Inferring and Predicting Invasive Species Dynamics - Focus on Xylella fastidiosa

Candy Abboud

AIX-MARSEILLE UNIVERSITÉ
ED 184 - Mathématiques et Informatique de Marseille
UR INRA Biostatistique et Processus Spatiaux (UR 546, BioSP)

Thèse pour obtenir le grade universitaire de docteure

Présentée par
Candy Abboud

Spécialité
Mathématiques Appliquées

Intitulée

Inférer et Prédire les Dynamiques D'espèces Invasives Focus sur *Xylella fastidiosa*

Jury de Soutenance

Rapportrices:	Anne Gégout-Petit	Professeure des universités (Université de Lorraine)
	Sophie Donnet	Chargée de recherche (AgroParisTech/INRA - Paris)
Présidente:	Florence Hubert	Professeure des universités (Aix Marseille Université)
Examinateur:	Jimmy Garnier	Chargé de recherche (Université Savoie Mont-Blanc)
Directeurs de thèse:	Samuel Soubeyrand	Directeur de recherche (INRA - Avignon)
	Éric Parent	Professeur (AgroParisTech/INRA - Paris)

To my beautiful angel and the source of my strength.

May your soul rest in peace Nana!

Abstract

The spread of invasive alien species to new areas has always been an appealing research topic for mathematicians as well as for biologists. In particular, many investigations are carried out to reconstruct the past dynamics of the alien species and to predict its future spread. In essence, the thesis research aims to provide a generic methodology (i.e. scalable to various invasive species) that improves the predictions of an invasive species dynamics for which no dedicated model is available and whose initial conditions (i.e. date and location of the introduction of invasive species) are unknown. In order to achieve this goal, we proceed in two complementary lines of research. The first one is to propose a model&data-based inference method of biological invasions, in the framework of the so-called mechanistic-statistical approach. This method allows us to jointly estimate the introduction point (date and location of the invasive species arrival) and other parameters of the dynamics related to diffusion, reproduction and death. It is hinged on (i) a partial differential equation that offers a phenomenological and concise description of the invasive species dynamics in a heterogeneous domain, (ii) a stochastic model that represents the observation process, which allows to fit the partial differential equation to the data and (iii) a statistical Bayesian inference procedure, the adaptive multiple importance sampling algorithm, for estimating model parameters. To gain in realism, the phenomenological deterministic model could be replaced by a stochastic model, as for example a stochastic partial differential equation or spatio-temporal point process. However, such models may induce additional difficulties in estimation because of the supplementary parameters and latent variables. Models issued from the framework of Piecewise-deterministic Markov Process could be an appealing and interesting alternative to balance the trade-off between model realism and estimation easiness. In the framework presented above, preference was given to the use of generic spatio-temporal propagation models since the main processes underlying the spread of an alien species are usually unknown. However, predictions that can be drawn from those models are not optimal because they are affected by the assumptions made in the corresponding models, and do not take into account the uncertainty about the model form. The approach I use to overcome this problem is the so-called Bayesian *model-averaging*. This method consists of combining predictions drawn from competing models in order to obtain a unique and ameliorated prediction. This technique has been previously used in environmental sciences. Nevertheless, it is not widespread in the field of epidemiology. One of the methodological goals of the PhD is to investigate its application and usefulness in predictive epidemiology.

The case study of my thesis is the phytopathogenic bacterium *Xylella fastidiosa* for which abundant spatio-temporal and binary post-introduction surveillance data were collected from an intensive surveillance plan implemented by governmental agencies after the first pathogen detection in Corsica in 2015. This quarantine pathogen that has significantly impacted olive production in Italy and that presents a drastic risk of change to the environment for its ability to reach a large variety of plants, is susceptible to cause in France a major sanitary crisis, as the one caused in Italy since 2013 where the socio-economical impacts are considerable.

Keywords: Mechanistic-statistical approach, partial differential equations, Bayesian inference, Bayesian model-averaging, predictive epidemiology, biological invasions, *Xylella fastidiosa*.

Résumé

L'invasion de territoires par des espèces allogènes a toujours été un sujet attrayant pour les mathématiciens aussi bien que pour les biologistes. En particulier, de nombreux travaux sont menés afin de reconstruire la dynamique passée d'espèces envahissantes. Fondamentalement, le projet de thèse porte sur la recherche d'une méthodologie générique (i.e. adaptable à diverses espèces invasives), permettant l'amélioration des prédictions d'une invasion biologique pour laquelle on ne dispose pas de modèle spécifique et dont les conditions initiales (i.e. la date et le lieu d'introduction de l'espèce invasive) sont inconnues. Pour atteindre cet objectif, on procède suivant deux axes de recherche complémentaires. Dans le premier axe, on s'intéresse à l'inférence des invasions biologiques à partir d'un modèle spatio-temporel de propagation et de données collectées, en suivant une approche mécanistico-statistique. Cette méthode permet d'estimer d'une façon jointe le point d'introduction (date et site de l'arrivée de l'espèce invasive) et d'autres paramètres de la dynamique reliés à la diffusion, la reproduction et la mortalité. Elle repose sur (i) une équation aux dérivées partielles offrant une représentation phénoménologique et concise d'une dynamique qui envahit un domaine hétérogène, (ii) un modèle stochastique représentant le processus d'observation permettant d'ajuster l'équation aux dérivées partielles aux données et (iii) une méthode d'inférence statistique Bayésienne, l'adaptive multiple importance sampling algorithm, pour estimer les paramètres du modèle. Pour gagner en réalisme, le modèle phénoménologique déterministe peut être remplacé par un modèle stochastique, comme par exemple une équation aux dérivées partielles stochastique ou un processus de points spatio-temporel. Cependant, de tels modèles peuvent induire des difficultés d'estimation du fait des paramètres supplémentaires et des variables latentes. Des modèles dérivés du cadre des processus de Markov déterministes par morceaux peuvent constituer une alternative intéressante en permettant un compromis entre réalisme du modèle et facilité d'estimation. Dans le cadre d'étude décrit ci-dessus, l'utilisation de modèles "tout-terrain" a été privilégiée puisque les déterminants de propagation d'une espèce localement nouvelle dans un nouvel environnement sont généralement incertains. Cependant, les prédictions pouvant être tirées de ces modèles ne sont pas optimales puisqu'elles dépendent fortement des hypothèses sous-jacentes au modèle et qu'elles ne prennent pas en compte les incertitudes pouvant l'entourer. Ma deuxième ligne de recherche consiste à proposer une approche permettant de prendre en compte les incertitudes entourant chaque modèle. La technique que j'emploie est celle du Bayesian model-averaging. Cette technique consiste à combiner les prédictions des modèles en compétition d'une façon à obtenir une prédiction unifiée améliorée. Cette technique a souvent été utilisée en sciences environnementales. Toutefois, elle n'est pas répandue dans le domaine de l'épidémiologie. L'un des buts méthodologiques de la thèse est d'en évaluer l'intérêt pour l'épidémiologie prédictive.

Le cas d'étude de ma thèse est celui de la bactérie phytopathogène *Xylella fastidiosa* pour laquelle des données de surveillance spatio-temporelles et binaires post-introduction ont été collectées à partir d'un plan de surveillance intense qui a été mis en place par l'État suite à la première détection de cette bactérie en Corse en 2015. Ce pathogène de quarantaine, qui a significativement impacté la production d'olives en Italie et présente un risque de modification drastique de l'environnement du fait de sa capacité à atteindre de nombreuses espèces végétales, a le potentiel de causer en France une crise sanitaire majeure en santé végétale, à l'image de celle qu'elle cause depuis 2013 en Italie où les impacts socio-économiques sont conséquents.

Mots-clés: Approche Mecanistico-statistique, équations aux dérivées partielles, inférence Bayesienne, Bayesian model-averaging, épidémiologie prédictive, invasions biologiques, *Xylella fastidiosa*.

Acknowledgments

First and foremost, I would like to sincerely thank my advisors, Samuel Soubeyrand and Eric Parent. If I have been successful, it is only because I have been fortunate to be guided by Giants. Samuel, I will never thank you enough. Throughout these last years, you have always been available to discuss and provide guidance on all aspects of my PhD. You knew how to increase my motivation and push me beyond my limits. You have taught me how to find an "optimistic" way to make a way out of no way. Eric, thank you so much for initiating me to Bayesian statistics with the finest pedagogy and expertise. Your constant encouragement has incited me to give my best to repay the confidence you have shown in me.

I would also like to address a big thank you to all the members of my thesis jury, who have devoted their time to read and comment my manuscript. I hope that one day we will cross paths again. In particular, I would like to thank the reviewers, Profs. Anne Gégout-petit and Dr. Sophie Donnet for their insightful comments and remarks. A special thought goes for Profs. Florence Hubert who initiated me during my masters to scientific computation and have stood by me throughout my scientific journey in France, always wearing a beautiful smile on her face.

I am sincerely indebted to a host of friends and colleagues, without whom I could not have accomplished this work. My warm gratitude goes in essence to SuperLoic (and the cluster), without whom this thesis would not be the same. Collaborating with Olivier B. and Rachid during my PhD was a great pleasure. It is also thanks to the BioSP family that I have been in excellent working conditions: Edith, Davide, Raphael, Thomas, Etienne, Denis, Florent, Florian(s), Herves, Lionel, Julien, Emily, Amélie, Virgile, Jean-Loup, Jeff, Olivier M., Claude, Pascal, Jérôme, Lucie, Marie, Alex, Maria G., Amel, Xhelbrim, Marie-Eve, Melina, Emilie, etc.. I would also like to thank my other colleagues at MIA-AgroParisTech. A big big thanks to the people who have helped me getting through hard times: Maryam – Ms. for every problem a shopping center, Maria – Ms. for every problem a clumsy solution, Patrizia – Ms. for every problem a cake, and Marine – Ms. for every problem a sushi meal. I would also like to thank the Lebanese mafia in France: Tonia, Berna, Michel, Hussein, Tania, Sarkis, Daher, Hanine, Sara, Noura, Ali, Georges and all my friends in the Foyer Franco-Libanais Marseille. A heartfelt thanks to my besties: Yara and Carla, my brother from another mother: Lara, my `Photoshop` tutors: Sarah and Perla, and my loyal friends in Lebanon: Nathalie, Peter, Zeina, Ghiwa, Sandy, Philippe, Joelle, Anna, etc..

I want to thank my family for supporting me and being a constant source of love and encouragement throughout my life. I thank God for all his blessings and for having such a warm and supportive BIG family. To my parents, Jean and Marlene, who made a lot of sacrifices – I owe you everything and I hope that I have met your high expectations. To my sister and brother, Caren and Carl – being far away from home was not easy because of you but you were the source of my constant motivation, I love you to the moon and back. To my uncle, Pierre – I will never find the perfect words to describe how much I love you and how much I am thankful to all what you have done for me every minute of every day. To my aunts, Sousou and Rora – ‘Bhebkoun add l dene’. To all my uncles and aunts, cousins, and grandparents, thank you for believing in me. Sylvie, I did not forget to mention you above. But, I wanted to thank you in this section to tell you that I am blessed for having you in my life. You know that you will always have this very special place in my family and in my heart.

And finally, to my companion and backbone, Serge – looking back to the last years, I have no clue how I was supposed to get through everything without you. Thank you for all the words of encouragement and motivation and for picking me up every time I felt down. Through everything, you kept me laughing and you kept me inspired. I am so lucky to have you by my side. Till forever...

Table of Contents

List of Figures

List of Acronyms

AIC

Akaike's information criterion. 48

AMIS

Adaptive multiple importance sampling. 15, 18, 45, 46, 129, 131, 133, 138, 141–145, 148, 149

BFGS

Broyden-Fletcher-Goldfarb-Shanno. 42

BIC

Bayesian information criterion. 48, 142

BMA

Bayesian *model-averaging*. 16, 17, 49, 50, 142, 143, 145

DAG

Directed acyclic graph. 12, 138, 139

DIC

Deviance information criterion. 48, 142

EFSA

European Food Safety Authority. 150

IC

Predictive information criterion. 48, 142

IS

Importance sampling. 44, 45, 149

LAD

Least absolute deviations. 43

LS

Least-squares. 41, 43

MC

Markov Chain. 42, 44

MCMC

Markov Chains Monte Carlo. 41, 43, 45, 46, 140, 141, 149

ML

Maximum Likelihood. 41–43, 46, 141

ODE

Ordinary Differential Equation. 30, 33, 38

PACA

Provence-Alpes-Côtes d'Azur. 12, 23, 149, 150, 152

PCR

polymerase chain reaction. 128

PDE

Partial Differential Equation. 15–17, 33–35, 38, 41, 138, 140–142, 144–146

PDMP

Piecewise-Deterministic Markov Process. 12, 15, 38, 127, 138, 139, 144–146, 148

SDE

Stochastic Differential Equation. 37, 41, 145

WAIC

Widely Applicable Information Criterion. 49

Xf

Xylella fastidiosa. 11, 12, 16, 17, 23–28, 127–129, 131, 133, 137, 140, 142–145, 147, 149–153

1. Introduction

1.1. Aims and Motivation

Emergence of exogenous pathogens may induce severe sanitary and socio-economical crises. The cost for pathogen eradication or containment generally increases with the delay between the establishment of the pathogen in a new territory and its detection [Jones and Baker, 2004, Faria et al., 2014, Soubeyrand et al., 2018]. This further increases motivation for understanding the pathogen epidemiology, designing eradication or containment strategies, and assessing their efficiency. In particular, reconstructing the past dynamics of the pathogen [Boys et al., 2008, Roques et al., 2016, Soubeyrand and Roques, 2014] and predicting its future extent [Chapman et al., 2015, Peterson et al., 2003], are key steps to conduct such investigations.

In essence, the thesis research aims to provide a generic methodology (i.e., a method scalable to various invasive species) that improves the predictions of an invasive species dynamics for which no dedicated model is available and whose initial conditions (i.e., date and location of the introduction of invasive species) are unknown. In order to achieve this goal, we proceed in two complementary lines of research.

The first one is to propose a model&data-based inference method of a pathogen dynamics, in the framework of the so-called mechanistic-statistical approach [Berliner, 2003, Roques et al., 2011, Soubeyrand et al., 2009a,b, Soubeyrand and Roques, 2014, Wikle, 2003a,b]. This method allows us to jointly estimate the introduction point (date and location of the pathogen arrival) and other parameters of the dynamics related to diffusion, reproduction, and death. It is grounded on (i) a Partial Differential Equation (PDE) which have been extensively used for modeling, in a phenomenological and concise way, spatio-temporal population dynamics [Skellam, 1951, Okubo, 1980, Shigesada et al., 1995, Gatenby and Gawlinski, 1996, Shigesada and Kawasaki, 1997b, Turchin, 1998, Okubo and Levin, 2002], (ii) a stochastic model that represents the observation process and describes the link between data and the mechanistic representation of the dynamics, and (iii) a statistical Bayesian inference procedure, the Adaptive multiple importance sampling (AMIS)[Cornuet et al., 2012], for estimating model parameters. In order to relax hypotheses made on the dynamics, the deterministic phenomenological model could be replaced by a stochastic population-dynamic model, from individual-based models [Renshaw, 1993, Kareiva and Shigesada, 1983] to aggregated models [Soubeyrand et al., 2009b]. However, such models induce extra difficulties in estimation because of the additional parameters and latent variables. Models issued from the framework of PDMP model could be an appealing and interesting alternative to balance the trade-off between model realism and estimation easiness. In the framework presented above, preference has been given to the use of generic spatio-temporal propagation models since the detailed processes underlying the spread of an alien species in a new environment are usually unknown. However, predictions that can be drawn from those models are not optimal because they are affected by the rather strong underlying assumptions, and do not take into account the uncertainty about the model form. This limitation can be circumvented by considering a family of candidate models and applying a model selection strategy [Burnham et al., 1995] or a model aggregation strategy [Hoeting et al., 1999].

The second research line is to propose an approach that allows taking into account the uncertainty about the model form. As part of the aggregation strategy, the Bayesian *model-averaging* (BMA) approach has been proposed by Leamer [1978], to account for parameter and model uncertainties [Hoeting et al., 1999]. This approach consists of averaging over all candidate models in a Bayesian way for weighting models [Raftery, 1996, Volinsky et al., 1997], combining multiple predictions and combining estimations to infer shared parameters [Roberts, 1965, Madigan and Raftery, 1994, Wintle et al., 2003]. Despite ample literature on BMA and its usefulness [Viallefont et al., 2001, Raftery et al., 2010, Parkinson and Liddle, 2013, Eicher et al., 2011, Sidman et al., 2008, Yeung et al., 2005, Oehler et al., 2009, Yin and Yuan, 2009, Boone et al., 2005, 2008, Wintle et al., 2003, Raftery et al., 2005], it has only been marginally applied in the context of predictive epidemiology.

The case study of the thesis is the dynamics of the phytopathogenic bacterium Xf in Corsica, France. For this real case study, abundant spatio-temporal and binary post-introduction surveillance data have been collected from an intensive surveillance plan implemented by governmental agencies after the first *in situ* detection of Xf in 2015 in Corsica. This quarantine pathogen in Europe has significantly impacted olive production in Puglia, Italy, and presents a drastic risk of environmental degradation due to its ability to reach a large variety of plant species. It is currently present in a large part of Corsica island and more marginally in Southeastern mainland France [Denancé et al., 2017a, Soubeyrand et al., 2018, Martinetti and Soubeyrand, 2019]. It is susceptible to cause a major sanitary crisis in France, as the one caused in Italy since 2013 where the socio-economical impacts have been considerable due to the grabbing-up and death of a significant proportion of olive trees. Moreover, in summer 2019, the first two cases of olive trees infected by Xf in France were detected in Menton and Antibes, which lifted risk evaluation related to this pathogen (https://agriculture.gouv.fr/la-contamination-par-xylella-fastidiosa-de-2-oliviers-confirmee-en-paca).

1.2. Research questions

The following summarizes the methodological questions addressed in this thesis:

- ❖ How to infer initial conditions of a PDE modeling the propagation of an invasive species, where the initial conditions represent the introduction of the species in question?
- ❖ How to extend this modeling and inferring framework, when various competing mathematical representations of an invasive species dynamics are considered?
- ❖ How to extend this modeling and inferring framework, when multiple introductions potentially occur?

These methodological questions allow me to answer the following epidemiological questions:

- When and where the strains of Xf that triggered the epidemic observed since 2015 in Corsica were introduced in this region?
- Which measure of the propagation capacity of Xf can be obtained from spatio-temporal surveillance data?
- What is the impact of winter temperatures on the dynamics of Xf in Corsica?
- What will be the spatial distribution of Xf in the future?

1.3. Chapters Contents and Manuscript Organization

This manuscript is organized as follows:

Chapter 2 gives the state of the art, in which mathematical tools have been outlined in a manner to justify our choices for meeting epidemiological challenges. First, the epidemiological context is detailed with a particular focus on the case study Xf and the existing data in France, in particular in Corsica. Then, the mathematical context is reviewed under three major headlines: modeling in population dynamics, inferring population dynamics given a mathematical model, and accounting for model uncertainties using model selection and model-averaging techniques.

Chapter 3 develops the first line of research. A mechanistic–statistical approach is proposed to tackle the problem of recovering the location and time of the introduction of an exogenous pathogen in a territory based on post-introduction data. In this chapter, we assume that only one introduction spot triggered the pathogen propagation and eventual subsequent introductions are considered to have negligible effects on the dynamics. In this framework, we adapt the process of statistical analysis presented by McCullagh [2019] (see Figure 1.1). The adaptation of this process is mapped out in Figure 1.2. Thus, a class of models is carefully considered, paying attention to the type and structure of data. Then, models parameters are estimated using the adaptive multiple importance sampling algorithm, and model choice criteria are used to select information from the model that is best linked to data. Finally, we conduct goodness-of-fit tests, as model checking to check the adequacy between the model and observed data.

Chapter 4 extend the process of statistical analysis presented in Figure 1.1 to raise issues from the second research line (see Figure 1.2). The application of BMA is investigated in the context of pathogen-dynamics prediction using PDEs-based models. The models are grounded on a family of reaction-diffusion equations (including those presented in Chapter 3) with different spatially heterogeneous diffusion and reproduction terms. Our aim is to compute, from post-introduction data, the BMA posterior distribution of a certain quantity of interest, such as the introduction time or location of the pathogen or its future spatial extent. This approach is tested on simulated data and then applied to make predictions concerning the dynamics of the phytopathogenic bacterium Xf in Southern Corsica, France.

Chapter 5 provides a forward-looking statement of the first research line developed in Chapter 3. The models used in Chapter 3 are generally not adapted to describe the dynamics of populations that expand their range not only by neighborhood dispersal but also because of new introductions in disease-free areas and by long-distance dispersal, which can correspond to rare but significant events. Chapter 5 explores these features by incorporating into the framework proposed in Chapter 3, the possibility of considering multiple introductions. This should allow considering data at larger spatial scales, for which the hypothesis of a single introduction is generally not adequate. Model parameters are estimated as in Chapter 3, using the AMIS algorithm.

In Chapter 6, summary, discussions and perspectives are made to conclude and propose improvements for furthering the work.

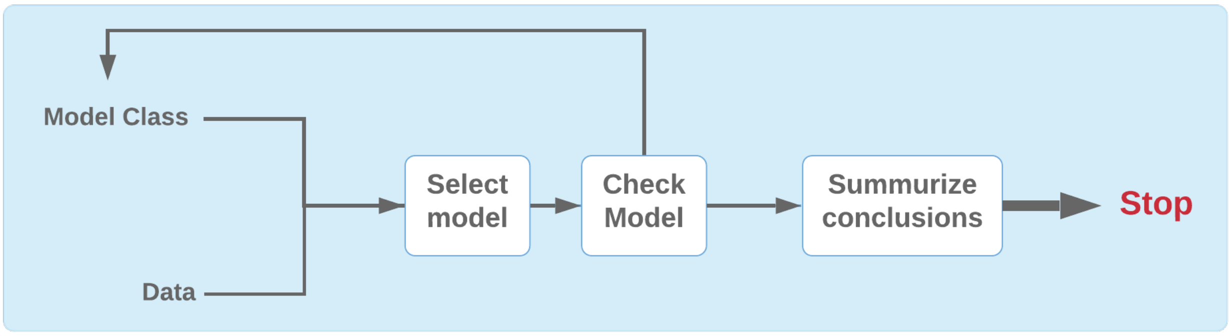


Figure 1.1.: The process of statistical analysis as presented by [McCullagh, 2019] (Chapter 12, page 392).

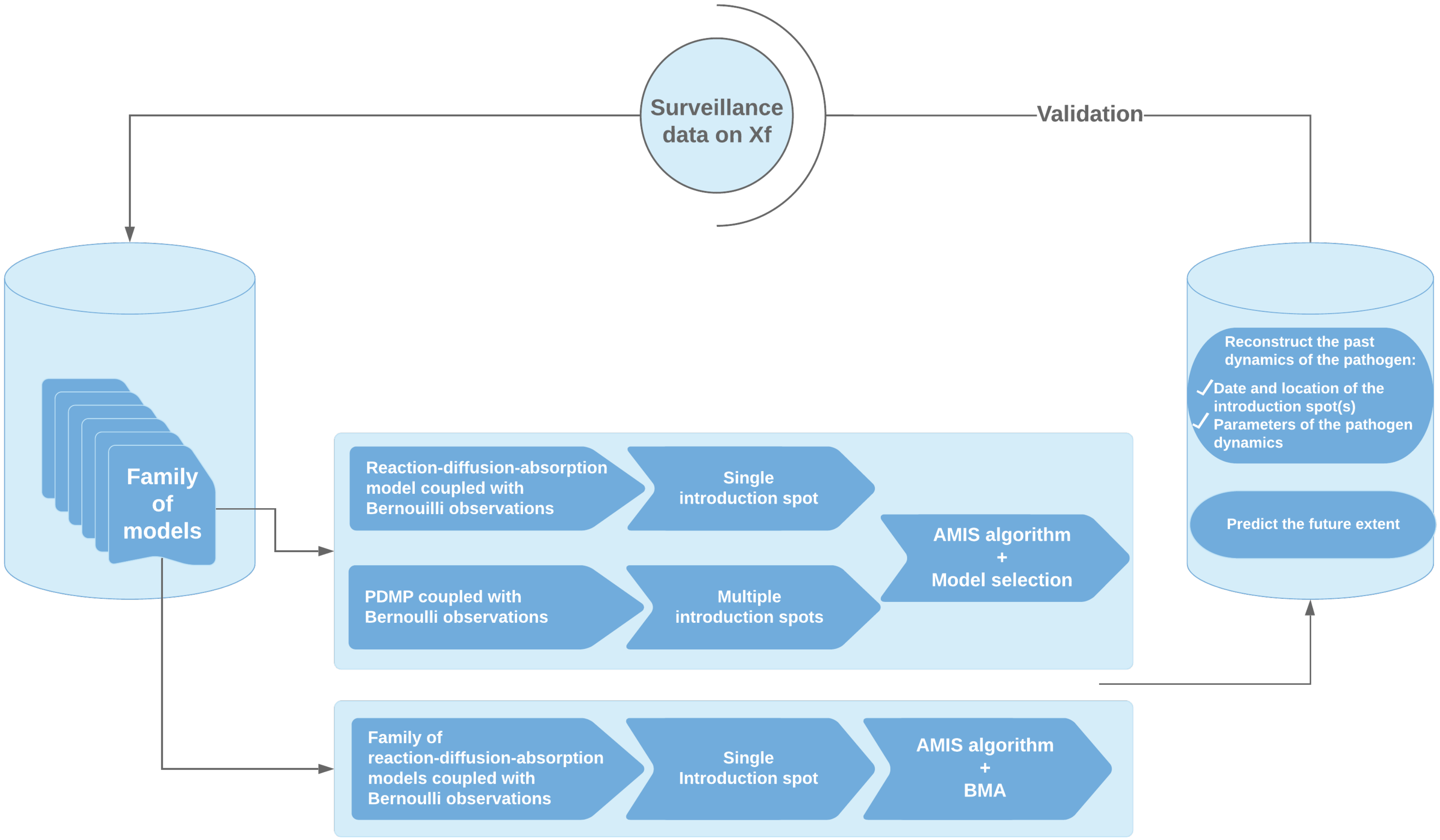

Surveillance data on Xf
Validation
Family of models
Reaction-diffusion-absorption model coupled with Bernouilli observations
Single introduction spot
PDMP coupled with Bernoulli observations
Multiple introduction spots
AMIS algorithm + Model selection
Family of reaction-diffusion-absorption models coupled with Bernoulli observations
Single Introduction spot
AMIS algorithm + BMA
Reconstruct the past dynamics of the pathogen:
✓Date and location of the introduction spot(s)
✓Parameters of the pathogen dynamics
Predict the future extent

Figure 1.2.: Summary of the thesis research lines and their respective components.

