural Information Processing Systems*, pages 10–18.

Summarizing text samples

"This document, which was written to obtain the *habilitation à diriger des recherches* (i.e. the accreditation to supervise research), illustrates what kind of researcher I am: a researcher who carries out his own research and who contributes to the research of colleagues; a researcher who tends to explore various fields, techniques and issues, but who is consistently interested in recurrent topics."

"Since the beginning of my PhD studies, I have participated in the development of [quantitative analysis for epidemiology] and tried to bring original ideas by carrying out research at the interplay between statistics, modeling, probability, plant epidemiology and, occasionally, animal epidemiology. Carrying out such multidisciplinary research led me to be a researcher in applied statistics."

"In my research practice, I am not focused on a given methodology, but I exploit diverse statistical and modeling tools and explore some of them in depth. The main tools I have used are spatial and spatio-temporal point processes, continuous-time Markov and semi-Markov processes, state-space models and estimation algorithms."

"Most of the works presented in this manuscript sketch the functioning of systems under study and provide elements to improve how these systems are understood. This approach will remain the core of my future research. However, I will also orientate my research towards statistics for predictive epidemiology. Thus, in the analysis of a data set, the question will not only be *What happened?* but also *What will happen?*"