2. State of the art: Mathematical tools for meeting epidemiological challenges

This chapter introduces the epidemiological context and outlines mathematical tools needed for the entire manuscript. This chapter provides a bibliographical review described in a manner that justifies our choices for meeting the epidemiological context.

Table of contents

2.1. Epidemiological context

2.1.1. Brief Review on Epidemiology of Emerging Infectious Diseases

Epidemiology was first defined as the science of disease in the population [Plank, 1963]. Since then, it has regularly updated giving the evolution in the field. Nowadays, epidemiology is defined as "the study of distribution and determinants of health-related states or events (including diseases), and the application of this study to the control of diseases and other health problems" [Friis and Sellers, 2004, Friis, 2017].

Four major types of epidemiological investigations can be distinguished: (i) *descriptive* epidemiology which aims to identify the spatial and/or temporal distribution of a disease in a population, (ii) *analytical* epidemiology with a view to study its determinants, (iii) *evaluative* epidemiology that qualifies the impact and evaluate the results obtained from health and disease prevention programs and, (iv) *predictive* epidemiology which propose handy mathematical tools to describe, infer and forecast disease propagation.

Mainly, we focus in this thesis on plant epidemiology motivated by the severe damage that can be triggered by the introduction of plant pathogens into a natural ecosystem [Pimentel et al., 2005]. As for humans and animals, occurrence of a plant disease requires interaction among three essential components: a susceptible host plant, a virulent pest (e.g., virus, bacteria, fungi and parasitic plants) and a favorable environment (e.g., related with temperature, moisture, presence of a vector and wildlife) [Stevens, 1960]. However, the development of the disease is largely affected by time factors [Agrios, 2005] as per example the season of the year, the duration and frequency of favorable temperature, and the appearance of vectors.

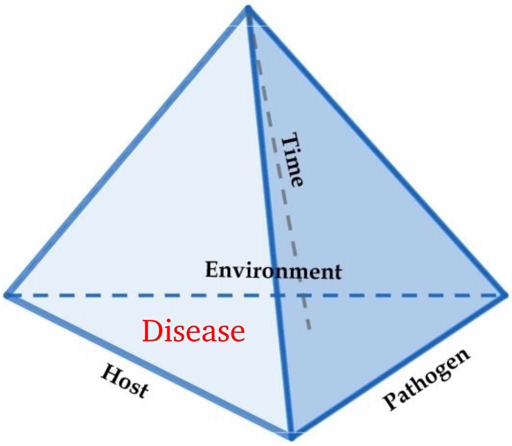


Figure 2.1.: Interrelations of the four components: host, pathogen, environment and time.

Thus, for the pathogen to propagate, there must exit an infected host plant and the pathogen should be able to enter a susceptible plant and multiply in the newly infected host plant which is conditioned by environmental and time factors (see Figure 2.1). The transmission of pathogens may be either direct (e.g., carried externally or internally on the seed or planting material like cuttings and sets) or indirect (e.g., via wind dispersal, water dissemination, animals, insects, and human behavior).

Assessment of epidemiological parameters such as initial growth, incidence and prevalence is required for disease characterization purposes and will help in predicting disease dynamics. In what follows, we recall the definition of the main quantities of interest.

The incidence rate represents the rate of occurrence of new cases per individual from population at risk in a given period [Rothman et al., 2008]. It is one of the essential criteria for determining the prevalence and speed of manifestation of a disease:

$$\text{Incidence Rate} = \frac{\text{Number of new cases found at a specified time}}{\text{Total population size}}.$$

The prevalence rate is the proportion of infected plants in a population that have a disease at a defined time or over a specified period:

$$\text{Prevalence Rate} = \frac{\text{Number of existing cases of disease at a specified time}}{\text{Total population size}}.$$

Conversely to incidence, prevalence includes all existing cases in the population at the specified time, whereas incidence is computed based on new cases only.

The dispersal rate describes the expected proportion of disease disseminating agents to leave an area. The so-called dispersal function typically gives the probability distribution of the distance traveled by the disease.

The latent period is the delay between exposure to a pathogen and the appearance of disease symptoms [Porta, 2008].

The infectious period is the period during which a pathogen produces dispersal units that come into contact with a host [Frantzen, 2007].

The basic reproduction number denoted by R_0 is defined as the expected number of secondary cases produced by a single infection in a susceptible population never infected before. In other words, R_0 is the initial growth rate when the population is considered on a generation basis [Diekmann and Heesterbeek, 2000]. This dimensionless number allows the characterization of diseases according to their potential to cause epidemics:

- If $R_0 > 1$, the disease is intensely transmitted. The higher the value of R_0, the faster the rise of the incidence rate.
- If $R_0 < 1$, the number of infected plants is expected to decline after the introduction.

The mortality rate (or death rate) is the measure of the frequency of occurrence of death in a specific population during a given interval of time [Porta, 2008]. It is commonly expressed in terms of per 1000 individuals:

$$\text{Mortality Rate} = \frac{\text{Number of deaths during a specified time}}{\text{Number of individuals at risk of dying during the period}} \times 1000.$$

The fatality rate is the proportion of cases of a specified condition which are fatal within a specified time [Porta, 2008]. It is commonly expressed as a percentage:

$$\text{Fatality Rate} = \frac{\text{Number of deaths during a specified time}}{\text{Number of individuals with the disease during the period}} \times 100.$$

2.1.2. *Xylella fastidiosa*

2.1.2.1. Definition

Xf is a phytopathogenic xylem-limited bacterium. Literally, the genus '*Xylella*' is taken from xylem and refers to the fact that this bacterium is limited to the vascular tissues ensuring the flow of crude sap in the plant; the specific epithet '*fastidiosa*' means highly critical and refers to the nutritional fastidiousness of the organism, particularly on primary isolation [Wells et al., 1987].

2.1.2.2. Status in France

Xf has probable origins from countries of the American continent [Pierce, 1892]. In fact, available genetic data indicate that three of five subspecies of Xf have origins in different geographic areas: fastidiosa subsp.fastidiosa originated in Central America, subsp.multiplex in North America, and subsp.pauca in South America [Sicard et al., 2018]. This bacterium has been reported for the first time in Europe in 2013, on olive trees in Italy [Saponari et al., 2013]. The situation in Italy evolved rapidly into an important outbreak significantly impacting olive culture and causing major sanitary crisis (massive grubbing-up of olive trees in Apulia). Subsequently, Xf has been of concern in many European countries, including France that reported the first detection in July 2015, on an ornemental plant, *Polygala myrtifolia*, in South Corsica, and is currently mostly present in this island and in the PACA region [Denancé et al., 2017b, Soubeyrand et al., 2018]. The observed disease incidence rate (%) computed monthly for all hosts, since the first detection *in situ* in France is shown in Figure 2.2. The decrease in the observed disease incidence rate does not mean that there is an actual decrease in the disease incidence, because detected positive cases are destroyed and therefore subtracted from the overall disease incidence [Soubeyrand et al., 2018].

2.1.2.3. Transmission Modes

Xf is transmitted from one plant to another one by various xylem sap-feeding insects [Frazier, 1965] as Froghoppers, Leafhoppers, Spittlebugs, Homoptera, and Auchenorrhyncha. The Forghopper *Philaenus spumarius* is currently the only confirmed vector of Xf in Europe [Maria Saponari, 2014, Cornara et al., 2017], and in experimental conditions, *Neophilaenus campestris* and *Philaenus italosignus* confirmed their physiological ability to carry the bacterium out [Cavalieri et al., 2018]. However, all xylem sap-feeding insects remain susceptible vectors of the disease [Purcell, 1990].

Besides, plantation (using infected planting material like cuttings and sets), multiplication and movements of infected seedlings may constitute an important risk factor for the diffusion of the bacterium. Figure 2.3 describes the direct and indirect transmissions of Xf and influencing factors of propagation.

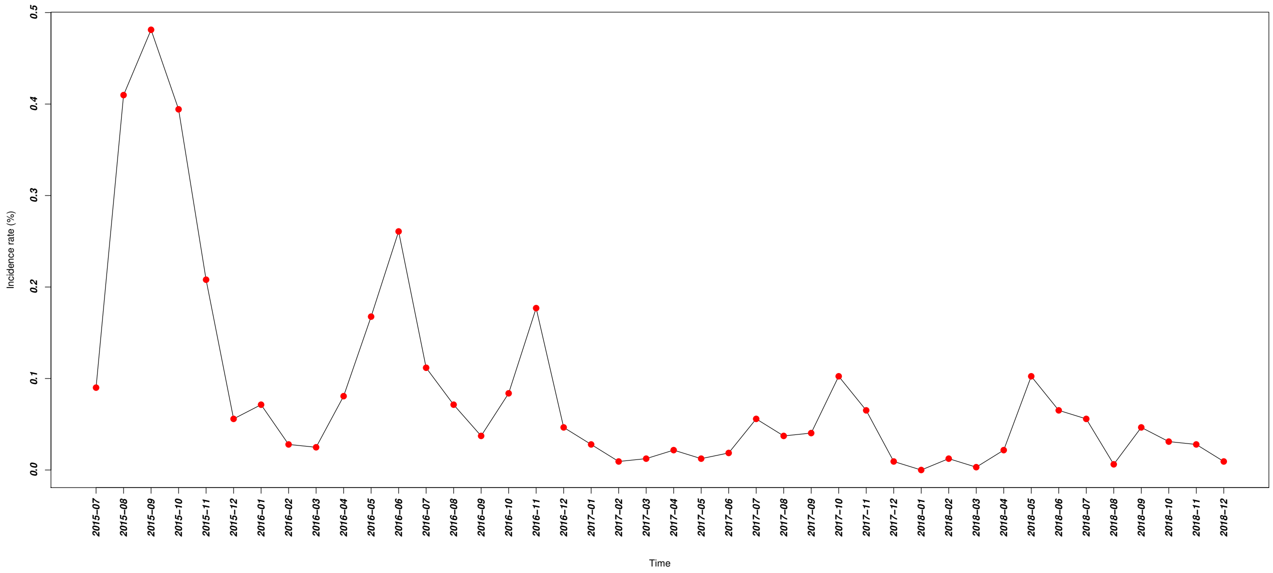


Figure 2.2.: Observed incidence rate (%) of Xf computed for all hosts, on a monthly basis since the first *in situ* detection in France.

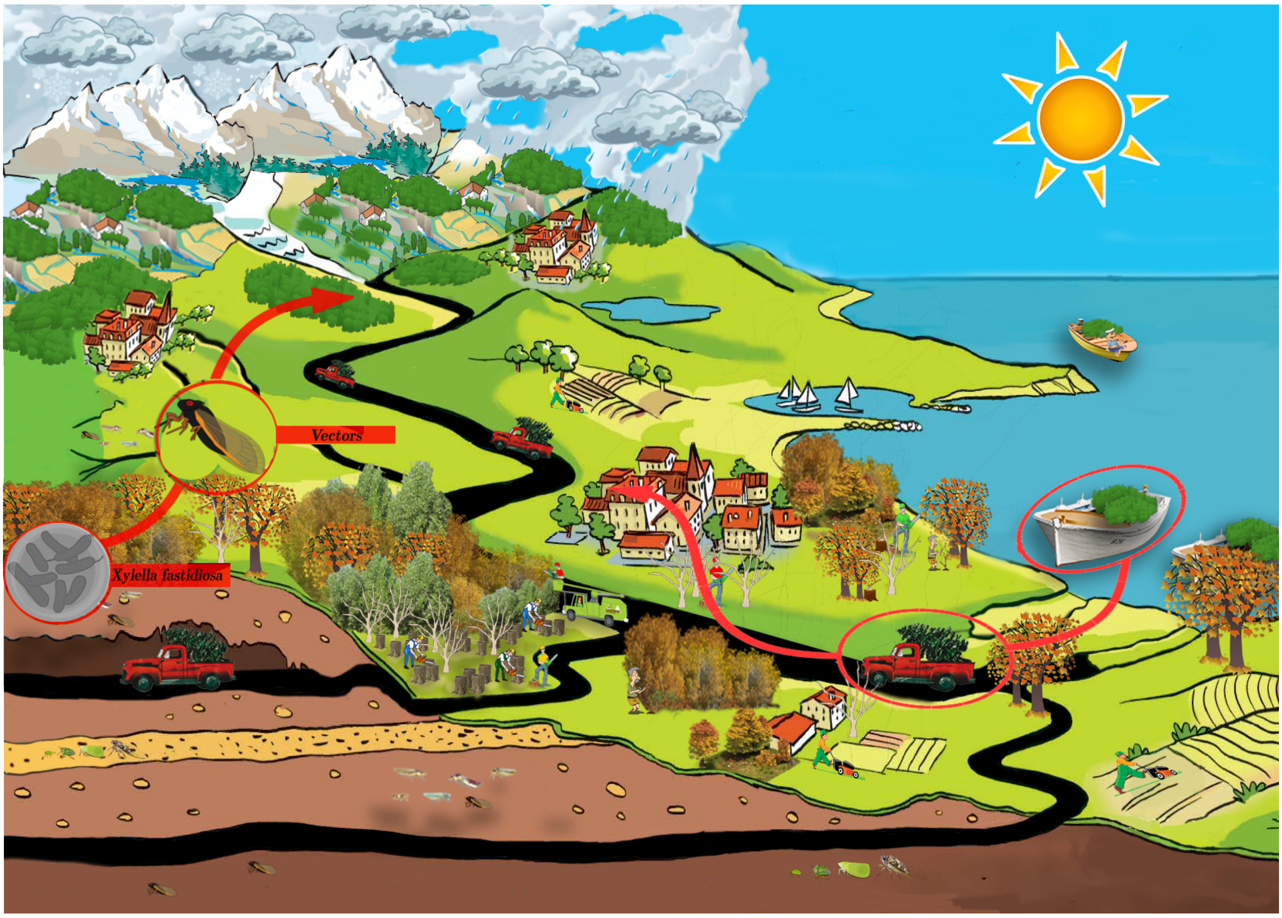


Figure 2.3.: Description of the transmission of Xf from plant to plant by xylem sap-feeding insects and by movements of infected plants in different environmental situations (high/low latitudes, high/low&winter temperatures).

2.1.2.4. Factors of propagation

Many influencing factors affect the propagation of Xf such as the phenology of insect vectors [Purcell, 1981, Almeida et al., 2005] and increasing globalization trade [Hulme, 2009], which levels up the risk of infected plant transportation. As for other bacteria, the propagation of Xf is also affected by many environmental factors such as latitude [Costello et al., 2017], inoculation date [Davis et al., 1980] and temperature [Daugherty et al., 2009, 2017]. Winter temperature has been inferred as a chief environmental factor governing the dynamics of Xf and the level of disease severity caused by Xf [Costello et al., 2017, Feil et al., 2003, Henneberger, 2003, Purcell, 1977, Purcell et al., 1980]. For instance, isolines for the average minimum daily temperature in January have been shown to be consistent with regions in the United States that are exposed to different levels of severity of Pierce's disease of grape caused by Xf [Anas et al., 2008, Feil and Purcell, 2001]. Most of the analyses on climatic and environmental factors were performed for strains present in the Americas, mostly on grapevines. However, climatic limits and severity of impacts largely depend on the host-pathogen-ecosystem combination [EFSA Panel on Plant Health (PLH) et al., 2019]. Recent studies on climatic suitability have been carried out in European conditions and have corroborated the influence of temperature among other climatic factors [Godefroid et al., 2018, Martinetti and Soubeyrand, 2019].

2.1.2.5. Hosts & Impacts

Xf is considered as one of the most dangerous plant bacteria in the world due to significant potential agricultural and socio-economical impacts likely arising with its emergence because there is still no treatment able to eliminate the bacteria from the plant except for grubbing up and destroying infected plants [EFSA Panel on Plant Health (PLH) et al., 2019]. When in infected plants, Xf forms bacterial aggregates in the vascular tissues of the xylem preventing the flow of crude sap, which results in the dryness and eventually the death of the infected plant. Over 360 plant species have been reported as susceptible to be attacked by the bacterium. The list of hosts in Europe is regularly updated (see Commission database of host plants found to be susceptible to Xf in the Union territory). This wide range includes wild plants, cultivated plants, forest species and ornamental plants (e.g., olive trees, vines, fruit trees, lavender, cherry, oaks). However, groups identified within each subspecies target more specific plants. So far, four subspecies of Xf have been frequently observed in Europe: *multiplex, pauca, sandyi* and *fastidiosa* [EFSA Panel on Plant Health (PLH) et al., 2019]. Until now, 59 plant species have been reported with Xf in France (see Figure 2.4) for the *multiplex* (Sequence type: ST6, ST7) and *pauca* (Sequence type: ST53) subspecies [EFSA Panel on Plant Health (PLH) et al., 2019] including economically valuable plants.

2.1.2.6. Epidemio-surveillance & control strategies

To avoid a potential socio-economical crisis in France similar to the one happening in Italy since 2013, the French General Directorate of Food (DGAL) has implemented enhanced control and surveillance measures after the first *in situ* detection of Xf in Corsica. These measures have been modified several times based on updated knowledge on Xf and according to European directives. Mainly, the strategy focuses on the following actions:

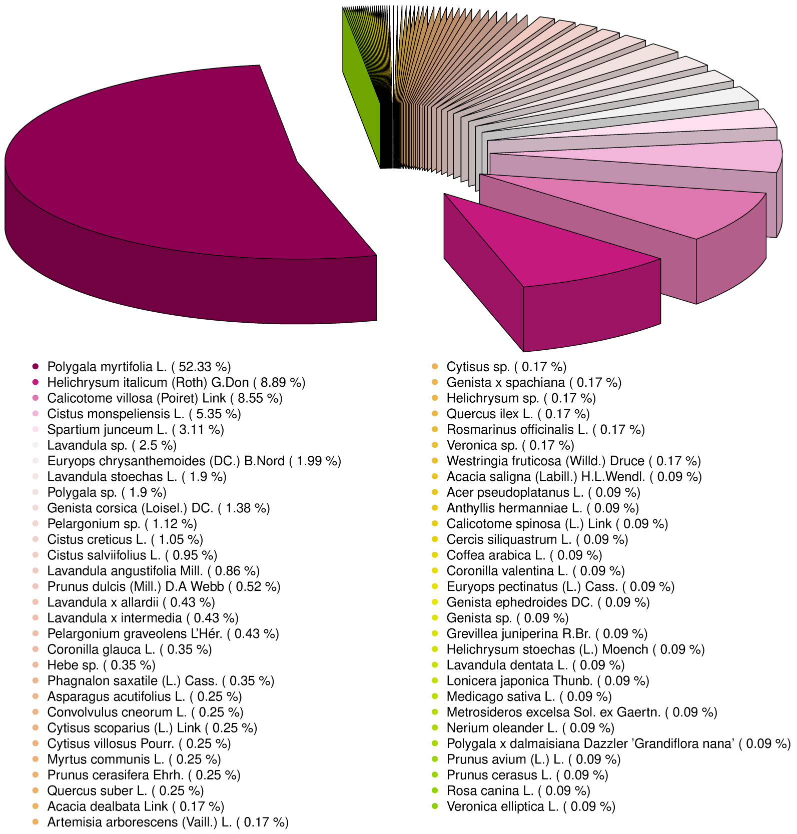


Figure 2.4.: Infection prevalence according to plant species identified in France until April 2019.

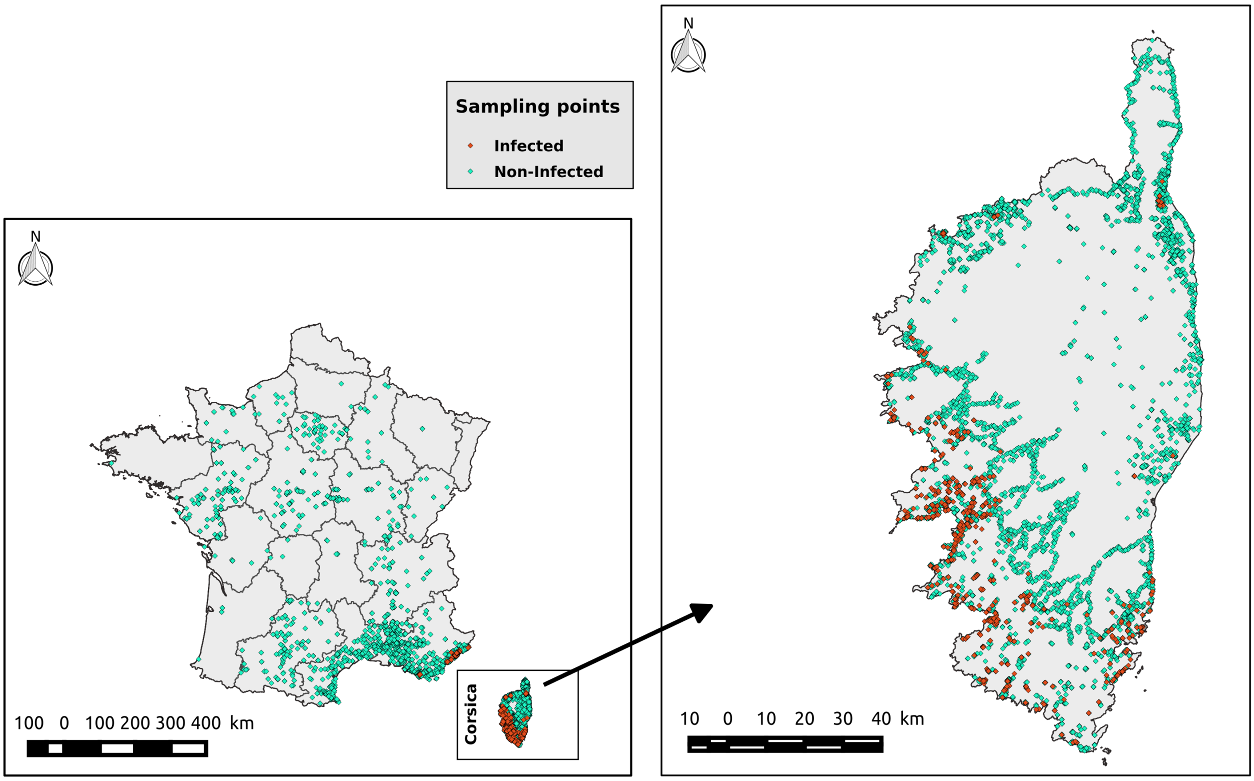


Figure 2.5.: Locations of plants, sampled from July 2015 to April 2019, that have been detected as positive (red diamonds) or negative (cyan diamonds) to Xf in France (on the left), and zoom in on Corsica (on the right).

The enhanced surveillance of the national territory which generated a data set consisting of a spatio-temporal point pattern (i.e., the locations and dates of plant samples) marked by a binary variable indicating the result of the diagnostic test (i.e., indicating if the plant sample is positive or negative to Xf). Until April 2019, approximately 32200 plants were sampled, among which 1600 have been diagnosed as infected [with a real-time polymerase chain reaction (PCR) technique; Denancé et al., 2017b]. Available data for each sampled plant are its spatial coordinates, its sampling date (which is unique) and its health status at the sampling date. Coordinates and health statuses at the sampling times are shown in Figure 2.5.

The eradication of detected outbreaks by creating a buffer and demarcated areas around the infected plant. Figure 2.6 shows:

- the buffer area, within which all infected and host plants are cut off and destroyed after they have been treated with insecticide in order to prevent the dissemination of vectors.
- the demarcated area, in which surveillance is implemented with the organisation of inspections and sampling in order to assess the disease-free status of the area. The control of the movement of host plants outside demarcated zones is prevented to protect disease-free areas.

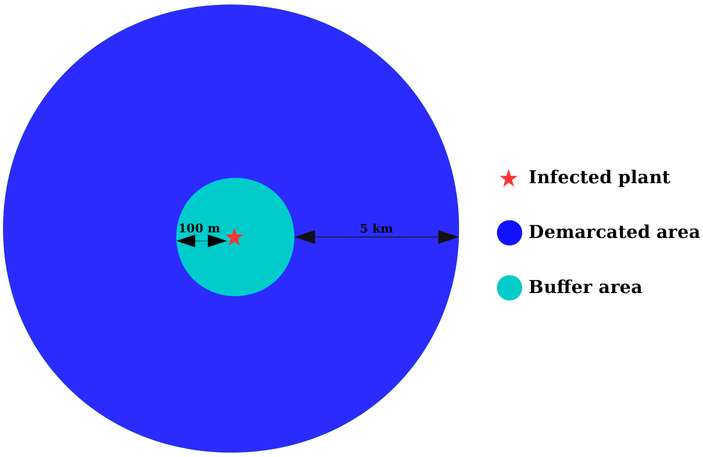


Figure 2.6.: Representation of buffer and demarcated areas around a newly detected infected plant.

In Corsica, the extended presence of the bacteria (see Figure 2.5) prompted the French authorities to request the shift from an eradication strategy to a containment strategy, since the eradication of Xf from the island was considered as impossible. The main difference between both strategies is that only the infected plants are uprooted.

Beyond this surveillance and control strategy, France and other European countries promoted the understanding of epidemiological characteristics of Xf from several perspectives. This understanding has been particularly developed via modeling works focused on the European situation [Strona et al., 2017, White et al., 2017, Bosso et al., 2016, Godefroid et al., 2018, Soubeyrand et al., 2018, Martinetti and Soubeyrand, 2019, Kyrkou et al., 2018]. Providing more insights on the spread of this pathogen is specifically judicious to optimize surveillance and control strategies because the price to eradicate the pathogen increases with the duration since the introduction and establishment of the pathogen [Soubeyrand et al., 2018]. Denancé et al. [2017a] assessed the introduction of the two sequence types found in Corsica around 1965 and 1980, respectively, using a phylogenetic approach. Likewise, Soubeyrand et al. [2018] dated the introduction around 1985 (95%-posterior interval: $[1978, 1993]$) with a statistical analysis of temporal data. One of the interesting questions in this context is how the disease was introduced in France, in particular in Corsica, where the situation is most critical. Dating and localizing the introduction can bring an appropriate answer to this question. However, inferring the introduction (location and date) from post-introduction data (i.e., data collected over a temporal window covering a period after the introduction time) providing the sanitary statuses of sampled plants, requires the inference of the spatio-temporal disease dynamics, and *vice versa*, because this dynamics links the introduction and the observations. Thus, reconstructing the past dynamics of Xf (i.e., the spatio-temporal introduction jointly with other epidemiological parameters related to growth, dispersal, and death that govern the post-introduction dynamics) can help understand its origins and therefore ameliorate the ability to inform about its future spatial extent. For addressing the above epidemiological challenges, mathematical tools (see Figure 2.7) can contribute to the acquisition of knowledge about Xf and support the decision process of decision makers (e.g., the French Ministry of Agriculture).

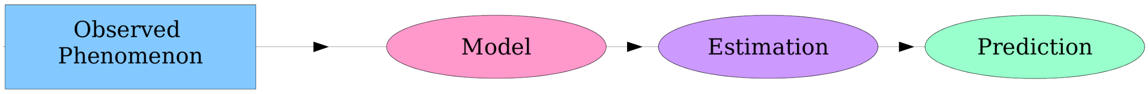


Figure 2.7.: Pipeline from the observation of the phenomenon of interest to prediction.

2.2. Mathematical context

The spread of invasive alien species to new areas has long been an important research topic not only for biologists but also for mathematicians because of their impact on the environment, indigenous species, and health of humans, animals, and plants [Andow et al., 1990, 1993, Baker, 1991, Hengeveld, 1989, Kermack and McKendrick, 1927, Richardson and Bond, 1991, Simberloff, 1989, Anderson et al., 1996, Shigesada and Kawasaki, 1997b, Weinberger, 1978]. An early work in predictive epidemiology dates back to Bernoulli [1760] after the inoculation of the smallpox epidemic. However, predictive epidemiology has started on a fast track through the pioneering works of Ross [1911] on malaria and Hamer [1906] on measles. Since then, models have evolved to fill in the gaps in the experimentation fields. Nowadays, mathematical modeling has become an inseparable part of the experimental activity, especially to help in getting a better comprehension of the emerging dynamics. In particular, extensive researches have been conducted in the intents of reconstructing the past dynamics [Boys et al., 2008, Roques et al., 2016, Soubeyrand and Roques, 2014] of alien species and predicting their future spatial extents [Chapman et al., 2015, Peterson et al., 2003]. In this section, we review mathematical models in population dynamics, as well as inference and prediction approaches that can be developed in the so-called mechanistic-statistical framework used to describe, infer and predict physical, ecological, and epidemiological processes [Berliner, 2003, Lanzarone et al., 2017, Roques et al., 2011, Soubeyrand et al., 2009a,b, Wikle, 2003a,b]. This framework is hinged on a mechanistic model for the dynamics of interest, a probabilistic model for the observation process, and a statistical procedure for estimating parameters and predicting the dynamics.

2.2.1. Population Dynamics Modeling in Mathematical Epidemiology

Mathematical models for population dynamics are based on diverse mathematical tools adapted to different resolutions at which the population dynamics are considered (e.g., individuals, groups, presence in quadrats, and numbers of individuals in districts), and different levels of perceptions (e.g., the population itself, its averaged characteristics, or more generally aggregated functions of the population patterns). Mathematical models are either deterministic or stochastic. Deterministic models perform the same way for a given set of parameters and initial conditions. Concise deterministic models classically proposed for population dynamics only incorporate the main epidemiological and environmental drivers of the dynamics. These drivers and secondary drivers can be implicitly handled by replacing the deterministic model by a stochastic version, incorporating some inherent randomness, that would contribute to the deduction of flexible realizations. For instance, in the comparison of deterministic and stochastic models for circadian oscillations, [Gonze et al., 2002] shows that, in the presence of noise in a small population, stochastic simulations are needed to get more realistic realizations. In what follows, we will briefly review some modeling approaches, from compartmental models to aggregate models, that are widely used to offer insights into the mechanisms of spatio-temporal dynamics of populations. We mainly focus on deterministic models, because in contrast to stochastic modeling approaches, they provide a phenomenological and concise description of the population dynamics, and can be fitted to data in a reasonable time span. The advantage of this approach is that it can be rapidly applied for endorsing a fast reaction after the detection of a new invasive species.

2.2.1.1. Deterministic Compartmental Models

Compartmental models arose in the early 20th century with the pioneering works of public health doctors [Ross, 1911, Hamer, 1906, Kermack and McKendrick, 1991] for describing the dynamics of infectious diseases. In these models, the population is partitioned into a small number of different compartments connected by a flow of individuals. Each compartment contains individuals that have identical statuses with respect to the disease in question.

Directed flow graphs are usually used to represent these models (see Figures 2.8, 2.9, 2.10, and 2.11). The nodes represent the compartments and the arrows are weighted by transmission functions. Typically, compartmental models are built with Ordinary Differential Equation (ODE). However, one can also build these models in a stochastic framework [Andersson and Britton, 2012]. In what follows, we revisit some standard compartmental models in the deterministic framework of ODE.

2.2.1.1.1. Standard SI Deterministic Model

Hamer [1906] built the simplest epidemiological model composed by two mutually exclusive compartments: the susceptible (S), which is the class of individuals who are susceptible to infection, and the infected (I), which consists of individuals whose the level of parasite is sufficiently large, and who have a potential in transmitting the infection to other susceptible individuals (see Figure 2.8). The propagation of the infection starts after a susceptible individual comes into positive direct contact with an infected and infectious individual. Therefore, the more susceptibles in S and infectious in I, the more there is newly-infected cases.

Hamer's model suggests that once infected, an individual belonging to the compartment S becomes infected/infectious and stays permanently in the compartment I. This model is grounded on the following assumptions:

($\mathcal{H}1$) At time t, compartments S and I contain respectively $S(t)$ and $I(t)$ individuals;

($\mathcal{H}2$) The population is closed, i.e., the population size $N = S(t) + I(t)$ for each $t \in \mathbb{R}^+$ is constant;

($\mathcal{H}3$) $S(0) = N - 1$ which means that at time $t = 0$ all the population is susceptible except the first infected individual;

($\mathcal{H}4$) β is the transmission intensity of the disease per unit of time;

($\mathcal{H}5$) Demographic factors are excluded.

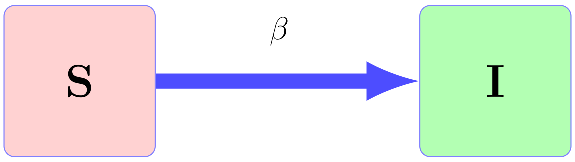


Figure 2.8.: Standard SI model

Having compartmentalized the population, Hamer's model is described by the following set of differential equations that specify how the sizes of compartments change over time:

$$\begin{cases} \dfrac{dS}{dt}(t) = -f(S(t), I(t)), \quad t \in \mathbb{R}^+, \\ \\ \dfrac{dI}{dt}(t) = f(S(t), I(t)), \quad t \in \mathbb{R}^+, \end{cases} \tag{2.1}$$

where $f(S, I)$ is the disease incidence over an infinitesimal time period dt. For instance, we can assume that $f(S, I) = \beta SI$, β being a positive constant representing contact rate. Thus, (2.1) is a differential system of the form:

$$\begin{cases} \dfrac{dS}{dt}(t) = -\beta S(t)I(t), \quad t \in \mathbb{R}^+, \\ \\ \dfrac{dI}{dt}(t) = \beta S(t)I(t), \quad t \in \mathbb{R}^+. \end{cases} \tag{2.2}$$

▶ **Indication for Solving System (2.2)**
Knowing that $N = S(t) + I(t)$ for $t \in \mathbb{R}^+$, we obtain the following logistical system:

$$\begin{cases} \dfrac{dS}{dt}(t) = -\beta S(t)(N - S(t)), \, t \in \mathbb{R}^+, \\ \\ \dfrac{dI}{dt}(t) = \beta(N - I(t))I(t), \, t \in \mathbb{R}^+. \end{cases} \tag{2.3}$$

System (2.2) can now be easily solved. An analytical solution can be obtained by integrating both of the system's equations.

2.2.1.1.2. Standard SIR Deterministic Model

The SI model is considered to be simple. Most of all, the SI model is mostly adapted to incurable diseases. However, for many common diseases as the flu, the carrier of the disease can recover. To take into account those individuals, Kermack and McKendrick [1927] proposed to add the compartment R to the SI model (see Figure 2.9). The model is now based on the following assumptions:

($\mathcal{H}1$) At time t, compartment R contains $R(t)$ individuals;

($\mathcal{H}2$) The population is closed, i.e., the population size $N = S(t) + I(t) + R(t)$ for each $t \in \mathbb{R}^+$ is constant;

($\mathcal{H}3$) $S(0) = N - 1$ and $R(0) = 0$, which means that at time $t = 0$ all the population is susceptible except the first infected individual;

($\mathcal{H}4$) β is the transmission intensity of the disease per unit of time;

($\mathcal{H}5$) γ is the recovery rate per unit of time;

($\mathcal{H}6$) Demographic factors are excluded.

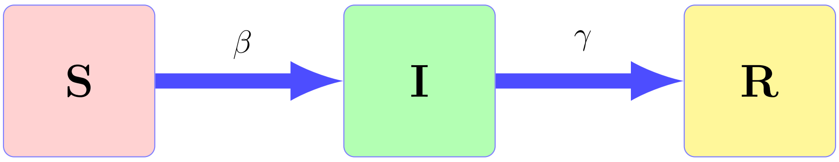


Figure 2.9.: Standard SIR model

Having compartmentalized the population, the standard SIR model is described by the following set of ordinary differential equations that specify how the sizes of compartments change over time:

$$\begin{cases} \dfrac{dS}{dt}(t) = -\frac{\beta}{N}S(t)I(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dI}{dt}(t) = \frac{\beta}{N}S(t)I(t) - \gamma I(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dR}{dt}(t) = \gamma I(t),\ t \in \mathbb{R}^+. \end{cases} \tag{2.4}$$

As for the differential system (2.2), system (2.4) can easily be solved analytically for all $t \in \mathbb{R}^+$. Besides, one can get more in realism by assuming that hypothesis $(\mathcal{H}5)$ is not held and that birth and death rates are equal and are denoted by b (see Figure 2.10). Then, the differential system is formulated as follows:

$$\begin{cases} \dfrac{dS}{dt}(t) = bN - bS(t) - \frac{\beta}{N}S(t)I(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dI}{dt}(t) = \frac{\beta}{N}S(t)I(t) - \gamma I(t) - bI(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dR}{dt}(t) = \gamma I(t) - bR(t),\ t \in \mathbb{R}^+. \end{cases} \tag{2.5}$$

Unlike system (2.4), system (2.5) has to be solved numerically.

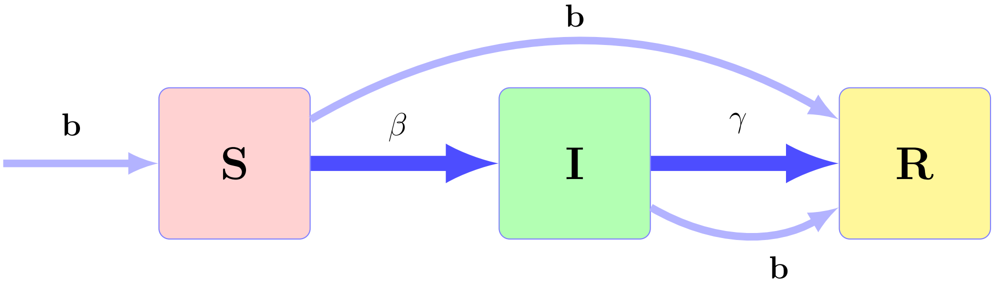


Figure 2.10.: SIR model including demographic factors

2.2.1.1.3. Standard SEIR Deterministic Model

The standard SIR model of Kermack and McKendrick [1991] has been extended to handle diseases with incubation period, during which the infected individual is not infectious yet. Thus, the individual is in a specific compartment, called E for exposed. The simplest SEIR model can be built based on the following hypotheses:

($\mathcal{H}$**1**) At time t, compartments S, E, I and R contain respectively S(t), E(t), I(t), and R(t) individuals;

($\mathcal{H}$**2**) The population is closed, i.e., the population size $N = S(t) + E(t) + I(t) + R(t)$ for each $t \in \mathbb{R}^+$ is constant;

($\mathcal{H}$**3**) β is the transmission intensity per unit of time;

($\mathcal{H}$**4**) α is the incubation rate (i.e., the rate of latent individuals becoming infectious) per unit of time;

($\mathcal{H}$**5**) γ is the rate of recovery per unit of time;

($\mathcal{H}$**6**) Demographic factors are excluded but could be added as in the system (2.5).

Thus, the above assumptions lead to the following system of differential equations that describes the standard SEIR model:

$$\begin{cases} \dfrac{dS}{dt}(t) = -\frac{\beta}{N}S(t)I(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dE}{dt}(t) = \frac{\beta}{N}S(t)I(t) - \alpha E(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dI}{dt}(t) = \alpha E(t) - \gamma I(t),\ t \in \mathbb{R}^+, \\[2ex] \dfrac{dR}{dt}(t) = \gamma I(t),\ t \in \mathbb{R}^+. \end{cases} \tag{2.6}$$

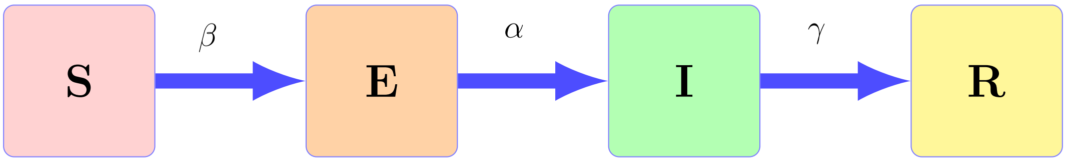


Figure 2.11.: Standard SEIR model

2.2.1.2. Partial Differential Equations

ODE models presented in the section above offer a temporal description of population dynamics, but do not allow for a spatial representation of its dynamics. Conversely, PDE models incorporate the spatial aspect into the model using a spatial variable denoted by x, and account for spatio-temporal interactions between population individuals. In the current section, the modeled quantity is population density u. Modeling the dynamics of the pathogen itself, would at least require modeling of susceptible hosts and infected hosts. To describe the transition between the discrete behaviour of an ODE system and the continuous behaviour of a PDE equation we use the SI model presented in Section 2.2.1.1.

Assume that x is the location of an individual in a 1-dimensional space. The second equation of System (2.3) satisfies:

$$\frac{I(x,t+dt)-I(x,t)}{dt} = \beta(N-I(x,t))I(x,t),\, t\in\mathbb{R}^+, \tag{2.7}$$

hence, Equation (2.7)$\times dx$ yields:

$$I(x,t+dt)\frac{dx}{dt} - I(x,t)\frac{dx}{dt} = \beta(N-I(x,t))I(x,t)dx,\, t\in\mathbb{R}^+. \tag{2.8}$$

When one adds transport terms relative to the movement of an individual in a spatial domain: moving from and to its vicinity, or staying at the same place (see Figure 2.12), one obtains:

$$\begin{aligned} dxI(x,t+dt) = {} & I(x,t)dx - 2D\frac{dt}{dx}I(x,t) \\ & + I(x+dx,t)D\frac{dt}{dx} \\ & + I(x-dx,t)D\frac{dt}{dx} \\ & + \beta(N-I(x,t))I(x,t+dt)dxdt,\, t\in\mathbb{R}^+. \end{aligned} \tag{2.9}$$

Thus, Equation (2.9)$\times\dfrac{1}{dxdt}$ one obtains the following discrete equation:

$$\frac{I(x,t+dt)-I(x,t)}{dt} = D\frac{I(x+dx,t)-2I(x,t)+I(x-dx,t)}{dx^2} + f(I),\, t\in\mathbb{R}^+,$$

such that,

$$f(I) = \beta(N-I(x,t))I(x,t+dt),$$

where dx and dt are respectively spatial and temporal variations, D is the diffusion rate, $D\dfrac{dt}{dx}$ is considered to be the probability of a host in the compartment I to move, and $f(I)$ is the so-called population growth term.

Hence, when $dt\to 0$, and $dx\to 0$, the above equations satisfies a PDE of the form:

$$\frac{\partial I}{\partial t} = D\frac{\partial^2 I}{\partial x^2} + f(I),\, t\in\mathbb{R}^+.$$

Let u be the probability of a host to be infected at time t in a location x. Based on equation (2.3), one can write that $u=\dfrac{I}{N}$ (N is assumed to be constant). Indeed,

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u),\, t\in\mathbb{R}^+.$$

This equation will be central in what follows.

2.2.1.2.1. Reaction-diffusion Equations for Modeling Short-distance dispersal

When one aims to model dispersal phenomena such as spatio-temporal dynamics of populations, reaction-diffusion equations are frequently used and have been exploited in many domains, especially in medicine, ecology, and epidemiology [Gatenby and Gawlinski, 1996, Roques, 2013b, Murray and Kulesa, 1996]. Reaction-diffusion equations are PDE of parabolic type [Evans, 1998]. Here, we describe some reaction-diffusion equations, in which dispersal is considered as a random diffusion process.

Random diffusion at the population level can be derived from random walks at the individual level. Random walks are often used to describe invasions by species that move via short-distance dispersal. Basic random walk models describe the path of an individual moving in a spatial domain via a succession of random steps. Typically, in a uni-dimensional space, as illustrated in Figure 2.12 and shown above, the individual located at x can move to the left and reach $x-d$ with probability $\mathbb{P}_L$, move to the right and reach $x+d$ with probability $\mathbb{P}_R$ or stay at the same place with probability $\mathbb{P}_S = 1 - \mathbb{P}_L - \mathbb{P}_R$. Such a microscopic and individual-based description of movements can be used to obtain diffusion equations at the population level [Roques, 2013b, Shigesada and Kawasaki, 1997a, Skellam, 1951]. In particular, the 1D random walk without directional bias and with constant and non-persistent increments leads to the following form of the diffusion equation: $\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}$, where u is the density of population.

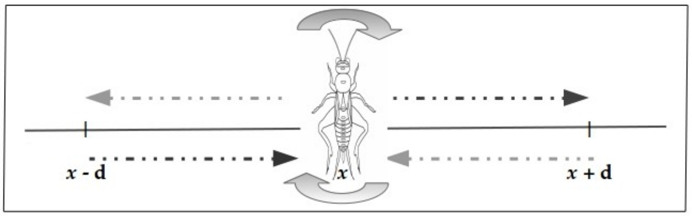


Figure 2.12.: Uni-dimensional random walk model.

In 1937, Fisher analyzed the rate of advance of advantageous genes with a PDE [Fisher, 1937], which has been generalized into:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + \underbrace{u(r - bu)}_{f(u)}, \quad t \geq 0, \tag{2.10}$$

where $u = u(t, x)$ is the frequency of the advantageous gene at time t and spatial location x in a uni-dimensional space; $D > 0$ is the coefficient measuring the rate of dispersal; r stands for the intrinsic growth rate of the species; and b corresponds to the coefficient measuring the effect of intra-specific competition; $f(u)$ is the population growth term.

In line with Fisher's work, Skellam [Skellam, 1951] proposed two-dimensional PDE for describing population dynamics. The so-called Skellam model, in particular, allowed him to study population spread with Malthusian growth theoretically. This model incorporates two terms, namely the population dispersal term and the population growth term, and assumes that there is no intra-specific competition:

$$\frac{\partial u}{\partial t} = D\Delta u + ur, \quad t \geq 0, \tag{2.11}$$

where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the 2-dimensional diffusion operator of Laplace.

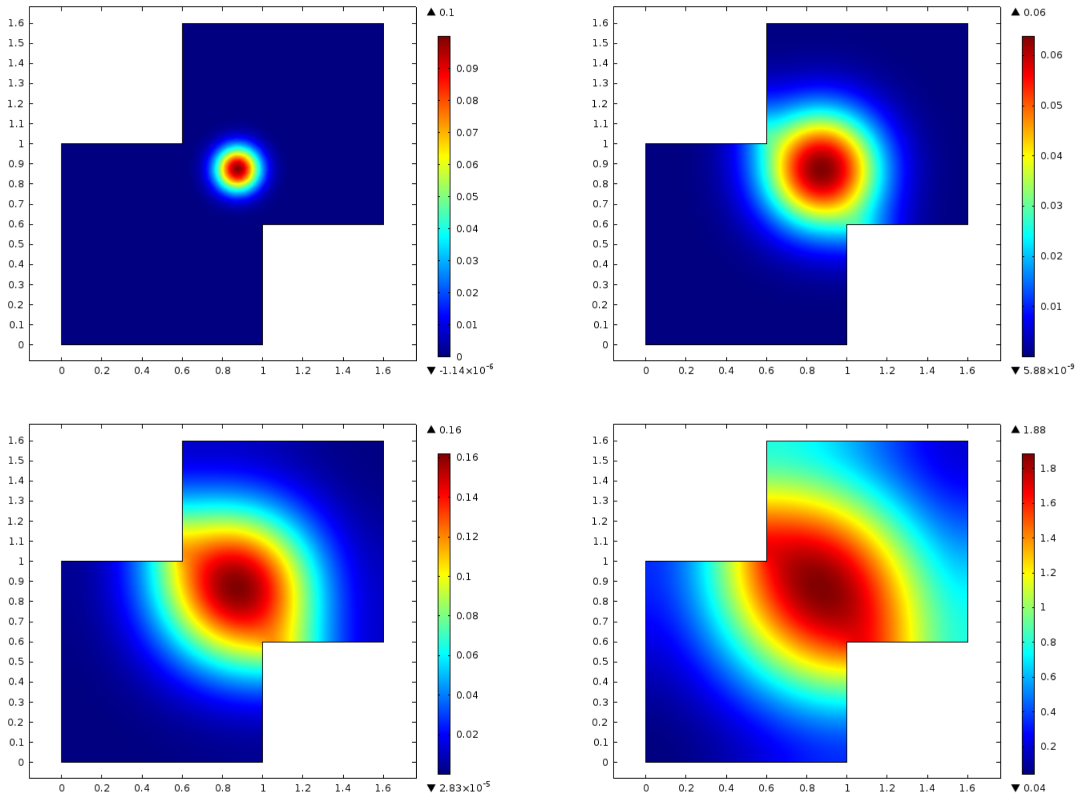


Figure 2.13.: Numerical solution $u(t, \mathbf{x})$ of Skellam model (2.11) in a bi-dimensional space (where $\mathbf{x} = (x, y)$) with Neumann boundary conditions, at time 0 (top left), 3 (top right), 6 (bottom left) and 12 (bottom right). The dispersal coefficient and the intrinsic growth rate were fixed at $(D, r) = (5 \times 10^{-3}, 0.5)$. The initial condition was $u(0, \mathbf{x}) = 0.1 \exp(-(10\|\mathbf{x} - \tilde{\mathbf{x}}_0\|)^2)$, where $\tilde{\mathbf{x}}_0 = (\tilde{x}_0, \tilde{y}_0) = (0.8, 0.8)$.

Figure 2.13 presents the solution of Equation (2.11) in a two-dimensional space, for specific values of parameters, initial conditions and boundary conditions.

Positive wavefront type-solutions exist for Equation (2.11). A simplified form of a traveling wave (in a unidimensional space) is a function of the form:

$$u(t, x) = U(x - ct),$$

where $c \in \mathbb{R}$ is the speed of the front $U \in \mathcal{C}^2(\mathbb{R})$. Note that a traveling wavefront can be defined not only when $t > 0$ but also for any $t \in \mathbb{R}$.

Skellam showed that the rate of spread at the front of the population range asymptotically approaches $c_0 = 2\sqrt{rD}$ when a small population is initially introduced at the origin. Furthermore, Luther [Luther, 1906] and Kolmogorov et al. [Kolomogorov and Piscouno, 1937] were the first to prove the existence of wavefront type-solutions for a diffusion equation with a logistic growth term $f(u) = ru(1 - u)$ (Fisher-KPP). Kolmogorov et al. showed that some initial distributions converge asymptotically to a traveling wave propagating to the right with a well defined, constant speed $c = 2\sqrt{rD}$. When the growth term includes an Allee effect as follows: $f(u) = ru(1 - u)(u - \theta)$, where $\theta \in]0, 1[$ is the Allee effect parameter, then there exists a unique positive wavefront-type solution with $\lim\limits_{x \longrightarrow -\infty} U = 1$, $\lim\limits_{x \longrightarrow +\infty} U = 0$. In addition, the speed of the front is [Hadeler and Rothe, 1975, Rothe, 1981, Lewis, MA and Kareiva, P, 1993]:

$$c = \sqrt{2rD}(\frac{1}{2} - \theta). \tag{2.12}$$

2.2.1.2.2. Integro-differential Equations for Modeling Long-distance Dispersal

The models introduced above are generally not adapted to describe the dynamics of populations that expand their range not only by neighborhood dispersal but also by long-distance dispersal, which can correspond to rare but significant events. An additional modeling approach for dispersal phenomena is the use of alternative representations of disease propagation that takes into account this twofold dispersal process [Alfaro et al., 2013]. In this case, the homogeneous diffusion can be replaced by a kernel-based term within an integro-differential equation. In this framework, the analogue of equation (2.10) will be written as follows:

$$\frac{\partial u}{\partial t} = J \star u - u + \underbrace{u(r - bu)}_{f(u)}, \quad t \geq 0, \tag{2.13}$$

with $J \star u$ a function of the form:

$$J \star u(t, x) = \int_{\mathbb{R}} J(|x - y|)u(t, y)dy,$$

where J is a dispersal kernel and the term $J(|x - y|)$ is the probability density of the dispersal distance $|x - y|$. This approach applied to population dynamics [Fife, 1979] generally allows a finer quantification of local and long-distance dispersal, and yields better predictions [Higgins and Richardson, 1999, Nathan et al., 2008, Fayard et al., 2009, Gilioli et al., 2013, White et al., 2017].

2.2.1.3. Stochastic Models for Population Dynamics

Diverse stochastic models have been proposed to represent randomness in population dynamics. Stochastic models can be built by adding stochastic components to deterministic equations in order to include some inherent randomness, but diverse approaches for constructing stochastic population dynamics models have been developed, for instance: Stochastic Differential Equation (SDE) used to model trajectories of individuals [Gloaguen et al., 2018]; branching processes used to model the growth and adaptation of populations [Méléard, 2011]; coalescent processes used to generate genealogies which relate a random sample of individuals resulting from a basic forward-time population model [Kingman, 1982, Gill et al., 2012]; temporal point processes used to build birth-death models [Champagnat et al., 2006]; spatio-temporal point processes used to model the temporal evolution of the spatial pattern of individuals forming a population [Soubeyrand et al., 2011]; stochastic Markovian areal processes used to model large-scale dynamics [Soubeyrand et al., 2009b]; and regressions (eventually including auto-regressive components) used to take into account the effect of environmental variables on population characteristics [Bordier et al., 2017]. Compared to their deterministic analogues, if any, stochastic population-dynamic models generally allow relaxing hypotheses made on the dynamics and generating more diverse realizations. However, these models tend to be less tractable in the analysis of model properties and the estimation of unknown components. Suppose that we are interested in fitting a spatio-temporal population dynamics model to data. There is, like in many other application fields, a trade-off between model realism and estimation complexity. For example, fitting a population dynamics model essentially constructed from a PDE containing a few parameters [Soubeyrand and Roques, 2014] is generally easier than fitting a (more flexible and realistic) hierarchical stochastic spatio-temporal Markovian model including a few parameters but numerous latent variables [Soubeyrand et al., 2009b]. In this example, two extreme cases are considered:

- a model with deterministic behavior and few degrees of freedom, which may yield poor goodness-of-fit;
- a model with stochastic behavior and lots of degrees of freedom, which may induce identifiability issues.

All models are wrong, but some are useful [Box, 1976]. To construct useful models and avoid the two extreme cases described above, allowing rapid, realistic, relevant and consistent inferences are required. Spatio-temporal PDMP can play this role.

2.2.1.4. Piecewise-Deterministic Markov Processes[1]

PDMP were introduced by Davis [1984] as a general family of non-diffusive stochastic models. PDMP are càdlàg Markov Processes (i.e., right continuous with left limits Markov processes), whose behavior is governed by a deterministic continuous motion (the flow) disrupted at random times by discrete random events (the jumps). Such an event can be either a discontinuity in the trajectory of the stochastic processes or merely a change in the rule of the continuous dynamics [Davis, 1984, Azaïs and Bouguet, 2018]. This class of models is often built to model temporal processes and is rarely encountered in the literature in a spatio-temporal framework. Nevertheless, spatio-temporal PDMP can be occasionally encountered in the theoretical and quantitative population dynamics literature, but these models are generally not called PDMP. For instance, spatio-temporal PDMP have been built at the population level [Shigesada et al., 1995], the metapopulation level (which is a set of populations) [Soubeyrand et al., 2009a] and the individual level [Caillerie, 2017]. These processes are illustrated in Figure 2.14.

Let $\mathcal{X} \in \mathbb{R}^n$ be an open subset, $\partial\mathcal{X}$ its boundary, $\bar{\mathcal{X}}$ its closure and $\mathcal{B}(\mathcal{X})$ the set of real-valued, bounded, measurable functions of $\mathcal{X}$. A PDMP $X(t) = \{X_t;\, t > 0\}$ with values in $\mathcal{X}$ is defined by three basic elements:

- a deterministic continuous flow $\Phi : \mathcal{X} \times \mathbb{R} \to \mathcal{X}$, which drives the dynamics of the process between the jumps;
- a jump rate $\lambda : \mathcal{X} \to \mathbb{R}_+$ which triggers the jump mechanism of the process (the larger $\lambda(x)$, the higher the probability to jump);
- a jump kernel $Q : \mathcal{X} \to \mathcal{X} \times \mathcal{B}(\mathcal{X})$, which rules the directions and the amplitudes of the jumps of X.

A classical PDMP is the one in which the flow is driven by an ODE supposed to have a unique solution Φ and the jump is governed by a Poisson process. With these tools, one can define

[1] Here, I use the notations of [Azaïs and Bouguet, 2018], which are also consistent with the notations of Chapter 5.

the sample paths of X recursively. Given $T_0 = 0$ and $X_0 \in \mathcal{X}$, let S_1 be a positive random variable such that, for all $t > 0$,

$$\mathbb{P}(S_1 \geq t) = \exp\left(- \int_0^t \lambda(\Phi(X_0, s))ds \right) \mathbb{1}_{\{t<t^+(X_0)\}}, \tag{2.14}$$

where $t^+(x) = \inf\{t > 0 : \Phi(x, t) \in \partial\mathcal{X}\}$, with the convention $\inf \emptyset = +\infty$.
Then, let $T_1 = T_0 + S_1$ be the first jump time and let U_1 such that $\mathbb{P}_{U_1} = Q(\Phi(X_0, S_1), \cdot)$. We can define, for $t \in [T_0, T_1]$,

$$X_t = \begin{cases} \Phi(X_0, t) & \text{if } t \in [T_0, T_1), \\ U_1 & \text{if } t = T_1, \end{cases} \tag{2.15}$$

and so on for the subsequent intervals $[T_1, T_2]$, $[T_2, T_3]$, ...

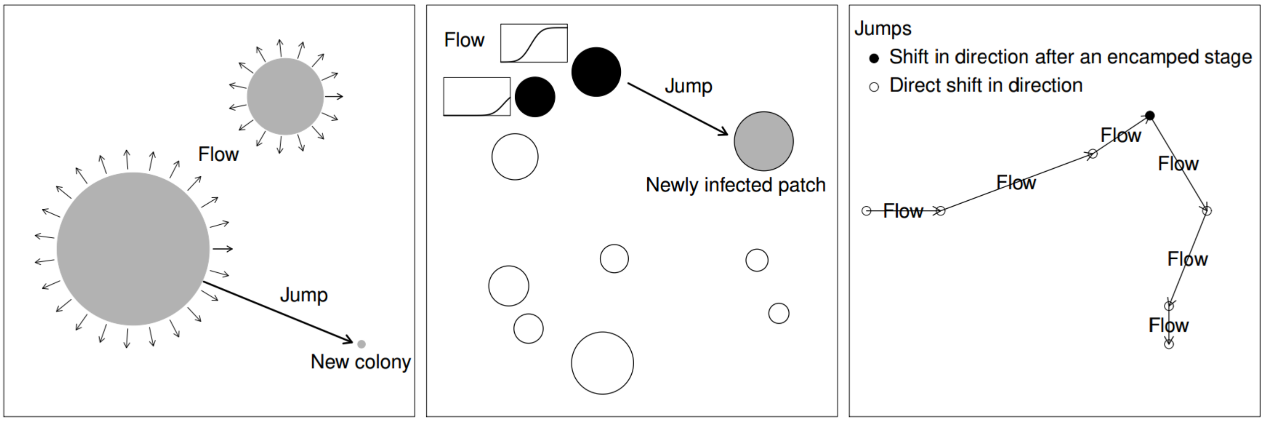


Figure 2.14.: Illustrations of the flows and jumps for the coalescing colony model (left), the metapopulation epidemic model (center) and the simple velocity-jump model (right) proposed by [Shigesada et al., 1995], [Soubeyrand et al., 2009a] and [Caillerie, 2017], respectively.

2.2.1.5. Key Points of Mathematical Models in Population Dynamics

- ❖ Relatively concise deterministic models do not provide a full description of all the biological and environmental drivers of population dynamics, but can be fitted to data in a reasonable time span. The advantage of these models is that they can be rapidly applied for endorsing a fast reaction after the detection of a new pathogen.

- ❖ Stochastic models of population dynamics can account for some inherent randomness but tend to be less tractable models from analysis and estimation perspectives. The advantage of opting for a stochastic modeling approach is that it allows getting more realistic and flexible realizations.

- ❖ The difficult task will be to find a convenient trade-off between deterministic and stochastic models, bearing in mind that the purpose of the thesis research is to provide a generic methodology (i.e., a method scalable to various invasive species) that provides predictions for an invasive species dynamics, for which no dedicated model is available and whose initial conditions (i.e., date and location of the introduction of the invasive species) are unknown. Stratified dispersal models or PDMPs appear as alternative models to achieve rapid, realistic and consistent inferences.

2.2.2. Inferring Population Dynamics From Mathematical Models

2.2.2.1. Overview

Mathematical modeling approaches can contribute to understand population dynamics because one can interpret the parameters in the models and the relationships expressed between and among variables. Inferring population dynamics provides further relevant and useful knowledge about model parameters that can be deduced from data. This is also known as model calibration or solving inverse problems. In the following material, we will only review the main methods applied to infer models in population dynamics described by differential equations.

Traditionally, the estimation of parameters and any derived quantities have been carried out using the Least-squares (LS) approach which have enjoyed an early history of application [Weisberg, 1985]. A recent work of Li et al. [2019] inferred parameters using the LS method in a nonlinear mean-reversion SDE model driven by Brownian motion for a population growth model. With the increasing use of computers in mathematical analysis, the LS approach was progressively replaced by the maximum-likelihood (ML) method [Burnham et al., 1995]. In the present time, the Maximum Likelihood (ML) method is very widespread [Müller et al., 2004, Timmer et al., 2004, Baker et al., 2005, Luzyanina et al., 2008, Roques et al., 2016]. For instance, Roques et al. [2016] used the ML approach to infer diffusion rates of a PDE-based population dynamics model, Luzyanina et al. [2008] estimated cell birth rate in a cell population dynamics model also based on a PDE. In general, this approach provides point estimates of model parameters and the sampling variance-covariance matrix or other quantities related to estimation uncertainty.

Besides, in the past decade, Bayesian methods have been increasingly used in the field of population dynamics. For example, Gillespie and Golightly [2010] estimated parameters in a stochastic population growth model using an Markov Chains Monte Carlo (MCMC) method; Heydari et al. [2014] proposed a Bayesian parameter estimation for stochastic logistic growth models; Gilioli et al. [2012] proposed a Bayesian inference method to estimate parameters in a stochastic predator-prey system. Bayesian inference approaches have also been applied to PDE-based population dynamics models. For instance, Spence et al. [2016] estimated parameters of a PDE model for aquatic communities. Roques et al. [2011] estimated the local fitness parameters and the diffusion parameter of a reaction-diffusion model of population dynamics. Lanzarone et al. [2017] estimated the mortality rates coupling a system of PDE to the MCMC algorithm. A joint estimation of such propagation characteristics (diffusion rates, growth rates, mortality rates) and the initial condition (date and site of the introduction) was proposed by Soubeyrand and Roques [2014] using a MCMC algorithm with a simple reaction-diffusion model, and was applied to simulated data (i.e., data collected over a temporal window covering a period after the introduction time). When one only has at disposal post-introduction data, and if one aims to estimate the introduction point, it is required to also estimate the propagation characteristics of the invasive species (and *vice versa*), as suggested in Soubeyrand and Roques [2014], because these characteristics link the introduction and the observations.

2.2.2.2. Inferential Statistical Methods

Inference about a certain parameter vector Θ can be made either in a frequentist framework, which consists in assessing an objective point estimate $\hat{\Theta}$ of Θ given an appropriate model, or in a Bayesian framework, which technically consists in assessing the posterior distribution $[\Theta|Y]$ of Θ conditional on data Y. Philosophically, a posterior probability is to be interpreted as a coherent judgment quantifying a subjective degree of uncertainty [Lindley, 2006]. The benefit of the Bayesian approach is primarily to allow the incorporation of prior expertise into the statistical analysis and the rigorous assessment of dependencies and uncertainties in estimation (via the joint posterior distribution of parameters). Moreover, most people better understand the direct probabilistic judgments about the unknowns provided by the Bayesian paradigm when reporting uncertainty [O'Hagan, 2008]. In addition, the Bayesian approach is computationally costly but leads to improved outcomes since it provides the joint probability distribution of the unknowns given the observations and the prior knowledge, thus providing complete information about the shape of the density and the uncertainty about parameters, e.g., see Lanzarone et al. [2017].

Henceforth, we will keep using Gelfand's bracket notation for probability distributions [Gelfand and Smith, 1990]. The posterior distribution of the unknown, hereafter dubbed Θ, is derived by Bayes' rule:

$$[\Theta|Y] = \frac{[Y|\Theta] \times [\Theta]}{[Y]},$$

where $[Y|\Theta]$ is the conditional distribution of the data Y given the unknown Θ (i.e., the likelihood function of the model); $[\Theta]$ is the prior distribution of Θ that depends on the application; the distribution of Y, $[Y] = \int [Y|\Theta][\Theta]d\Theta$, may be a formidable integral, depending on the dimension of the unknown Θ. However, modern Bayesian algorithms [Brooks, 2003] avoid its computation by making recourse to Markov Chain (MC) techniques only based on the un-normalized probability function $[Y|\Theta] \times [\Theta]$.

In what follows, we will present in a generic framework, the main statistical techniques used to infer model parameters in both frequentist and Bayesian approaches, showing the pros and cons of each approach.

2.2.2.2.1. Maximum-likelihood approach

Nowadays, the ML approach is the most widespread frequentist approach. This approach provides an objective[2], omnibus theory for estimation of model parameters and the sampling covariance matrix [Burnham et al., 1995]. The likelihood is a real-valued function denoted by $\mathcal{L}(Y;\Theta) = [Y|\Theta]$ and is given by evaluating the joint probability of the observed data sample $Y = (y_1, y_2, \ldots, y_I)$ of size I given Θ. The primary goal of the ML approach is to compute the global maximum of the likelihood function :

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{L}(Y;\Theta).$$

The likelihood is often a function of a large dimensional vector Θ and may have a complicated surface with several local maxima, all of which may lead the optimization to get stuck in local maxima [Dattner et al., 2017]. The most commonly used method to find the maximum likelihood is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [Broyden, 1969, Fletcher, 1970, Goldfarb, 1970, Shanno, 1970] method, which belongs to Quasi-Newton methods. The BFGS algorithm is implemented in computer software packages such as `Matlab`. In complex settings, the search for acceptable optimum will require complex optimization al-

[2] Unlike Bayesian inference, the frequentist approach allows for a subjective choice of the estimator. However, when this choice is made up, the frequentist approach is objective itself.

gorithms, such as the simulated annealing algorithm [Kirkpatrick et al., 1983], in order to converge to a global maximum.

2.2.2.2.2. Least-squares Approach

The LS approach [Legendre, 1805] targets the estimation of model parameters by minimizing the squared discrepancies between observed data Y and their expected values u_i under a specified model parameterized by Θ:

$$\tilde{\Theta} = \min_{\Theta}\Big(\sum_{i=1}^{I}(y_i - u_i)^2\Big).$$

LS estimation corresponds to maximum likelihood estimation when the noise is normally distributed with equal variances.

Other fitting measures are sometimes preferred, for example, Least absolute deviations (LAD) [Portnoy and Koenker, 1997], which is more robust to outliers. In addition, it is worth noting that general frameworks, such as contrast estimation [Lánska, 1979, Dacunha-Castelle and Duflo, 1982], embedding ML, LS, and LAD have been proposed to pool several frequentist estimation approaches into a single setting providing general results about convergence properties of estimators.

2.2.2.2.3. Markov Chain Monte Carlo Algorithms

Since the 1990s, the MCMC techniques have in majority been oriented towards Bayesian inferences [Berliner, 2003, Lanzarone et al., 2017, Roques et al., 2011, Soubeyrand et al., 2009a,b, Wikle, 2003a,b]. These computer-intensive techniques include a class of algorithms for drawing samples from probability distributions [Gilks et al., 1996]. In the Bayesian context, the chains provided by MCMC are supposed to converge to a stationary distribution: the posterior distribution $[\Theta|Y]$. Once the stationarity is reached, the chain can be used to sample from the posterior distribution of the parameters. The MCMC techniques can adequately explore the space of the target distribution and find the regions with high probability, provided that the proposal distribution is well-chosen (neither diffusive nor focused) to fastly reach the stationarity of the chain. In general, MCMC methods are more appropriate for large dimensional problems than classical importance sampling methods [Robert and Casella, 1998, Parent and Bernier, 2007].

Here, we briefly present the widely used versions of MCMC : Gibbs Sampler and the Metropolis-Hastings algorithms. Conditions leading to the theoretical convergence of these two widely used versions of MCMC , are described in Roberts and Smith [1994].

➤ Metropolis-Hastings Sampler: Metropolis et al. [1953] proposed a sampler to construct the Markov chains. Later, this sampler have been generalized by Hastings [1970]. Generally speaking, the Metropolis-Hastings algorithm designs a Markov process by constructing transition probabilities from the proposal density which can be any chosen density from which we can draw samples. **Pseudo-code #1** describes the steps of the Metropolis-Hastings algorithm.

➤ Gibbs Sampler: The Gibbs sampler [Casella and George, 1992] is a method for sampling from distributions over at least a parameters vector of two dimensions. It can be viewed as a special case of the Metropolis-Hastings algorithm in which the proposal distributions are full conditional distributions. For this sampler to be usable, these conditional distributions should be tractable, i.e., straightforward to sample from. Should this not be the case, and in high dimensional problems, it is required to use other samplers such as the Metropolis-Hastings algorithm. **Pseudo-code #2** describes the steps of the Gibbs sampler.

Pseudo-code #1: Metropolis-Hastings Sampler

```
Suppose that $\pi(\Theta)$ is the density or probability mass function of the target distribution, e.g. $\pi(\Theta) = [\Theta \mid Y]$
  1. Initialize the chain by setting an initial state $\Theta_0$ for the parameter vector $\Theta$, such that $\pi(\Theta_0) \neq 0$
  2. At iteration $m \geq 1$
     a) Sample $\Theta^*$ from the proposal distribution $\Theta^* \to g(\Theta^*, \Theta_{m-1})$ and $u$ from $\mathcal{U}([0,1])$
     b) Compute the probability of acceptance $\alpha$:
          $$\alpha = \min\left\{1, \frac{\pi(\Theta^*)}{\pi(\Theta_{m-1})}\frac{g(\Theta_{m-1}|\Theta^*)}{g(\Theta^*|\Theta_{m-1})}\right\}.$$
          – If $u \leq \alpha$, then $\Theta_m = \Theta*$
          – If $u > \alpha$, then $\Theta_m = \Theta_{m-1}$
```

Pseudo-code #2: Gibbs Sampler

```
  1. Initialize the chain by setting an initial state $\Theta_0 = (\Theta_0^1, \Theta_0^2, \dots, \Theta_0^J)$ for the parameter vector $\Theta$, such that: $[Y|\Theta_0] \times [\Theta_0] \neq 0$
  2. At iteration $m \geq 1$
     – Generate $\Theta^1_m$ from the distribution $[\Theta^1|\Theta^2_{m-1}, \dots, \Theta^j_{m-1}, \dots, \Theta^J_{m-1}, Y]$
     – Generate $\Theta^2_m$ from the distribution $[\Theta^2|\Theta^1_m, \Theta^3_{m-1}, \dots, \Theta^j_{m-1}, \dots, \Theta^J_{m-1}, Y]$
     – Generate $\Theta^j_m$ from the distribution $[\Theta^j|\Theta^1_m, \Theta^2_m, \dots, \Theta^{j-1}_m, \Theta^{j+1}_{m-1}, \dots, \Theta^J_{m-1}, Y]$
     – Generate $\Theta^J_m$ from the distribution $[\Theta^J|\Theta^1_m, \Theta^2_m, \dots, \Theta^j_m, \dots, \Theta^{J-1}_m, Y]$
```

2.2.2.2.4. Classical Importance Sampling

Among MC methods, Importance sampling (IS) [Rubin, 1987, Ripley, 1987] consists of generating an initial sample of size M from a proposal distribution, and using this sample to provide an empirical approximation of the parameters posterior distribution via a final weighted sample with respect to the integrated likelihood. This final sample $\{\tilde{\Theta}_m\}_{m=1}^{\tilde{M}}$ is obtained by resampling $\tilde{M}$ values with replacement from the initial sample of size M generated from the proposal distribution, where the normalized importance weight $\tilde{w}_m$ is the sampling probability of Θ_m. M and $\tilde{M}$ should be chosen large enough to best satisfy the law of large numbers. **Pseudo-code #3** describes the steps of the IS.

The appeal of IS remains in its solid theoretical foundations [Robert and Casella, 1998, Parent and Bernier, 2007] (e.g., non-biased estimator, controlled variance) and its simple implementation. In addition, the IS offers a framework to easily estimate the integrated likelihood by averaging over the unnormalized weights computed in equation (2.16). However, to design efficient IS algorithms, the proposal distribution should be chosen as close as possible to the posterior distribution. The posterior distribution being unknown, the crucial choice of the proposal is a difficult task [Gelman et al., 1996, Roberts et al., 1997].

Pseudo-code #3: IS

```
1. Generate an initial sample $\{\Theta_m\}_{m=1}^{M}$ from the proposal distribution $\Theta \to g(\Theta)$
2. Compute the unnormalized importance weights as follows:
       $$w_m = \frac{[\Theta_m|Y]}{g(\Theta_m)}, \quad (2.16)$$
3. Normalize the weights:
       $$\tilde{w}_m = \frac{w_m}{\sum_{k=1}^{M} w_k}, \quad (2.17)$$
4. Sample with replacement $\{\tilde{\Theta}_m\}_{m=1}^{\tilde{M}}$ in $\{\Theta_m\}_{m=1}^{M}$ weighted by $\{\tilde{w}_m\}_{m=1}^{M}$.
```

2.2.2.2.5. Adaptive Multiple Importance Sampling Algorithm

The AMIS is an iterated importance sampling scheme [AMIS; Cornuet et al., 2012]. AMIS consists of iteratively generating parameter vectors under an adaptive proposal distribution and assigning weights to the parameter vectors. The main aim of the AMIS algorithm is to overcome the difficulty related to the choice of the proposal distribution encountered in classical MCMC and IS, by tuning the coefficients of the proposal distribution picked in a parametric family of distributions, generally the Gaussian one, at the end of each iteration.

In this framework, at each iteration, new coefficient values for the proposal distribution are determined using the current weighted posterior sample [Bugallo et al., 2015], then the posterior sample is augmented by generating new replicates from the newly tuned proposal distribution and the weights of the cumulated posterior sample are recomputed. The algorithm is described in **Pseudo-code #4**.

The AMIS algorithm provides a weighted posterior sample $\{\{\Theta_m^l, w_m^l\}_{l=1}^{L}\}_{m=1}^{M}$ of size ML, which provides an empirical approximation of the posterior distribution $[\Theta|Y]$. Conditions leading to the convergence in probability of the posterior mean of any function (integrable with respect to the posterior distribution) of the parameters are described in Cornuet et al. [2012], Marin et al. [2012] and are satisfied in our case. If in practice, the convergence of

AMIS to the true posterior cannot be numerically demonstrated (because the true posterior is not known), one can still assess its stabilization.

Like other adaptive importance sampling algorithms [Bugallo et al., 2015], AMIS can be easily parallelized and its tuning parameters are automatically adapted across the algorithm iterations, contrary to the basic MCMC and the ML approach. It has however to be noted that AMIS has to be appropriately initialized, which can be relatively easily done in practice by evaluating the marginal posterior distributions over 1D grids. Still, in regard with the computational cost, ML estimation could be an attractive option, even if the control of estimation uncertainty is usually more convincing in the Bayesian framework. Using AMIS can yield gains in computation time with respect to MCMC. From an example in population genetics, Cornuet et al. [2012] observed that AMIS was six times faster than MCMC for providing similar posteriors with slightly better repeatability in the case of AMIS (without parallelization). The authors mentioned that AMIS is particularly interesting in cases where the likelihood is computationally expensive because all particles simulated during the process are recycled, which decreases the number of calls of the likelihood function.

Pseudo-code #4: AMIS

```
1. Set initial values $\mu_0$ and $\Sigma_0$ for the mean vector and the variance matrix of the multi-normal proposal distribution $\mathcal{N}(\mu_0, \Sigma_0)$, whose probability density function is denoted by $\Theta \to g_{\mu_0,\Sigma_0}(\Theta)$.

2. At iteration $m = 1, \cdots, M$,

    a) Generate a new sample $\{\Theta_m^l : l = 1 \cdots, L\}$ from the proposal distribution $\mathcal{N}(\mu_{m-1}, \Sigma_{m-1})$.

    b) Compute the un-normalized importance weights for the new sample as in Equation (2.18), and re-compute the un-normalized weights for the previously generated samples as in Equation (2.19):

       $$\tilde{w}_m^l = \frac{[Y|\Theta_m^l] \times [\Theta_m^l]}{\frac{1}{m}\sum_{j=1}^{m} g_{\mu_{j-1},\Sigma_{j-1}}(\Theta_m^l)}, \quad l = 1, \cdots, L, \qquad (2.18)$$

       $$\tilde{w}_\varepsilon^l = \frac{[Y|\Theta_\varepsilon^l] \times [\Theta_\varepsilon^l]}{\frac{1}{m}\sum_{j=1}^{m} g_{\mu_{j-1},\Sigma_{j-1}}(\Theta_\varepsilon^l)}, \quad \varepsilon = 1, \cdots, m-1, \ l = 1, \cdots, L, \qquad (2.19)$$

       where $g_{\mu_{j-1},\Sigma_{j-1}}$ is the probability density function of the multi-normal distribution with mean vector $\mu_{j-1}$ and variance matrix $\Sigma_{j-1}$.

    c) Normalize the weights:

       $$w_\varepsilon^l = \frac{\tilde{w}_\varepsilon^l}{\sum_{i=1}^{m}\sum_{j=1}^{L} \tilde{w}_i^j}, \quad \varepsilon = 1, \cdots, m, \ l = 1, \cdots, L.$$
```

d) Adapt coefficient values for the next proposal distribution as follows:

$$\mu_m = \sum_{l=1}^{L}\sum_{\varepsilon=1}^{m} w_\varepsilon^l \Theta_\varepsilon^l.$$

$$\Sigma_m = \sum_{l=1}^{L}\sum_{\varepsilon=1}^{m} w_\varepsilon^l (\Theta_\varepsilon^l - \mu_\varepsilon)(\Theta_\varepsilon^l - \mu_\varepsilon)^t.$$

2.2.2.2.6. A brief comment on frequentist and Bayesian approaches

Here, we briefly comment the difference between the frequentist and Bayesian approaches in terms of interpretation of results.

- The frequentist approach describes the uncertainty about the value obtained through an estimation procedure. In order to assess the reliability of the algorithmic-computation procedure, one needs to make simulation studies using replicates of observation generated by the model, and repeat many times the estimation procedure.

- The Bayesian approach assesses the remaining uncertainty about the unknowns, conditionally to a unique observed dataset, and given a fundamental belief about their possible variations. This is, therefore, a probabilistic judgment similar to a personnel challenge with respect to the hypothetical value of these unknowns. There exist numerous algorithms that allow computing the probability distribution relative to this probabilistic judgment.

2.2.2.3. Key Points of Inferring Population Dynamics From Mathematical Models

- The Bayesian approach is generally computationally more costly than a frequentist approach, but leads to improved outcomes since it systematically provides the joint probability distribution of the unknowns given the observations and the prior knowledge, thus providing complete information about the shape of the density and the uncertainty about parameters.

- ML and LS estimation could be an attractive option, to regard with computational cost, even if the control of estimation uncertainty is usually more convincing in the Bayesian framework. However, the likelihood function is often multivariable and may have a complicated surface. Thus, the LS and ML approach may get stuck in local maxima.

- The MCMC techniques can adequately explore the parameter space to find the regions with high posterior probability. In general, MCMC methods require fewer iterations for the same level of accuracy than basic importance sampling methods, provided that the proposal distribution is well-chosen to allow for fast stationarity of the chain.

- Using AMIS is expected to allow gains in computation time with respect to MCMC. In addition, AMIS can be easily parallelized and the parameters of the proposal distribution are automatically adapted across the algorithm iterations, contrary to the basic MCMC approach.

2.2.3. Model Selection and *Model-averaging*

To statistically reconstruct the past and predict the future dynamics of pathogens, this dynamics should be represented in a phenomenological and concise way. However, this approach necessitates to ignore some processes and sources of variability involved in the epidemiological dynamics, and models with various structures are likely to be considered as candidate models. When the goal of the study is to make inferences and draw predictions, the use of a single model is prone to prediction error because this model might not take into account crucial drivers of the dynamics. This limitation can be circumvented by considering a family of candidate models and applying a model selection strategy [Burnham et al., 1995] or a model aggregation strategy [Hoeting et al., 1999].

2.2.3.1. Model Selection Methods

Numerous methods have been proposed as part of the model selection strategy, whether it is for an explanatory or a predictive purpose. In this section, we focus on information criteria: The Akaike's information criterion (AIC) [Akaike, 1973] satisfies:

$$\text{AIC} = -2\log[Y|\hat{\Theta}] + 2k, \tag{2.20}$$

where k is the number of model parameters, and $\hat{\Theta}$ is the maximum likelihood estimate of the parameter vector Θ in the support $\mathcal{S}(\Theta)$ of Θ:

$$\hat{\Theta} = \underset{\Theta\in\mathcal{S}(\Theta)}{\text{argmax}}[Y|\Theta].$$

The Bayesian information criterion (BIC) [Schwarz et al., 1978] satisfies:

$$\text{BIC} = -2\log[Y|\hat{\Theta}] + k\log I, \tag{2.21}$$

where I is the sample size.
The Deviance information criterion (DIC) satisfies:

$$\text{DIC} = \bar{\mathcal{D}} + p_{\text{eff}}, \tag{2.22}$$

where $\bar{\mathcal{D}}$ is the posterior mean of the deviance $\mathcal{D}(\Theta) = -2\log[Y|\Theta]+C$ (where C is a constant that cancels out when one compares different models) and p_{eff} is the effective number of parameters of the model. The difference in the two versions of the DIC considered here lies in the calculation of p_{eff}. In the first version proposed by Spiegelhalter et al. [2002],

$$p_{\text{eff}} = p_{\mathcal{D}} = \bar{\mathcal{D}} - \mathcal{D}(\bar{\Theta}), \tag{2.23}$$

where $\bar{\Theta}$ is the posterior mean of Θ: $\bar{\Theta} = \mathbb{E}[\Theta|Y]$. In the second version proposed by Gelman et al. [2003],

$$p_{\text{eff}} = \frac{1}{2}\mathbb{V}(\mathcal{D}(\Theta)|Y), \tag{2.24}$$

where $\mathbb{V}(\mathcal{D}(\Theta)|Y)$ is the posterior variance of $\mathcal{D}(\Theta)$.
The Predictive information criterion (IC) of Ando [2011], which is supposed to solve overfitting issues, satisfies:

$$\text{IC} = \bar{\mathcal{D}} + 2p_{\mathcal{D}} := 3\bar{\mathcal{D}} - 2\mathcal{D}(\bar{\Theta}). \tag{2.25}$$

The Widely Applicable Information Criterion (WAIC) [Watanabe, 2010, 2013], which is an estimate of the expected log pointwise predictive density for a new dataset, satisfies:

$$\text{WAIC} = \text{lppd} - \text{p}_{\text{WAIC}},$$

where lppd is the log pointwise predictive density and p_{WAIC} is an estimate of the effective number of free parameters in the model. Let $(Y_q)_{q=1,\cdots,Q}$ be a division of the data Y. The lppd is the logarithm of the predictive density integrated over the posterior distribution of the model parameters summed over all the observations and is expressed as follows:

$$\text{lppd} = \sum_{q=1}^{Q} \log \int [Y_q|\Theta] \times [\Theta|Y] d\Theta. \tag{2.26}$$

To compute the lppd we use a sample of size n drawn from the posterior distribution of Θ. n has to be chosen large enough to best approach the former distribution. Thus lppd is calculated as follows:

$$\text{lppd} = \sum_{q=1}^{Q} \log \left(\sum_{i=1}^{n} [Y_q|\Theta^i] w_i \right). \tag{2.27}$$

Two forms for p_{WAIC} have been proposed in the literature. In the first approach, the p_{WAIC} is measured as follows:

$$\text{p}_{\text{WAIC}} = 2 \sum_{q=1}^{n} \left(\log \Big(\mathbb{E} \left[Y_q|\Theta \right] \Big) - \mathbb{E} \Big(\log \left[Y_q|\Theta \right] \Big) \right). \tag{2.28}$$

In the second approach, the p_{WAIC} is measured using the variance of individual terms in the log predictive density summed over the Q subsets of data:

$$\text{p}_{\text{WAIC}} = \sum_{q=1}^{Q} \mathbb{V} \Big(\log[Y_q|\Theta] \Big). \tag{2.29}$$

2.2.3.2. Bayesian *Model-averaging* Method

As part of the aggregation strategy, the BMA approach has been proposed by Leamer [1978] to reduce and account for parameter and model uncertainties. This approach consists on averaging over all the candidate models in a Bayesian way for weighting models [Raftery, 1996, Volinsky et al., 1997], combining multiple predictions and combining estimations to infer shared parameters [Roberts, 1965, Madigan and Raftery, 1994, Wintle et al., 2003]. Theoretically, BMA provides better average predictive ability, as measured by a logarithmic scoring rule, than using any single model [Madigan and Raftery, 1994]. The BMA efficiency has been largely explored, in particular with respect to its theoretical properties [Rubin and Schenker, 1986, Madigan and Raftery, 1994, Raftery and Zheng, 2003], leave-one-out predictive performance [Madigan et al., 1995, Lamon and Clyde, 2000, Fernández et al., 2002] and numerical performance [George and McCulloch, 1993, Clyde and George, 2000, Viallefont et al., 2001]. While BMA is an intuitively attractive solution to the problem of accounting for model uncertainty, it presents several difficulties related to its numerical implementation [Hoeting et al., 1999]. By dint of some pioneering work implementing BMA [Madigan and Raftery, 1994, Raftery, 1996, Volinsky et al., 1997], BMA has been applied in numerous study domains such as medicine [Oehler et al., 2009, Yin and Yuan, 2009], ecology [Boone et al., 2005, 2008, Wintle et al., 2003], meteorology [Raftery et al., 2005], genetics [Yeung et al., 2005], economical and political sciences [Eicher et al., 2011, Sidman et al., 2008], engineering and physical sciences [Raftery et al., 2010, Parkinson and Liddle, 2013] and epidemiology [Viallefont et al., 2001]. Despite ample literature on BMA and its usefulness, it has only been marginally applied in the context of predictive epidemiology.

Briefly, the BMA consists in computing the average of the posterior distribution of the variable of interest Δ provided under all the competing models $\{\mathcal{M}_j; j = 1, \cdots, J\}$ and weighted by the posterior model probabilities [Raftery, 1996, Hoeting et al., 1999]. Δ is typically a set of shared parameters or a future observation. The BMA posterior distribution of Δ given training data Y satisfies:

$$[\Delta|Y] = \sum_{j=1}^{J}[\Delta|Y, \mathcal{M}_j] \times [\mathcal{M}_j|Y].$$

The posterior model probability of $\mathcal{M}_j$ is

$$[\mathcal{M}_j|Y] = \frac{[Y|\mathcal{M}_j] \times [\mathcal{M}_j]}{\sum_{j'=1}^{J}[Y|\mathcal{M}_{j'}] \times [\mathcal{M}_{j'}]}. \tag{2.30}$$

The integrated likelihood $[Y|\mathcal{M}_j]$ of $\mathcal{M}_j$, which may be a formidable integral, depending on the dimension of the unknown Θ, satisfies:

$$[Y|\mathcal{M}_j] = \int [Y|\Theta_j, \mathcal{M}_j] \times [\Theta_j|\mathcal{M}_j] d\Theta_j, \tag{2.31}$$

where Θ_j is the vector of parameters of $\mathcal{M}_j$; $[Y|\Theta_j, \mathcal{M}_j]$ is the likelihood under model $\mathcal{M}_j$; the prior distribution of Θ_j is denoted by $[\Theta_j|\mathcal{M}_j]$ and $[\mathcal{M}_j]$ is the prior probability of $\mathcal{M}_j$. Thus, we clearly see with these formulas how Bayesian model averaging takes into account uncertainties about the model form. The posterior mean of Δ is likewise a weighted average of the posterior means in the separate component models,

$$\mathbb{E}[\Delta|Y] = \sum_{i=1}^{n}[\mathcal{M}_j|Y] \times \mathbb{E}[\Delta|\mathcal{M}_j, Y]. \tag{2.32}$$

Similarly, the posterior variance may be expressed via the formula:

$$\mathbb{V}[\Delta|Y] = \sum_{j=1}^{J}[\mathcal{M}_j|Y] \times \left(\mathbb{V}[\Delta|\mathcal{M}_j, Y] + \left(\mathbb{E}[\Delta|\mathcal{M}_j, Y] - \mathbb{E}[\Delta|Y]\right)^2\right). \tag{2.33}$$

Based on the ample literature on *model-averaging*, we expect this technique to provide ameliorated predictions and a more realistic estimate of the uncertainty related to model predictions than any single model. In addition, the BMA can be used for a direct model comparison by computing model posterior probabilities which are equal to model weights [Hoeting et al., 1999, Draper, 1995, Wintle et al., 2003].

2.2.3.3. Key Points of Model Selection and *Model-averaging*

- Models with various structures are likely to be considered as candidate models for a given epidemics.
- When the goal of the study is to make inferences and draw predictions, the use of a single model is prone to prediction error because this model might not take into account crucial drivers of the dynamics. This limitation can be circumvented by considering a family of candidate models and applying a model selection strategy or a model aggregation strategy.
- Model selection strategy is an intuitive, relatively fast strategy. However, ignoring uncertainties involved in model selection leads to over optimistic outputs.
- BMA is a computer-intensive technique that is expected to give more realistic inferences and ameliorated predictions than a single model.

3. A mechanistic-statistical approach to infer for spatio-temporal population dynamics

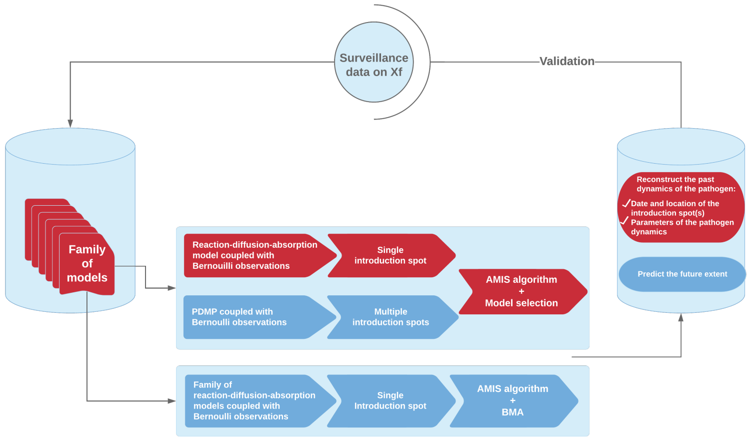
This chapter introduces a published article[1] cited hereafter:

Table of contents

[1] C. Abboud, O. Bonnefon, E. Parent, and S. Soubeyrand. Dating and localizing an invasion from post-introduction data and a coupled reaction–diffusion–absorption model. *Journal of Mathematical Biology*, 2019a

3.1. Graphical Summary

Major Components of The Chapter



Surveillance data on Xf

Validation

Family of models

Reaction-diffusion-absorption model coupled with Bernouilli observations

Single introduction spot

PDMP coupled with Bernoulli observations

Multiple introduction spots

AMIS algorithm + Model selection

Reconstruct the past dynamics of the pathogen:

✓Date and location of the introduction spot(s)

✓Parameters of the pathogen dynamics

Predict the future extent

Family of reaction-diffusion-absorption models coupled with Bernoulli observations

Single Introduction spot

AMIS algorithm + BMA

How to · · ·

I- Biological Questions

- Tackle the problem of recovering the location and time of the introduction of a pathogen for which only post-introduction data are available.
- Get more insights on the pathogen epidemiology in order to adapt surveillance strategies.

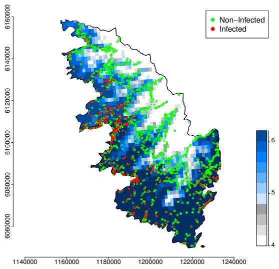
II- Methodological Questions

- Estimate the initial conditions using observational partial data.
- Jointly estimate the introduction spot and the propagation characteristics, because these characteristics link the introduction to the observation.
- Choose an adequate statistical inference procedure that is capable of ensuring the adequacy between model and data in a reasonable time span.

Methodological Ingredients

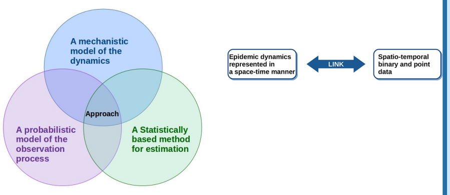
Post-introduction Data on Xf

- ∼**8000** plants sampled **between 2015-2017** of which **800** have been diagnosed as **infected** (real-time PCR).
- For those ∼ 8000 plants, **geographic coordinates** and **sampling dates** are available.
- T : average of the minimum daily temperature over January and February b/n 1995 and 2003 (Map of T with 1 km grid resolution on the right).



Mechanistic-statistical Approach

- Couple the reaction-diffusion-absorption model with the probabilistic model describing the observation process.
- Use Bayesian inference procedure via the AMIS algorithm.
- Jointly infer initial conditions and parameters of the dynamics and select the threshold of temperature using various information criteria.



A mechanistic model of the dynamics

A probabilistic model of the observation process

A Statistically based method for estimation

Approach

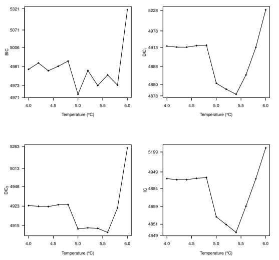
Epidemic dynamics represented in a space-time manner

LINK

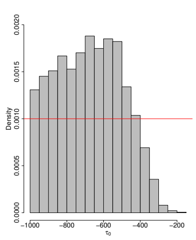
Spatio-temporal binary and point data

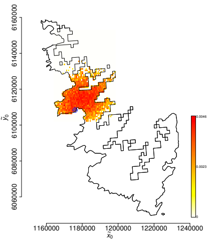
Main Results

Selection of the threshold of temperature

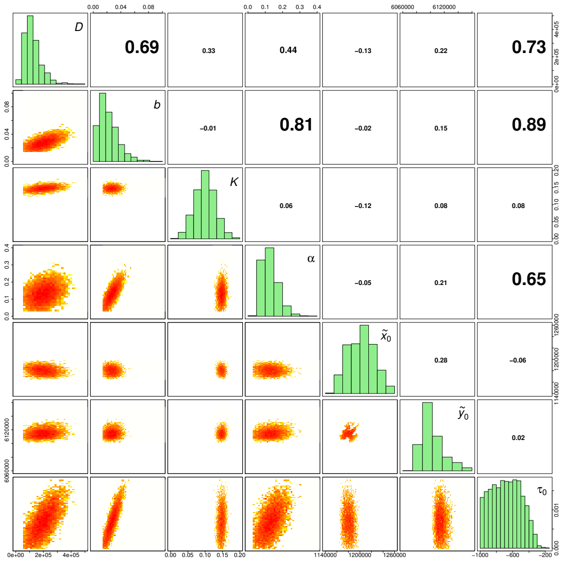


Estimation of the marginal posterior distribution of the date and site of the introduction point





Estimation of the marginal posterior distribution of parameters (panels in the diagonal) and 2D posterior distributions of parameters (lower triangle panels). Correlation cœfficients are provided in the upper triangle panels.



Conclusion & Perspectives

Conclusion

- Initial conditions and model parameters related to diffusion, reproduction, and mortality are jointly estimated in a Bayesian framework to assess for parameter uncertainties. The threshold of temperature was selected using different selection criteria.
- Goodness-of-fit tests have been conducted to check the adequacy between the model and the observed data.
- The conducted analysis tend to show that the introduction of Xf in South Corsica has probably occurred near Ajaccio in 1959, long time before its first detection. Results obtained for the introduction date are consistent with the results in the literature.

Perspectives

- Use alternative representation of disease propagation to account for more epidemiological and environmental drivers such as long distance dispersal events and seasonality.
- Refine the definition of the spatial partition by using additional relevant environmental variables, other than the daily winter temperature.
- Incorporate into the model the possibility of multiple introductions of the pathogen.
- Account for uncertainties about model forms.

3.2. Article

Journal of Mathematical Biology
https://doi.org/10.1007/s00285-019-01376-x

Mathematical Biology

Dating and localizing an invasion from post-introduction data and a coupled reaction–diffusion–absorption model

Candy Abboud[1] · Olivier Bonnefon[1] · Eric Parent[2,3] · Samuel Soubeyrand[1]

Received: 16 July 2018 / Revised: 17 April 2019

Abstract
Invasion of new territories by alien organisms is of primary concern for environmental and health agencies and has been a core topic in mathematical modeling, in particular in the intents of reconstructing the past dynamics of the alien organisms and predicting their future spatial extents. Partial differential equations offer a rich and flexible modeling framework that has been applied to a large number of invasions. In this article, we are specifically interested in dating and localizing the introduction that led to an invasion using mathematical modeling, post-introduction data and an adequate statistical inference procedure. We adopt a mechanistic-statistical approach grounded on a coupled reaction–diffusion–absorption model representing the dynamics of an organism in an heterogeneous domain with respect to growth. Initial conditions (including the date and site of the introduction) and model parameters related to diffusion, reproduction and mortality are jointly estimated in the Bayesian framework by using an adaptive importance sampling algorithm. This framework is applied to the invasion of *Xylella fastidiosa*, a phytopathogenic bacterium detected in South Corsica in 2015, France.

Keywords Partial differential equation · Reaction–diffusion · Diffusion–absorption · Bayesian inference · Mechanistic-statistical approach · Biological invasions · Disease dynamics · *Xylella fastidiosa*

Mathematics Subject Classification 62F15 · 65M06 · 35K10

This research was funded by a Ph.D. grant INRA-Région PACA (Emplois Jeunes Doctorants 2016–2019), the HORIZON 2020 XF-ACTORS Project SFS-09-2016 and the INRA-DGAL Project 21000679. We thank DGAL, Anses, SRAL, FREDON, LNR-LSV and certified laboratories for data collection and data availability. We thank Afidol for their endorsement in the Ph.D. grant.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00285-019-01376-x) contains supplementary material, which is available to authorized users.

Candy Abboud
candy.abboud@inra.fr

Extended author information available on the last page of the article

Published online: 16 May 2019

Springer

1 Introduction

Biological invasions have long been an important topic for biologists and mathematicians because of their impact on the environment, indigenous species, and health of humans, animals and plants (Andow et al. 1990, 1993; Baker 1991; Hengeveld 1989; Kermack and McKendrick 1927; Richardson and Bond 1991; Simberloff 1989; Anderson et al. 1996; Shigesada and Kawasaki 1997; Weinberger 1978). Biological invasions are generally viewed as the result of a process with four stages: arrival, establishment, spread and concentration (Reise et al. 2006; Vermeij 1996). Each stage of the invasion process has been a core topic in mathematical modeling since the mid-twentieth century (Fisher 1937; Mollison 1977; Okubo 1980; Shigesada et al. 1995; Skellam 1951), and better understanding processes governing invasions is chiefly relevant for improving surveillance and control strategies. In particular, extensive researches have been conducted in the intents of reconstructing the past dynamics (Boys et al. 2008; Roques et al. 2016; Soubeyrand and Roques 2014) of alien species and predicting their future spatial extents (Chapman et al. 2015; Peterson et al. 2003). In this context, partial differential equations offer a rich and flexible framework that has been applied to a large number of invasions (Gatenby and Gawlinski 1996; Lewis, MA and Kareiva, P 1993; Murray 2002; Okubo and Levin 2002; Turchin 1998). Even though a partial differential equation does not describe all the processes involved in an ecological dynamics, it can help in understanding its important properties and inferring its major components, such as dates and sites of invasive-species introductions.

Consider as an example the emergence of *Xylella fastidiosa* (Xf), a phytopathogenic bacterium detected in South Corsica, France, in 2015 and currently present in a large part of this island (Denancé et al. 2017b; Soubeyrand et al. 2018). This plant pathogen has the potential to cause a major sanitary crisis in France, typically like in Italy, where a large number of infected olive trees dried and died, causing serious damages to olive cultivation. To avoid such a situation, the French General Directorate of Food (DGAL) implemented enhanced control and surveillance measures after the first *in situ* detection of Xf in Corsica, which generated a data set consisting of a spatio-temporal point pattern (i.e. the locations and dates of plant samples) marked by a binary variable indicating the result of the diagnostic test (i.e. indicating if the plant sample is positive or negative to Xf).

In this example, only post-introduction data are available (i.e. data collected over a temporal window covering a period after the introduction time), and we precisely propose in this article an approach for estimating the date and the site of the introduction using such observational partial data. It has however to be noted that estimating the introduction point from post-introduction data requires the estimation of the propagation characteristics of the invasive species (and *vice versa*) because these characteristics link the introduction and the observations. Thus, in this paper, we aim at jointly estimating the date and site of the introduction, and other parameters related to growth, dispersal and death that govern the post-introduction dynamics.

Such a joint estimation was proposed by Soubeyrand and Roques (2014) with a simple reaction–diffusion model and was applied to simulated data. It was developed in a mechanistic-statistical framework that has often been used to describe and infer ecological processes. This framework combines a mechanistic model for the dynamics of interest, a probabilistic model for the observation process and a statistical procedure for estimating model parameters (Berliner 2003; Lanzarone et al. 2017; Roques et al. 2011; Soubeyrand et al. 2009a, b; Wikle 2003a, b). We adapted this framework for dating and localizing the introduction of an invasive species by taking into account spatial heterogeneities in growth and mortality. Precisely, we built a mechanistic model yielding the probability for the invasive species to occupy any spatial units at any time. This spatio-temporal function, with values in [0, 1], satisfies (i) a reaction–diffusion equation that describes the spread of the alien species in a sub-domain of the study domain and (ii) a diffusion–absorption equation that describes the dispersal and the death of the alien species in the complementary sub-domain. Typically, the partition into the two sub-domains can be determined by environmental variables affecting the growth and mortality of the invasive species (e.g. host/non-host environment, low/high winter temperature, and presence/absence of nutrients). In addition, our model assumes that there is only one introduction point (in time and space) that governs the emergence of the invasive species and that eventual other introduction points have negligible effects on the dynamics.

Estimation of model parameters, including the time and the location of the introduction, is carried out in the Bayesian framework with the adaptive multiple importance sampling algorithm (AMIS; Cornuet et al. 2012). Our main motivation for using AMIS is the gain in computation time with respect to Markov chain Monte Carlo (MCMC) often used in the mechanistic-statistical framework (see references above). From an example in population genetics, Cornuet et al. (2012) observed that AMIS was 6 times faster than MCMC for providing similar posteriors with a slightly better repeatability in the case of AMIS (without parallelization). The authors mentioned that AMIS is particularly interesting in cases where the likelihood is computationally expensive (like in our case) because all particles simulated during the process are recycled, which minimizes the numbers of calls of the likelihood function. In addition, like other adaptive importance sampling algorithms (Bugallo et al. 2015), AMIS can be easily parallelized and its tuning parameters are automatically adapted across the algorithm iterations.

In our framework, the two sub-domains, where the reaction–diffusion and diffusion–absorption equations are defined, are obtained by thresholding a spatial variable. The threshold value is determined with a selection criterion. Four criteria are considered: the Bayesian information criterion (BIC; Schwarz et al. 1978), two versions of the deviance information criteria (DIC; Gelman et al. 2003; Spiegelhalter et al. 2002) and a predictive information criterion (IC; Ando 2011). In the Xf case study, the two sub-domains are defined by thresholding the average of the minimum daily temperature in January and February, the two coldest months of the year in Corsica. Indeed, winter temperature has been inferred as an important environmental factor governing the dynamics of Xf and the level of disease severity caused by Xf (Costello et al. 2017; Feil et al. 2003; Feil and Purcell 2001; Henneberger 2003; Purcell 1977; Purcell et al. 1980). For instance, isolines for the average minimum daily temperature in January have been shown to be quite consistent with regions in the United States that are exposed to different levels of severity of the Pierce’s disease of grape caused by Xf (Anas et al. 2008).

The paper is structured as follows. The hierarchical modeling framework coupling a partial differential equation and a Bernoulli observation is described in Sect. 2. Bayesian inference for parameter estimation grounded on the AMIS algorithm and model selection are also presented in this methodological section. The results obtained from surveillance data for Xf in the case study (Corsica) are provided in Sect. 3. In Sect. 4, we summarize and discuss our work.

2 The mechanistic-statistical approach

2.1 Process model

Models based on parabolic partial differential equations have often been used to describe biological invasions (Skellam 1951; Shigesada et al. 1995; Shigesada and Kawasaki 1997; Okubo 1980). Here, we are interested in the invasion of a pathogen, that spreads in a domain Ω included in $\mathbb{R}^2$. We assume that there is only one single introduction point in time and space that triggered the invasion and that eventual subsequent introductions have negligible effects on the dynamics and are therefore not incorporated into the model. Furthermore, to account for spatial heterogeneity in the reproduction regime of the pathogen, we divide Ω into two sub-domains, say Ω_1 and Ω_2, such that $\Omega = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 = \emptyset$ and different growth terms apply to Ω_1 and Ω_2.

More formally, the spread of the pathogen is described by a coupled model governing the probability $u(t, \mathbf{x})$ of a host located at site $\mathbf{x} = (x_1, x_2) \in \Omega$ to be infected at time t. This model is grounded on two particular types of parabolic partial differential equations: (i) a reaction–diffusion equation in Ω_1 where the growth is logistic (Verhulst 1838) and (ii) a diffusion–absorption equation in Ω_2 where only dispersal and death events occur. The probability $u(t, \mathbf{x})$ satisfies:

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\Delta u + bu\left(1 - \dfrac{u}{K}\right)\mathbb{1}(\mathbf{x} \in \Omega_1) - \alpha u \mathbb{1}(\mathbf{x} \in \Omega_2), & t \geq \tau_0,\ \mathbf{x} \in \Omega, \\ \nabla u(t, \mathbf{x}).n(\mathbf{x}) = 0, & t \geq \tau_0,\ \mathbf{x} \in \partial\Omega, \\ u(\tau_0, \mathbf{x}) = u_0(\mathbf{x}) \geq 0, & \mathbf{x} \in \Omega, \end{cases} \tag{1}$$

where $D > 0$ is the diffusion coefficient; b corresponds to the intrinsic growth rate of the pathogen infection in Ω_1; $K \in (0, 1]$ is a plateau for the probability of infection (i.e. an analogue to the carrying capacity of the environment); α is the decrease rate of the infection in Ω_2; $\Delta = \dfrac{\partial^2}{\partial x_1^2} + \dfrac{\partial^2}{\partial x_2^2}$ is the 2-dimensional diffusion operator of Laplace; $\mathbf{x} \mapsto \mathbb{1}(\mathbf{x} \in \Omega_i)$ is the characteristic function taking the value 1 if $\mathbf{x} \in \Omega_i$ and 0 otherwise; $\tau_0 \in \mathbb{R}$ is the introduction time of the pathogen. As explained in the introduction, the sub-domains Ω_1 and Ω_2 are defined by thresholding a spatial function, say T, with the threshold value $\tilde{T}$ that is hold fixed: $\Omega_1 = \Omega_1(T, \tilde{T}) = \{\mathbf{x} \in \Omega : T(\mathbf{x}) > \tilde{T}\}$ and $\Omega_2 = \Omega_2(T, \tilde{T}) = \{\mathbf{x} \in \Omega : T(\mathbf{x}) \leq \tilde{T}\}$.

In our framework, the initial condition u_0 models the introduction of the pathogen in the study domain. Here, the introduction represents the initial phase of the outbreak corresponding to the arrival of the pathogen and its local establishment. Thus, u_0 is not expressed as a Dirac delta function but as a kernel function centered around the central point of the introduction $\tilde{\mathbf{x}}_0 = (\tilde{x}_0, \tilde{y}_0) \in \Omega$. More precisely, the probability of a host at $\mathbf{x}$ to be infected at τ_0 satisfies:

$$u_0(\mathbf{x}) = p_0 \exp\left(-\frac{\|\mathbf{x} - \tilde{\mathbf{x}}_0\|^2}{2\sigma^2}\right) \tag{2}$$

where p_0 is the infection probability at $(\tau_0, \tilde{\mathbf{x}}_0)$, $\sigma^2 = \frac{r_0^2}{q}$, q is the 0.95-quantile of the χ^2 distribution with two degrees of freedom, and r_0 is the *radius* of the kernel. Thus, at τ_0, if we neglect border effects, 95% of the infected plants are located within the ball with center $\tilde{\mathbf{x}}_0$ and radius r_0. Assuming in addition reflecting conditions on the boundary $\partial\Omega$ of Ω, the system of equations (1) is well-posed (Evans 1998). In addition, by constraining p_0 in $[0, K]$, the principle of parabolic comparison (Protter, MH and Weinberger, HF 1967) implies that the solution of (1) is also in the interval $[0, K]$.

Remark We adopted a parsimonious approach consisting of modeling the probability of a host to be infected (i.e., the local quantity of infected host units over the local total quantity of host units) instead of the dynamics of the pathogen in the host population (i.e., the local quantities of susceptible, exposed, infectious and removed host units). This choice allowed us, in particular, to ignore eventual spatial heterogeneity in host abundance and to reduce the number of unknown parameters.

2.2 Data model

Let $t_i \in \mathbb{R}$ denote the sampling time of host $i \in \{1, \ldots, I\}$, $I \in \mathbb{N}^*$, $\mathbf{x}_i \in \Omega$ its location and $Y_i \in \{0, 1\}$ its sanitary status observed at time t_i (1 for infected, 0 for healthy). Conditionally on u, T and $\{(t_i, \mathbf{x}_i) : 1 \le i \le I\}$, the sanitary statuses Y_i, $i \in \{1, \ldots, I\}$, are assumed to be independent random variables following Bernoulli distributions with success probability $u(t_i, \mathbf{x}_i)$:

$$Y_i \mid u, T, \{(t_i, \mathbf{x}_i) : 1 \le i \le I\} \underset{\text{indep.}}{\sim} \text{Bernoulli}(u(t_i, \mathbf{x}_i)), \tag{3}$$

where u depends on parameters D, b, K, α, τ_0, $\tilde{\mathbf{x}}_0$, r_0, p_0 and $\tilde{T}$.

Remark This simple data model could be modified to account for factors classically encountered in epidemiology, e.g. false-positive and false-negative observations, and spatial and temporal dependencies not accounted for in the process model. In the real case study tackled in this article, each observed host was sampled only once. In a case where hosts could be sampled several times, a temporal dependence should be introduced in the observation process to account for, e.g., the within-host persistence of the pathogen.

2.3 Parameter estimation with an adaptive importance sampling algorithm

Inference about the parameter vector $\Theta = (D, b, K, \alpha, \tau_0, \tilde{\mathbf{x}}_0, r_0, p_0)$ is made in the Bayesian framework, which technically consists in assessing the posterior distribution $[\Theta|Y]$ of Θ conditional on sanitary statuses $Y = \{Y_i : 1 \leq i \leq I\}$. The parameter $\tilde{T}$ will be treated later in Sect. 2.4 via model selection. Philosophically, a posterior probability is to be interpreted as a coherent judgment quantifying a subjective degree of uncertainty (Lindley 2006).

In what follows, we will keep using Gelfand's bracket notations for probability distributions (Gelfand and Smith 1990). The posterior distribution of the unknown, hereafter dubbed Θ, is derived by Bayes' rule:

$$[\Theta|Y] = \frac{[Y|\Theta] \times [\Theta]}{[Y]},$$

where

$[Y|\Theta]$ is the conditional distribution of the data Y given the unknown Θ (i.e. the likelihood function of the model) that satisfies [using Eq. (3)]:

$$[Y|\Theta] = \prod_{i=1}^{I} u(t_i, \mathbf{x}_i)^{Y_i} (1 - u(t_i, \mathbf{x}_i))^{1-Y_i}; \tag{4}$$

$[\Theta]$ is the prior distribution of Θ that depends on the application and that will be specified in Sect. 3; the distribution of Y, $[Y] = \int [Y|\Theta][\Theta] d\Theta$, may be a formidable integral, depending on the dimension of the unknown Θ. However, modern Bayesian algorithms (Brooks 2003) avoid its computation by making recourse to Monte Carlo techniques only based on the un-normalized probability function $[Y|\Theta] \times [\Theta]$. Yet, the computation of $[Y|\Theta]$ itself requires the value of the solution u of Eq. (1) for any valid parameter vector Θ. This equation admits a unique solution for any fixed and valid Θ, but cannot be solved analytically. Hence, we make recourse to a standard finite-element method with the software `Freefem++` (Hecht 2012); see Sect. 2.5.

For the mechanistic-statistical model defined above, the posterior distribution $[\Theta|Y]$ cannot be expressed analytically due to its intractable normalizing constant, but one can draw a sample under this distribution using an adequate algorithm for Bayesian inference. The so-called posterior sample $[\Theta|Y]$ is then used to numerically characterize all that we know about Θ after data assimilation. Here, we use the adaptive multiple importance sampling (AMIS; Cornuet et al. 2012) algorithm, that consists of iteratively generating parameter vectors under an adaptive proposal distribution and assigning weights to the parameter vectors. To design efficient importance sampling algorithms, the auxiliary proposal distribution should be chosen as close as possible to the posterior distribution. However, the posterior distribution being unknown, the crucial choice of the proposal is a difficult task (Gelman et al. 1996; Roberts et al. 1997). The main aim of the AMIS algorithm is to overcome this difficulty by tuning the coefficients of the proposal distribution picked in a parametric family of distributions, generally the Gaussian one, at the end of each iteration.

In this framework, at each iteration, new coefficient values for the proposal distribution are determined using the current weighted posterior sample (Bugallo et al. 2015), then the posterior sample is augmented by generating new replicates from the newly tuned proposal distribution and the weights of the cumulated posterior sample are recomputed. The algorithm can be described as follows:

1. Set initial values μ_0 and Σ_0 for the mean vector and the variance matrix of the multi-normal proposal distribution $\mathcal{N}(\mu_0, \Sigma_0)$, whose probability density function is denoted by $\Theta \to g_{\mu_0,\Sigma_0}(\Theta)$.
2. At iteration $m = 1, \cdots, M$,
   (a) Generate a new sample $\{\Theta_m^l : l = 1 \cdots, L\}$ from the proposal distribution $\mathcal{N}(\mu_{m-1}, \Sigma_{m-1})$.
   (b) Compute the un-normalized importance weights for the new sample as in Eq. (5), and update the un-normalized weights for the previously generated samples as in Eq. (6):

$$\tilde{w}_m^l = \frac{[Y|\Theta_m^l] \times [\Theta_m^l]}{\frac{1}{m}\sum_{j=1}^{m} g_{\mu_{j-1},\Sigma_{j-1}}(\Theta_m^l)}, \quad l = 1, \cdots, L \tag{5}$$

$$\tilde{w}_\varepsilon^l = \frac{[Y|\Theta_\varepsilon^l] \times [\Theta_\varepsilon^l]}{\frac{1}{m}\sum_{j=1}^{m} g_{\mu_{j-1},\Sigma_{j-1}}(\Theta_\varepsilon^l)}, \quad \varepsilon = 1, \cdots, m-1, \ l = 1, \cdots, L, \tag{6}$$

   where $g_{\mu_{j-1},\Sigma_{j-1}}$ is the probability density function of the multi-normal distribution with mean vector μ_{j-1} and variance matrix Σ_{j-1}.
   (c) Normalize the weights:

$$w_\varepsilon^l = \frac{\tilde{w}_\varepsilon^l}{\sum_{i=1}^{m}\sum_{j=1}^{L} \tilde{w}_i^j}, \quad \varepsilon = 1, \cdots, m, \ l = 1, \cdots, L.$$

   (d) Adapt coefficient values for the next proposal distribution as follows:

$$\mu_m = \sum_{l=1}^{L}\sum_{\varepsilon=1}^{m} w_\varepsilon^l \Theta_\varepsilon^l$$

$$\Sigma_m = \sum_{l=1}^{L}\sum_{\varepsilon=1}^{m} w_\varepsilon^l (\Theta_\varepsilon^l - \mu_\varepsilon)(\Theta_\varepsilon^l - \mu_\varepsilon)^t.$$

The AMIS algorithm provides a weighted posterior sample $\{\{\Theta_m^l, w_m^l\}_{l=1}^L\}_{m=1}^M$ of size ML, which provides an empirical approximation of the posterior distribution $[\Theta|Y]$. Conditions leading to the convergence in probability of the posterior mean of any function (integrable with respect to the posterior distribution) of the parameters are described in Cornuet et al. (2012) and are satisfied in our case.

If in practice, the convergence of AMIS to the true posterior cannot be numerically demonstrated (because the true posterior is not known), one can assess its stabilization by evaluating the variation in the following deviation measure between the assessments of the posterior distribution at iteration $m-1$ and $m>1$:

$$\mathcal{M}_{\mathcal{G}}(m-1, m) = \max_{c \in \mathcal{G}} |p_m(c) - p_{m-1}(c)|,$$

where $p_m(c)$ denotes the assessment at iteration m of the posterior probability that Θ is in the sub-domain $c \subset \mathbb{R}^8$ of the parameter space, i.e.

$$p_m(c) = \sum_{m'=1}^{m} \sum_{l=1}^{L} w_{m'}^{l} \mathbb{1}(\Theta_{m'}^{l} \in c),$$

and $\mathcal{G}$ is a partition of a sub-space of the parameter space. The definition of $\mathcal{G}$ depends on the application and will be given in the Results section.

We implemented AMIS in the `R` statistical software, except for solving the PDE, which was performed by calling the `Freefem++` software from `R` each time a new parameter vector was proposed. Parallel computation was performed: the estimation procedure for a fixed value of $\tilde{T}$ took approximately 1.75 days with $(M, L) = (50, 10^4)$ and the use of 100 computer cores.

2.4 Choice of $\tilde{T}$ with a model selection procedure

Implementation constraints concerning the partition of the study domain which depends on the threshold $\tilde{T}$, led us to proceed by two separate steps: (i) to infer model parameters for different fixed values of $\tilde{T}$ and, then, (ii) to select the value of $\tilde{T}$ having the largest support of data (this amounts to selecting a model within a class of models characterized by $\tilde{T}$). Thus, for each element $\tilde{T}_a$ in $\{\tilde{T}_1, \ldots, \tilde{T}_A\} \subset \mathbb{R}^A$, $A \in \mathbb{N}^*$, we carried out the estimation procedure described in Sect. 2.3 by instancing $\tilde{T}$ at the value $\tilde{T}_a$ and letting it fixed. Then, the best value of $\tilde{T}$ is chosen by minimizing some criteria classically used for model selection: here we rely on the Bayesian Information criterion (BIC; Schwarz et al. 1978), two Deviance information criteria (DIC; Spiegelhalter et al. 2002; Gelman et al. 2003) and a predictive Information Criterion (IC; Ando 2011). We use different selection criteria in order to report the variability of the selected $\tilde{T}$ when different hypotheses are made about which the best model is, if any.

The BIC satisfies:

$$\mathrm{BIC} = -2\log[Y|\hat{\Theta}] + k \log I, \tag{7}$$

where I is the sample size, k is the number of model parameters (in our setting, k is the same for all the models), and $\hat{\Theta}$ is the maximum likelihood estimate of the parameter vector Θ in the support $\mathcal{S}(\Theta; \tilde{T}_a)$ of Θ defined by the prior distribution (in our setting, this support depends on the fixed value $\tilde{T}_a$ of $\tilde{T}$):

$$\hat{\Theta} = \operatorname*{argmax}_{\Theta \in \mathcal{S}(\Theta; \tilde{T}_a)} [Y|\Theta].$$

The DIC satisfies:

$$\mathrm{DIC} = \bar{\mathcal{D}} + p_{\mathrm{eff}}, \tag{8}$$

where $\bar{\mathcal{D}}$ is the posterior mean of the deviance $\mathcal{D}(\Theta) = -2\log[Y|\Theta] + C$ (where C is a constant that cancels out when one compares different models) and p_{eff} is the effective number of parameters of the model. The difference in the two versions of the DIC considered here lies in the calculation of p_{eff}. In the first version proposed by Spiegelhalter et al. (2002),

$$p_{\mathrm{eff}} = p_{\mathcal{D}} = \bar{\mathcal{D}} - \mathcal{D}(\bar{\Theta}), \tag{9}$$

where $\bar{\Theta}$ is the posterior mean of Θ: $\bar{\Theta} = \mathbb{E}[\Theta|Y]$. In the second version proposed by Gelman et al. (2003),

$$p_{\mathrm{eff}} = \frac{1}{2}\mathbb{V}(\mathcal{D}(\Theta)|Y), \tag{10}$$

where $\mathbb{V}(\mathcal{D}(\Theta)|Y)$ is the posterior variance of $\mathcal{D}(\Theta)$. The IC of Ando (2011), which is supposed to solve over-fitting issues, satisfies:

$$\mathrm{IC} = \bar{\mathcal{D}} + 2p_{\mathcal{D}} := 3\bar{\mathcal{D}} - 2\mathcal{D}(\bar{\Theta}). \tag{11}$$

In practice, the different terms appearing in the four criteria, namely $\hat{\Theta}$, $\bar{\Theta}$, $\bar{\mathcal{D}}$ and $\mathbb{V}(\mathcal{D}(\Theta)|Y)$, are replaced by their empirical values using the weighted posterior sample $\{\{\Theta_m^l, w_m^l\}_{l=1}^L\}_{m=1}^M$ provided by the application of the AMIS algorithm.

2.5 Numerical equation solving

For the application, computations for solving the PDE were carried out with the software `Freefem++` (Hecht 2012). A Finite Element Method was used. The nonlinearity has been treated with a Newton-Raphson algorithm applied to the variational formulation of Equation (1), by instancing the criterion of convergence at the value 10^{-10}. The solution was approximated by a piecewise linear and continuous function. The time resolution was based on an adaptive step size using a backward Euler method. Supplementary Figure S1 shows the spatial discretization composed of 4791 nodes that has been used in the application in Sect. 3. With this mesh, the average computation time for one simulation is 55 s. We explored the effect of the spatial discretization by comparing the numerical solutions of the equation obtained with the 4791 nodes mesh and with a finer mesh composed of 10703 nodes. The solutions were computed for the set of parameters corresponding to the posterior maximum (Supplementary Material S4 shows the time continuous dynamics for this set of parameters). Supplementary Figure S2 shows very close simulation results for both meshes. Moreover, we investigated the numerical error of system 1 by using the indicator, norm $||u||_{H^2}$ which is classically considered to control the H^1-error (Allaire 2008). Using the mesh composed of 4791 nodes leads to a numerical error around 0.02 corresponding to a satisfying accuracy for our application.

3 Application to the dynamics of *Xylella fastidiosa* in South Corsica

3.1 Surveillance data

For this application, we use spatio-temporal binary data on the presence of *Xylella fastidiosa* (Xf) collected in South Corsica, France, from July 2015 to May 2017. Over this period, approximately 8000 plants were sampled, among which 800 have been diagnosed as infected (with a real-time polymerase chain reaction (PCR) technique; Denancé et al. 2017b). Available data for each sampled plant are its spatial coordinates, its sampling date (which is unique) and its health status at the sampling date. Coordinates and health statuses at the sampling times are shown in Fig. 1.

3.2 Model specifications

As mentioned in the introduction, we use temperature data to divide the spatial domain into two sub-domains. We exploit a freely available database (PVGIS © European Communities, 2001–2008) providing, in particular, monthly averages of the daily minimum temperature reconstructed over a grid with spatial resolution of 1×1km (Huld et al. 2006); these monthly averages correspond to the period 1995-2003, but we used them as references over the period covered by our models. We use these data to build the average of the daily minimum temperature over January and February, say $T(\mathbf{x})$ for any location $\mathbf{x}$; see Fig. 1. $T(\mathbf{x})$ is then used to split the study domain into two parts: one part where $T(\mathbf{x}) \leq \tilde{T}$ and the growth of Xf is hampered by cold winter temperatures, and the other part where $T(\mathbf{x}) > \tilde{T}$ and the growth of Xf is not hampered. The threshold value $\tilde{T}$ will be selected in the set $\{4.0, 4.2, 4.4, \ldots, 6.0\}$, in Celsius degrees. Panels of Fig. 2 display the partitioning of the study domain induced by the different values of $\tilde{T}$.

The prior distribution for Θ combines vague uniform distributions and Dirac distributions:

$$\begin{aligned}[\Theta] = &\frac{1}{(10^8 - 50) \times 100 \times 1 \times 100 \times 1000 \times |\Omega_1|}\\ &\times \mathbb{1}(D{\in}[50; 10^8], b{\in}[0; 100], K{\in}]0; 1], \alpha{\in}[0; 100], \tau_0{\in}[-1000; 0], \tilde{\mathbf{x}}_0{\in}\Omega_1)\\ &\times \mathrm{Dirac}_{5000}(r_0) \times \mathrm{Dirac}_{0.1}(p_0),\end{aligned}$$

where $|\Omega_1|$ is the area of Ω_1 and $\mathrm{Dirac}_b(B)$ is equal to 1 if $B = b$, and 0 otherwise. The Dirac distribution for $\tilde{T}$ was chosen to deal with implementation issues

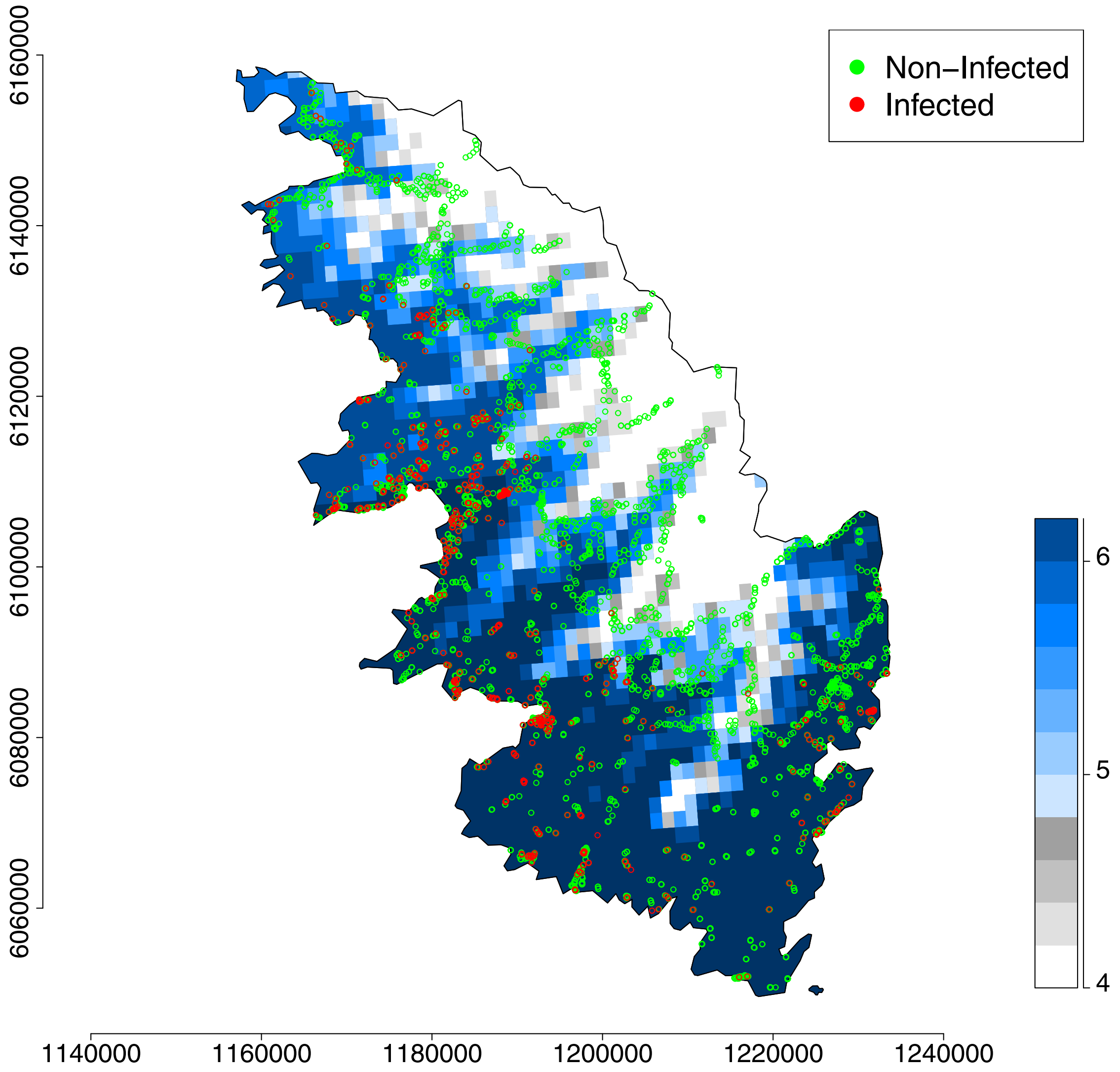


Fig. 1 Locations of plants, sampled from July 2015 to May 2017, that have been detected as positive (red dots) or negative (green dots) to *Xylella fastidiosa* in South Corsica, France, and map of the average of the daily minimum temperature (in Celsius degrees) over January and February reconstructed over a grid with spatial resolution of 1 × 1km (blue–white color palette) (color figure online)

explained in Section 2.4. We chose Dirac prior distributions for r_0 and p_0 in the aim of precisely defining what is an *introduction* (see Section 2.1) and imposing the same intensity level and spatial extent for the introduction in all the models in competition. For D, b, K and α, we specified vague uniform priors satisfying constraints of positivity. In addition, the plateau K had to be less than 1, as indicated in Sect. 2.1. For the introduction time τ_0, we chose a uniform distribution between -1000 months and 0 month before the first detection of Xf in South Corsica. Note that, using a temporal model and aggregated data, Soubeyrand et al. (2018) inferred an introduction date around -360 months before the first detection of Xf in South Corsica. Finally, the introduction location $\tilde{\mathbf{x}}_0$ was supposed to be uniformly distributed in Ω_1, the sub-domain where the conditions are favorable for the expansion of Xf.

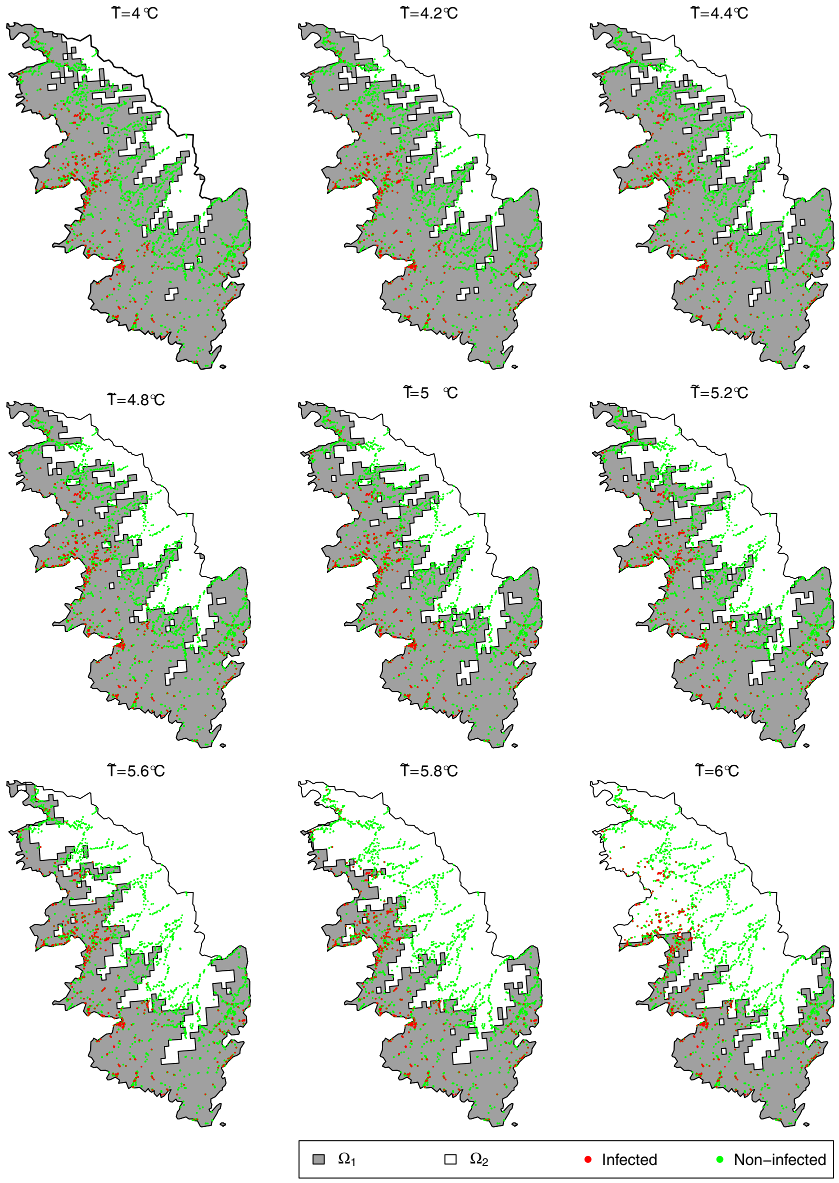


Fig. 2 Partition of the study domain Ω into the sub-domains Ω_1 and Ω_2 with respect to the value of $\tilde{T}$ in $\{4.0, 4.2, 4.4, \ldots, 6.0\}$, in Celsius degrees. Red and green dots give the locations of infected and non-infected samples (color figure online)

3.3 Selection of the temperature threshold

The spatio-temporal models corresponding to different values of $\tilde{T}$ ranging from 4 to 6°C were fitted to data using the estimation approach presented in Sect. 2.3 (with $(M, L) = (50, 10^4)$) and were compared with the four selection criteria introduced in Sect. 2.4. The values of the criteria are displayed in Fig. 3. The smallest BIC value was obtained for $\tilde{T} = 5.0\,°\text{C}$. The smallest DIC value based on the computation proposed by Spiegelhalter et al. (2002) and the smallest IC values were obtained for $\tilde{T} = 5.4\,°\text{C}$. The smallest DIC value based on the computation proposed by Gelman et al. (2003) was obtained for $\tilde{T} = 5.6\,°\text{C}$. Except the BIC, which only measures the adequacy between the model and data at the posterior mode of the parameter vector, each of the three other criteria takes quite close values around $\tilde{T} =$

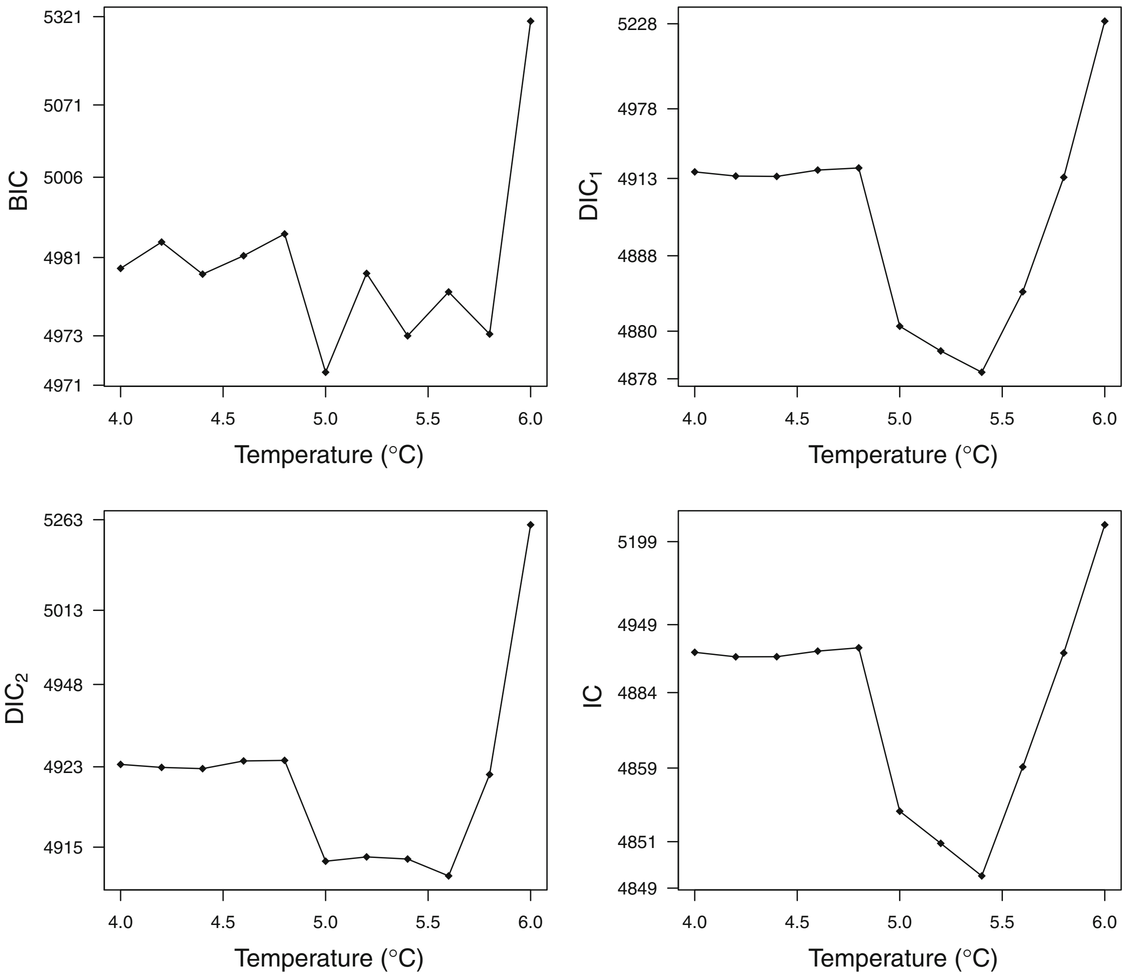


Fig. 3 Values of the four selection criteria (BIC, DIC$_1$ of Spiegelhalter et al. (2002), DIC$_2$ of (Gelman et al. 2003), IC of Ando (2011)) for different thresholds of temperature $\tilde{T}$ ranging from 4 to 6 °C. Non-linear transformations of the y-axis were applied to facilitate the identification of the lowest values of the criteria

5.4 °C (typically from 5.0 to 5.6 °C). In what follows, we present the results obtained with the model corresponding to the threshold $\tilde{T} = 5.4$ °C, which is a satisfying compromise.

3.4 Stabilization of the AMIS algorithm

Figure 4 gives the variation in $\mathcal{M}_{\mathcal{G}}(m-1, m)$ for different partitions $\mathcal{G}$ allowing us to assess the stabilization of all the 2D posterior distributions of parameters (see Sect. 2.3 for the definition of the deviation measure $\mathcal{M}_{\mathcal{G}}$). For each pair of parameters, $\mathcal{G}$ was defined as the set of infinite cylinders with rectangular bases whose orthogonal projection in the 2 dimensions of interest forms a 60×60 regular rectangular grid. In each dimension of interest, the endpoints of the grid were set at the minimum and maximum values of the corresponding parameter having a weight w_M^l larger than 10^{-5} (the 2D posterior distributions over these 60×60 grids are displayed in Fig. 5). Figure 4 shows the stabilization of all the 2D posterior distributions after iteration 21.

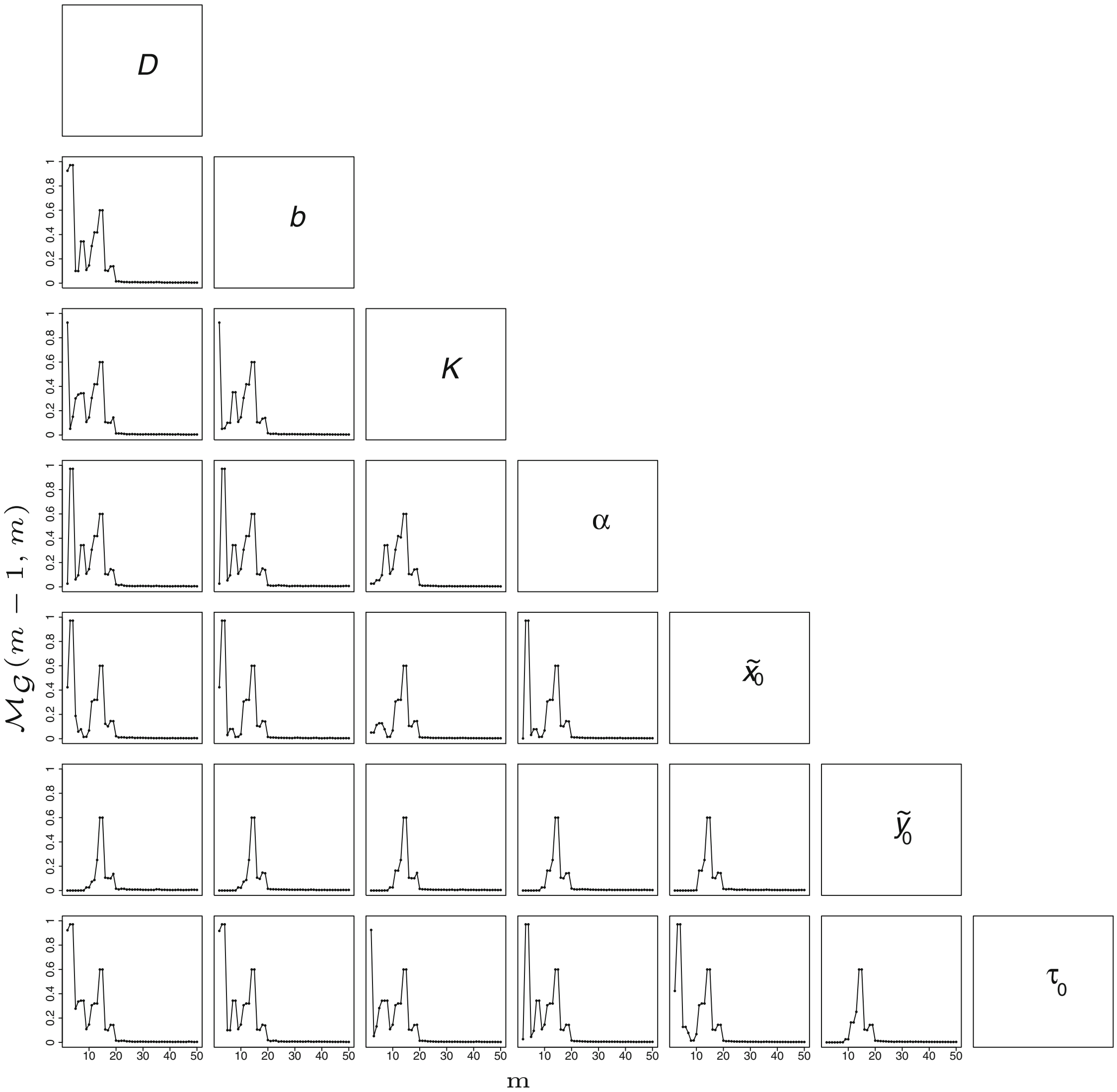


Fig. 4 Variation in the deviation measure $\mathcal{M}_{\mathcal{G}}(m-1, m)$ between the assessments of the posterior distribution at iteration $m-1$ and $m > 1$ of the AMIS algorithm. $\mathcal{M}_{\mathcal{G}}(m-1, m)$ is plotted for different partitions $\mathcal{G}$ allowing the assessment of the stabilization of all the 2D posterior distributions of parameters D, b, K, α, $\tilde{x}_0$, $\tilde{y}_0$ and τ_0

3.5 Posterior distribution of parameters

Marginal and 2D posterior distributions of parameters are displayed in Figs. 5, 6 and 7. The introduction of Xf tends to be relatively ancient (posterior median: −680 months before July 2015, i.e. introduction around 1959; posterior mean −681 months) but also relatively uncertain (posterior standard deviation: 179 months). This uncertainty has to be regarded in the light of the relatively high posterior correlation between τ_0 and the reaction–diffusion–absorption parameters D, b and α. Acquiring knowledge about D, b and α could help in eliciting informative priors for these parameters and obtain a less uncertain estimation of the introduction date. Based on our analysis, the introduction probably occurred around Ajaccio or the surrounding municipalities in the East, North and North-East (Right panel of Fig. 6). Figure 7 and Table 1 show posterior distributions and statistics of D, b, K and α. In particular, we observe that

Table 1 Posterior medians, means and standard deviations of parameters of the reaction–diffusion–absorption equation

Parameter	Unit	Median	Mean	Standard deviation
D	$\text{m}^2\,\text{month}^{-1}$	1.8×10^5	2.0×10^5	0.7×10^5
b	month^{-1}	0.026	0.027	0.008
K	probability	0.147	0.148	0.007
α	month^{-1}	0.12	0.13	0.05

the plateau for the probability of infection is around 15%. This relatively low estimate has to be considered with caution. First, it is relative to the population of plants that have been sampled. Second, it ignores the risk of false-negative observations. The inference about the diffusion parameter D allowed us to assess the length of a straight line move of the pathogen during a time unit, namely the month. This length is given by Eq. (12) (Turchin 1998; Roques et al. 2016):

$$D = \frac{(\text{length of a straight line move during one time step})^2}{4 \times \text{duration of the time step}}, \tag{12}$$

and has a posterior median equal to 155 meters per month (posterior mean: 155; posterior standard deviation: 27). These figures correspond to the move of the pathogen with different means, in particular via insects and transportation of infected plants, which are both modeled by the diffusion operator in Eq. (1).

3.6 Goodness-of-fit of the model

To check the adequacy between the selected model and observed data, we measured the accuracy of the probabilistic predictions provided by the model by using the Brier score (BS) (Brier 1950). This score is the mean of the square differences between (i) the observed health statuses Y_i^{obs}, $i = 1, \ldots, I$ (which is a realization of Y_i and takes values in $\{0, 1\}$), and (ii) the corresponding probabilities of infection $u(t_i, \mathbf{x}_i)$, which depend on Θ:

$$\text{BS} = \frac{1}{I} \sum_{i=1}^{I} \left(Y_i^{\text{obs}} - u(t_i, \mathbf{x}_i)\right)^2. \tag{13}$$

The Brier score varies between 0 and 1; lower the Brier score, better the goodness-of-fit; a systematic prediction of 0.5 leads to a Brier score equal to 0.25, which can be viewed as a threshold above which the model is clearly inadequate. In our application, the posterior median of BS is 0.0829 (95%-posterior interval: [0.0827,0.0830]).

The probabilistic predictions provided by the model can also be compared to simple but data-informed predictions via the Brier skill score (BSS):

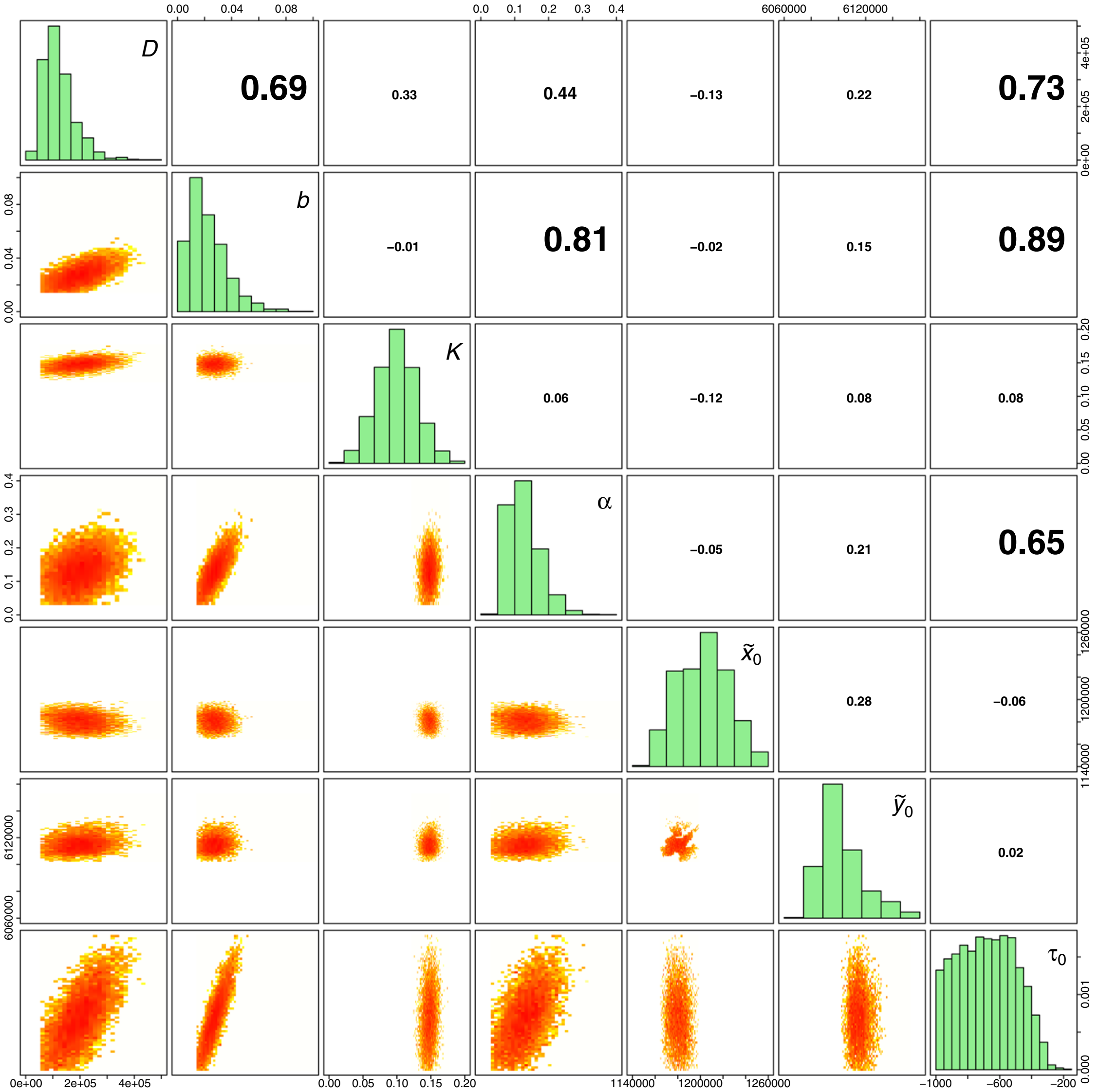


Fig. 5 Marginal posterior distributions of parameters (panels in the diagonal) and 2D posterior distributions of parameters over the 60×60 grids described in Sect. 3.4 (panels in the lower triangle). Figures in the upper triangle panels provide correlation coefficients (the larger the text size, the stronger the correlation)

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}},$$

where $\mathrm{BS}_{\mathrm{ref}}$ is the Brier score for a reference forecast. The BSS takes values between $-\infty$ and 1; A positive (resp. negative) BSS value indicates that the model-based prediction is more (resp. less) accurate than the reference forecast. The most common reference forecast is the so-called *climatology* forecast (Mason 2004) that is the mean $\bar{Y}^{\mathrm{obs}}$ of $\{Y_i^{\mathrm{obs}} : i = 1, \ldots, I\}$: $\mathrm{BS}_{\mathrm{ref}} = (1/I)\sum_{i=1}^{I}(Y_i^{\mathrm{obs}} - \bar{Y}^{\mathrm{obs}})^2$. In our application, the posterior median of BSS is 0.031 and its 95%-posterior interval is $[0.029, 0.032]$, which is entirely above zero. Hence, the model-based prediction tends to be significantly more accurate than the *climatology* forecast.

We extended the goodness-of-fit analysis by building and analyzing a local Brier score that allows us to check the adequacy of the model across space. The local Brier

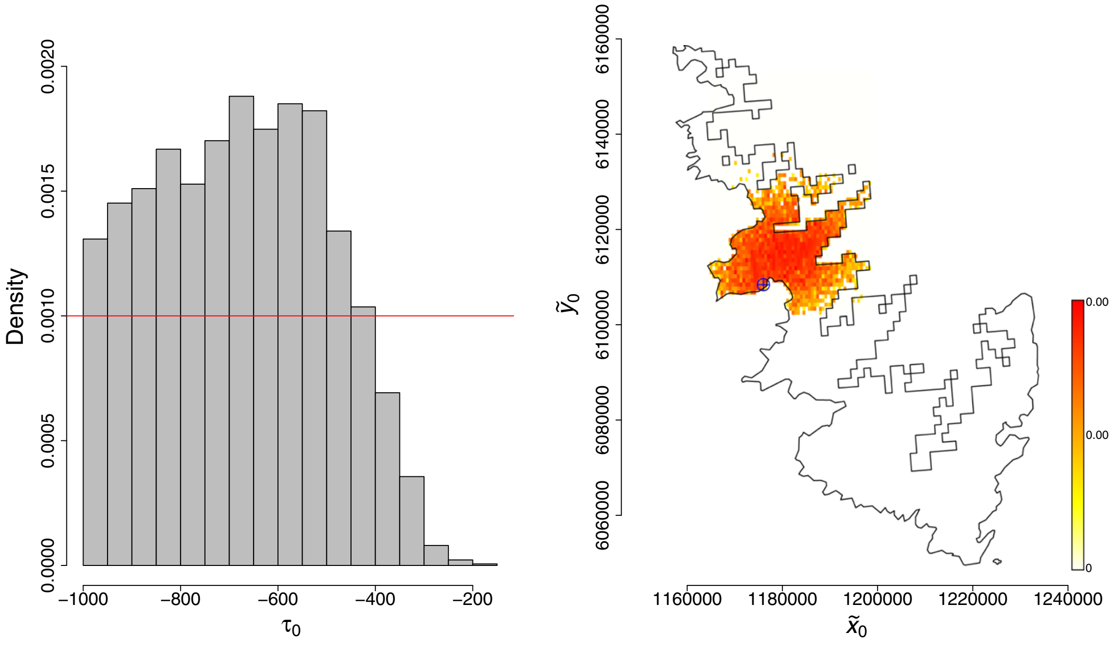


Fig. 6 Posterior distributions of the introduction time τ_0 (histogram) and the introduction point $\tilde{\mathbf{x}}_0$ (color palette). The prior for τ_0 was uniform over $[-1000, 0]$ (red line). The value of $\tilde{\mathbf{x}}_0$ having the largest weight in AMIS is indicated by a blue cross. The prior for $\tilde{\mathbf{x}}_0$ was uniform over the space delimited by the contours (color figure online)

score (LBS) computed at the location of observation $i \in \{1, \ldots, I\}$ is defined as a mean over the k-nearest neighbors:

$$\mathrm{LBS}_k(i) = \frac{1}{k+1} \sum_{i' \in \{i\} \cup \mathcal{V}_k(i)} \left(Y_{i'}^{\mathrm{obs}} - u(t_{i'}, \mathbf{x}_{i'})\right)^2, \tag{14}$$

where $\mathcal{V}_k(i)$ is the set of indices in $\{1, \ldots, I\}$ corresponding to the $k > 0$ observations nearest to $\mathbf{x}_i$ with respect to the Euclidean distance in $\mathbb{R}^2$. Figure 8 gives the distribution of the posterior means of the local Brier scores (Remark: each $\mathrm{LBS}_k(i)$ has a posterior mean because it depends on θ via the function u). 6.2% of these scores are above 0.25, which is a rather small percentage. Figure 9 displays locations where the LBS is larger than 0.25 with $k = 20$ (Supplementary Figure S3 provides similar information for k equal to 50, 100 and 150). This figure also indicates whether observations with LBS>0.25 were detected as positive or negative to Xf. None of the observations with LBS>0.25 are in Ω_2 where the growth of the pathogen is negative. Thus, discrepancies between data and the model are limited to Ω_1. In addition, in general, model discrepancies for positive samples and negative samples are located approximately at the same places. Therefore, there might be some spatially abrupt changes in the rate of infection that are not represented by our aggregated model.

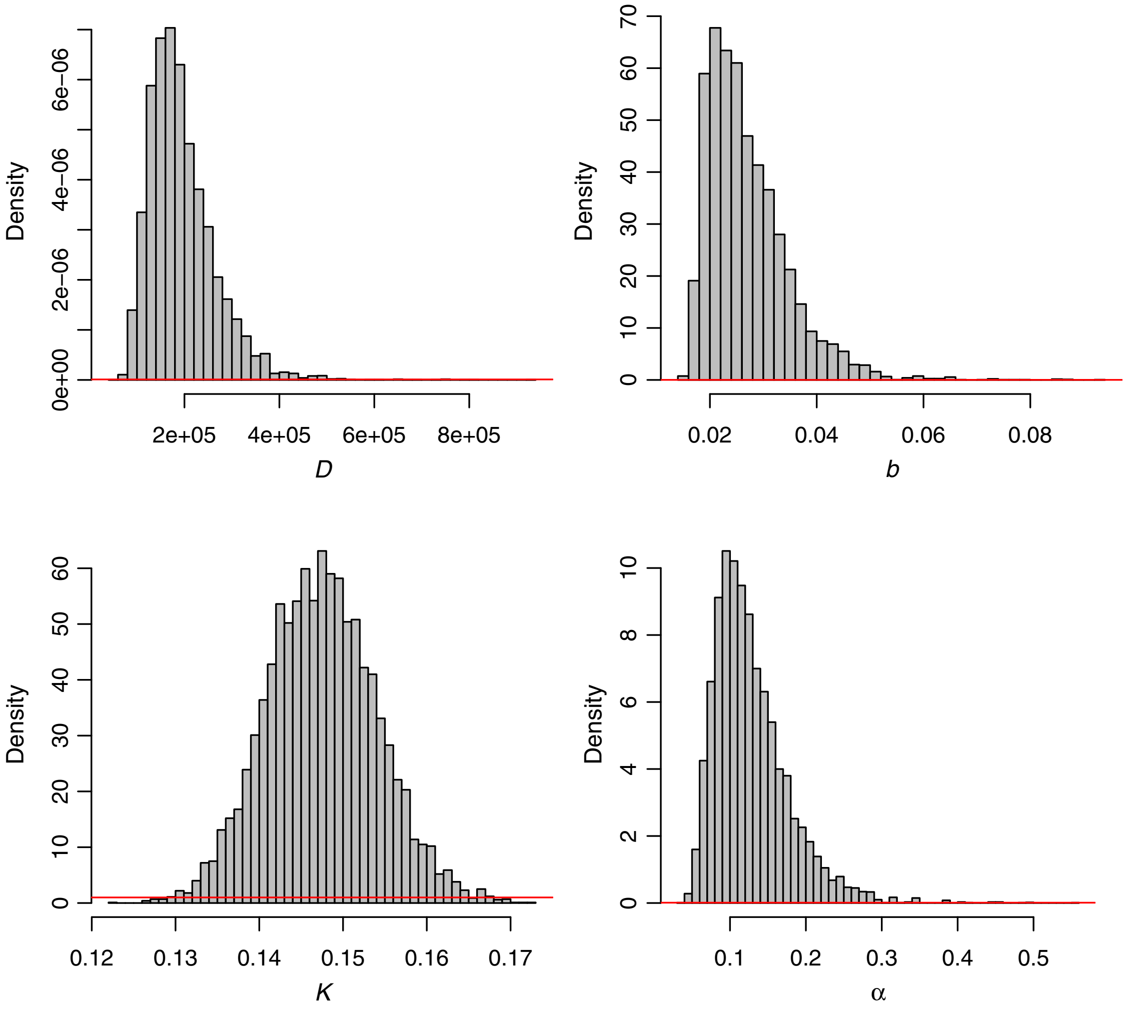


Fig. 7 Marginal posterior distributions of D, b, K and α (histograms) and corresponding prior distributions (red lines) over the supports covered by the posteriors (color figure online)

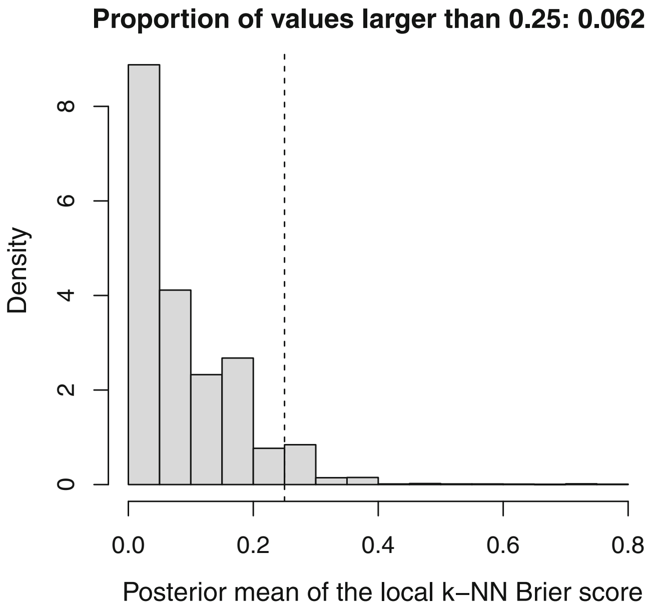


Fig. 8 Distribution of the posterior means of the local Brier scores with $k = 20$. The dashed line gives the 0.25 threshold

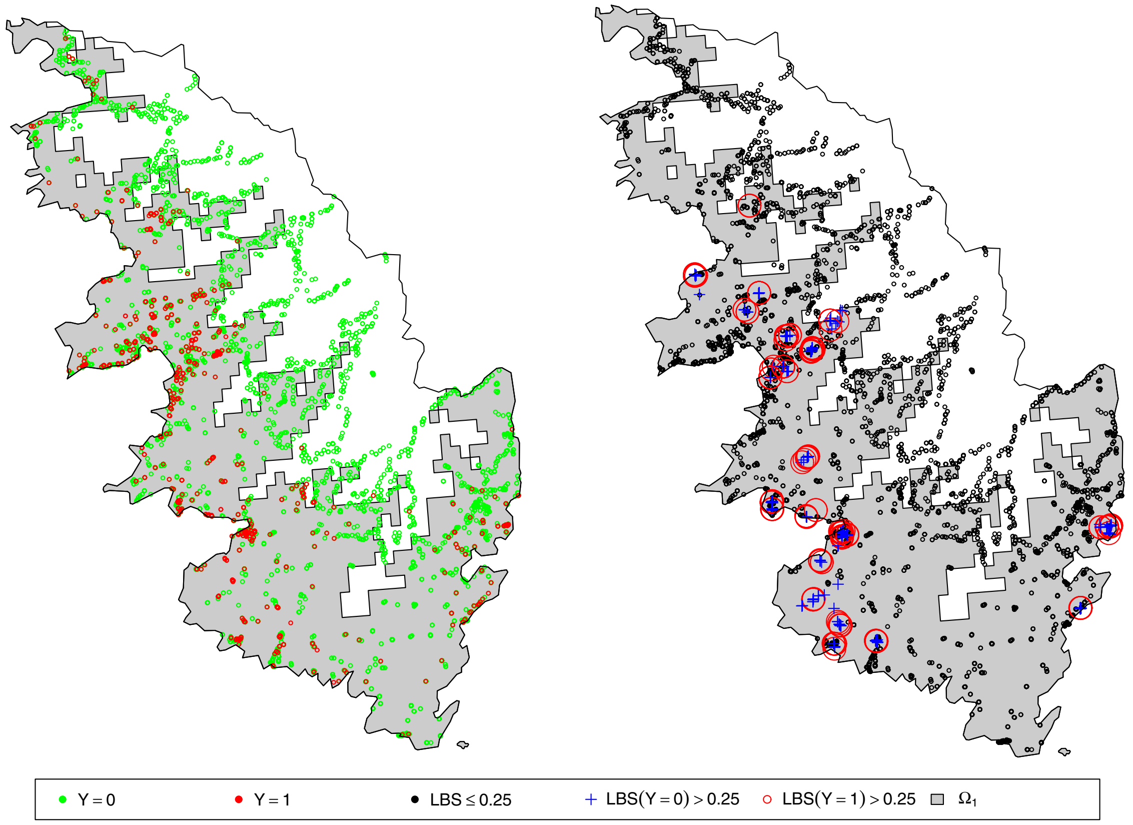


Fig. 9 Locations of samples diagnosed as positive and negative to *Xylella fastidiosa* (left) and samples with different levels of the local Brier score with $k = 20$ (right; black circles: $\mathrm{LBS}_{20}(i) \leq 0.25$; blue crosses: $Y_i^{\mathrm{obs}} = 0$ and $\mathrm{LBS}_{20}(i) > 0.25$; red circles: $Y_i^{\mathrm{obs}} = 1$ and $\mathrm{LBS}_{20}(i) > 0.25$). The gray surface gives the extent of Ω_1 (color figure online)

4 Discussion

Since the detection of Xf in Europe, several modeling approaches have been implemented to provide more insights on the spread of this invasive pathogen in European environments (Strona et al. 2017; White et al. 2017; Bosso et al. 2016; Godefroid et al. 2018; Soubeyrand et al. 2018; Martinetti and Soubeyrand 2018). In this paper, we mainly focus on dating and localizing the introduction of this invasive species. Nevertheless, inferring the parameters of the coupled reaction–diffusion–absorption equation is required since only post-introduction data are available. The conducted analyses using a Bayesian inference approach, tend to show that the introduction of Xf in South Corsica occurred probably near Ajaccio around 1959 (95%-posterior interval: [1933, 1986]), long time before its first detection. Our estimation of the introduction time is relatively consistent with the results obtained by Denancé et al. (2017a) who assessed the introduction of the two main strains found in Corsica around 1965 and 1980, respectively, using a phylogenetic approach. Likewise, our estimation is compatible to the result of Soubeyrand et al. (2018), who dated the introduction around 1985 (95%-posterior interval: [1978, 1993]) with a statistical analysis of temporal data (indeed, the posterior intervals obtained from both analyses overlap). To obtain a more accurate estimation of the introduction date, at least two tracks could be followed: coupling the analysis of spatio-temporal surveillance data and genetic data, as discussed in Soubeyrand et al. (2018), and, as suggested in the result section, gaining knowledge about parameters D, b and α whose estimations are correlated with the estimation of the introduction date (such a knowledge could be incorporated into the prior distribution and could lead to a narrower posterior distribution of τ_0).

To infer the posterior distribution of the parameter vector we proceed in two steps: (i) infer the parameters of the dynamics given the temperature threshold $\tilde{T}$ used for partitioning the study domain, and (ii) choose $\tilde{T}$ using different selection criteria. A possible extension of our work is to refine the definition of the spatial partition by not only using the minimum daily winter temperature but also other relevant environmental variables (Godefroid et al. 2018; Martinetti and Soubeyrand 2018). Thus, a parametric logistic regression function depending on these variables could be built for partitioning the study domain and its parameters should be jointly estimated with the other parameters. However, this perspective requires a faster estimation approach. Indeed, an important milestone towards an accurate inference about the parameter vector, is to accurately solve the partial differential equation, which requires non-negligible computation time. Fortunately, the AMIS algorithm is easily parallelized. However, jointly estimating the partition of the study domain (and not only selecting it as we did), would result on much larger computation times, especially if the partition depends on multiple spatial variables. To reduce the computational cost, approximating the input/output relation in the mechanistic model using meta-models necessitating less computer intensive calculations could be a valuable option, that could be incorporated in AMIS (Osio and Amon 1996; Giunta and Watson 1998). In particular, kriging meta-models show up to be an adequate solution for approximating deterministic models since they interpolate the observed or known data points (Simpson et al. 2001). An additional advantage that derives from the use of AMIS is that its tuning parameters are adapted across the algorithm iterations, contrary to the basic MCMC and the maximum likelihood (ML) approach frequently used in the mechanistic-statistical framework. It has however to be noted that AMIS has to be appropriately initialized, which can be relatively easily done in practice by evaluating the marginal posterior distributions over 1D grids. Still to regard with the computational cost, ML estimation could be an interesting option, even if the control of estimation uncertainty is more convincing in the Bayesian framework for a model such ours. Supplementary Section S3 and Figure S4 precisely investigate ML applied to our case study: using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for the maximization, the computation effort is reduced, but results tend to indicate that the optimization is stuck in local maxima. More complex optimization algorithms, such as the simulated annealing algorithm, could be applied to converge to a global maximum but much more computations would hence be required.

Obviously, the deterministic model [Eqs. (1–2)] that we proposed to describe the dynamics of the pathogen does not take into account all the epidemiological and environmental drivers of the dynamics. These drivers could be implicitly handled by replacing our model by a stochastic version that would result in more flexible realizations. Gonze et al. (2002) compared deterministic and stochastic models for circadian oscillations and showed that, in presence of noise in a small population, stochastic simulations are needed to get more realistic realizations. Although the population size for the case study of Xf is expected to be relatively large, stochastic population-dynamic models, from individual-based models (Renshaw 1993; Kareiva and Shigesada 1983) to aggregated models (Soubeyrand et al. 2009b), could allow to relax hypotheses made on the dynamics. In contrast, our parsimonious model, which only incorporates the main epidemiological and environmental drivers, provides a concise description of the dynamics of the pathogen, and can be fitted to data in a reasonable time span. The advantage of this approach is that it can be rapidly applied for endorsing a fast reaction after the detection of a new invasive pathogen.

Instead of replacing our model by a stochastic version, we could refine it by taking into account relevant supplementary epidemiological and environmental processes. For instance, the diffusion, the growth/decrease of the pathogen infection and the plateau for the infection probability (represented in our model by parameters D, b, α and K) could depend on the spatio-temporal distribution of insects transmitting the pathogen, host density, seasonality and other environmental factors. Incorporating such dependencies into the model and using sufficiently high-resolution maps for spatial factors could allow the modeling of rapid changes in the infection probability that have been observed in Sect. 3.6. This sort of model refinement probably requires, however, more data than we have for Xf. For example, mapping host-density for Xf is not an easy issue because of the large spectrum of host species and the large variability in species susceptibility. Similarly, estimating seasonal effects on the growth/decrease rate of the infection probability certainly requires a larger observation temporal window allowing the detection of seasonal trends (in our case study, observations, which are available during only 2 years long after the introduction, mostly give information on the accumulation of the disease across time, but not on within-year variations of the infection probability). Neglecting all these factors implies that our framework provides estimates of *efficient* parameters (e.g., we estimate an *efficient* diffusion coefficient because diffusion is *averaged* over time in our model, neglecting seasonality in the presence of insect vectors and in the transportation of plants).

An additional perspective for the framework that we proposed is the use of alternative representations of disease propagation. The homogeneous diffusion could be replaced by an heterogeneous diffusion as proposed above, but could also be replaced/augmented by a kernel-based term within an integro-differential equation (Bonnefon et al. 2014), a spatial contact model (Mollison 1977), a mixed dispersal kernel model (Clark et al. 1998), a stratified dispersal model (Shigesada et al. 1995) or a piecewise deterministic Markov process (Abboud et al. 2018). These approaches, allowing a finer quantification of local and long distance dispersal, are generally expected to yield better predictions (Higgins and Richardson 1999; Nathan et al. 2008; Fayard et al. 2009; Gilioli et al. 2013; White et al. 2017). For instance, White et al. (2017) model the spread of Xf in the (supposed) early stages of the invasion in Apulia, Italy, with a stratified dispersal approach. They predict that the long-distance dispersal component is a paramount driver of the rapid spread of the pathogen and has to be taken into account in the design of management strategies. They however advocate that field estimates of key parameters, such as infection growth rate, local and non-local dispersal parameters, should be estimated to decrease prediction uncertainty. The relatively simple framework that we propose precisely provides, using field data, estimates of such parameters and other quantities such as the temperature threshold, the date and the location of the pathogen introduction. Regarding the pathogen introduction, we assumed that there is only one introduction that triggered the invasion and that eventual subsequent introductions had negligible effects on the dynamics. In the aim of relaxing this assumption, stratified dispersal models and piecewise deterministic Markov processes (PDMP) discussed above can be designed to incorporate into the model not only long-distance dispersal but also multiple introductions. Distinguishing these two types of events from surveillance data is not easy in general, except if one has at disposal genetic data or contact tracing data, but can anyway be modeled separately with a mixture of two kernels (identifiability issues of the mixture components may however arise). Abboud et al. (2018) precisely discuss a PDMP embedding multiple introductions without implementing it in practice. This is one of the most attractive perspectives for furthering our work.

References

Abboud C, Senoussi R, Soubeyrand S (2018) Piecewise-deterministic Markov processes for spatio-temporal population dynamics. In: Azaïs R, Bouguet F (eds) Statistical inference for piecewise-deterministic Markov processes, ISTE edn. Wiley, New York

Allaire G (2008) Analyse numérique et optimisation. Les Éditions de l'École Polytechnique, Palaiseau

Anas O, Harrison UJ, Brannen PM, Sutton TB (2008) The effect of warming winter temperature on the severity of pierce's disease in the appalachian mountains and piedmont of the southeastern United States. Plant Health Prog 101094:450–459

Anderson RM, Donnelly CA, Ferguson NM, Woolhouse MEJ, Watt CJ, Udy HJ, Mawhinney S, Dunstan SP, Southwood TRE, Wilesmith JW, Ryan JBM, Hoinville LJ, Hillerton JE, Austin AR, Wells GAH (1996) Transmission dynamics and epidemiology of BSE in British cattle. Nature 382:779–788. https://doi.org/10.1038/382779a0

Ando T (2011) Predictive Bayesian model selection. Am J Math Manag Sci 31:13–38. https://doi.org/10.1080/01966324.2011.10737798

Andow D, Kareiva PM, Levin SA, Okubo A (1990) Spread of invading organisms. Landsc Ecol 4:177–188

Andow DA, Kareiva PM, Levin SA, Okubo A (1993) Spread of invading organisms: patterns of spread. In: Kim KC, McPheron BA (eds) Evolution of insect pests: the pattern of variations. Wiley, New York, pp 219–242

Baker HG (1991) The continuing evolution of weeds. Econ Bot 45:445–449

Berliner LM (2003) Physical-statistical modeling in geophysics. J Geophys Res Atmos 108:8776. https://doi.org/10.1029/2002JD002865

Bonnefon O, Coville J, Garnier J, Roques L (2014) Inside dynamics of solutions of integro-differential equations. Discrete Contin Dyn Syst B 19(10):3057–3085

Bosso L, Russo D, Febbraro MD, Cristinzio G, Zoina A (2016) Potential distribution of *Xylella fastidiosa* in Italy: a maximum entropy model. Phytopathol Mediterr 55:62–72

Boys RJ, Wilkinson DJ, Kirkwood TBL (2008) Bayesian inference for a discretely observed stochastic kinetic model. Stat Comput 18:125–135. https://doi.org/10.1007/s11222-007-9043-x

Brier GW (1950) Verification of forecasts expressed in terms of probability. OPTmonthey Weather Rev 78:1–3

Brooks S (2003) Bayesian computation: a statistical revolution. Trans R Stat Soc Ser A 15:2681–2697. https://doi.org/10.1098/rsta.2003.1263

Bugallo MF, Martino L, Corander J (2015) Adaptive importance sampling in signal processing. Digit Signal Process 47:36–49. https://doi.org/10.1016/j.dsp.2015.05.014

Chapman DS, White SM, Hooftman DA, Bullock JM (2015) Inventory and review of quantitative models for spread of plant pests for use in pest risk assessment for the EU Territory, vol 12. EFSA Supporting Publications, New York. https://doi.org/10.2903/sp.efsa.2015.EN-795

Clark JS, Fastie C, Hurtt G, Jackson ST, Johnson C, King GA, Lewis M, Lynch J, Pacala S, Prentice C, Schupp EW, Webb T III, Wyckoff P (1998) Reid's paradox of rapid plant migration: dispersal theory and interpretation of paleoecological records. BioScience 48:13–24. https://doi.org/10.2307/1313224

Cornuet J, Marin JM, Mira A, Robert CP (2012) Adaptive multiple importance sampling. Scand J Stat 39:798–812. https://doi.org/10.1111/j.1467-9469.2011.00756.x

Costello M, Steinmaus S, Boisseranc C (2017) Environmental variables influencing the incidence of Pierce's disease. Aust J Grape Wine Res 23:287–295. https://doi.org/10.1111/ajgw.12262

Denancé N, Cesbron S, Briand M, Rieux A, Jacques MA (2017a) Is *Xylella fastidiosa* really emerging in France? In: Costa J, Koebnik R (eds) 1st Annual conference of the EuroXanth—COST action integrating science on *Xanthomonadaceae* for integrated plant disease management in Europe, EuroXanth, Coimbra, Portugal, vol 7

Denancé N, Legendre B, Briand M, Olivier V, Boisseson C, Poliakoff F, Jacques MA (2017b) Several subspecies and sequence types are associated with the emergence of *Xylella fastidiosa* in natural settings in France. Plant Pathol 66:1054–1064. https://doi.org/10.1111/ppa.12695

Evans LC (1998) Partial differential equations, graduate studies in mathematics, vol 19. American Mathematical Society, Providence

Fayard J, Klein EK, Lefèvre F (2009) Long distance dispersal and the fate of a gene from the colonization front. J Evol Biol 22(11):2171–2182

Feil H, Purcell AH (2001) Temperature-dependent growth and survival of *Xylella fastidiosa* in vitro and in potted grapevines. Plant Dis 85:1230–1234. https://doi.org/10.1094/PDIS.2001.85.12.1230

Feil H, Feil WS, Purcell AH (2003) Effects of date of inoculation on the within-plant movement of *Xylella fastidiosa* and persistence of Pierce's disease within field grapevines. Phytopathology 93:244–251. https://doi.org/10.1094/PHYTO.2003.93.2.244

Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugen 7:355–369. https://doi.org/10.1111/j.1469-1809.1937.tb02153.x

Gatenby RA, Gawlinski ET (1996) A reaction–diffusion model of cancer invasion. Cancer Res 56:5745–5753

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman A, Roberts GO, Gilks WR et al (1996) Efficient metropolis jumping rules. Bayesian Stat 5:599–608

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Texts in statistical science series. Chapman & Hall/CRC, New York

Gilioli G, Pasquali S, Tramontini S, Riolo F (2013) Modelling local and long-distance dispersal of invasive chestnut gall wasp in europe. Ecol Model 263:281–290

Giunta A, Watson L (1998) A comparison of approximation modeling techniques-polynomial versus interpolating models. In: 7th AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, multidisciplinary analysis optimization conferences, St. Louis, MO, USA, p 4758. https://doi.org/10.2514/MMAO98

Godefroid M, Cruaud A, Streito JC, Rasplus JY, Rossi JP (2018) Climate change and the potential distribution of *Xylella fastidiosa* in Europe. bioRxiv https://doi.org/10.1101/289876

Gonze D, Halloy J, Goldbeter A (2002) Deterministic versus stochastic models for circadian rhythms. J Biol Phys 28:637–653. https://doi.org/10.1023/A:1021286607354

Hecht F (2012) New development in Freefem++. J Numer Math 20:251–266. https://doi.org/10.1515/jnum-2012-0013

Hengeveld R (1989) Dynamics of biological invasions. Springer, New York

Henneberger TS (2003) Effects of low temperature on populations of *Xylella fastidiosa* in sycamore. Ph.D. thesis, University of Georgia

Higgins SI, Richardson DM (1999) Predicting plant migration rates in a changing world: the role of long-distance dispersal. Am Nat 153(5):464–475

Huld TA, Šúri M, Dunlop ED, Micale F (2006) Estimating average daytime and daily temperature profiles within Europe. Environ Model Softw 21:1650–1661

Kareiva P, Shigesada N (1983) Analyzing insect movement as a correlated random walk. Oecologia 56:234–238. https://doi.org/10.1007/BF00379695

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. R Soc 115:700–721. https://doi.org/10.1098/rspa.1927.0118

Lanzarone E, Pasquali S, Gilioli G, Marchesini E (2017) A Bayesian estimation approach for the mortality in a stage-structured demographic model. J Math Biol 75:759–779. https://doi.org/10.1007/s00285-017-1099-4

Lewis MA, Kareiva P (1993) Allee dynamics and the spread of invading organisms. Theor Popul Biol 43:141–158. https://doi.org/10.1006/tpbi.1993.1007

Lindley D (2006) Understanding uncertainty. Wiley, New York. https://doi.org/10.1002/0470055480

Martinetti D, Soubeyrand S (2018) Identifying lookouts for epidemio-surveillance: application to the emergence of *Xylella fastidiosa* in France, submitted

Mason SJ (2004) On using "climatology" as a reference strategy in the brier and ranked probability skill scores. Mon Weather Rev 132:1891–1895. https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2

Mollison D (1977) Spatial contact models for ecological and epidemic spread. J R Stat Soc Ser B (Methodol) 39:283–326

Murray JD (2002) Mathematical biology. In: Interdisciplinary applied mathematics, vol 17, 3rd edn. Springer, New York

Nathan R, Schurr FM, Spiegel O, Steinitz O, Trakhtenbrot A, Tsoar A (2008) Mechanisms of long-distance seed dispersal. Trends Ecol Evol 23(11):638–647

Okubo A (1980) Diffusion and ecological problems: mathematical models, interdisciplinary applied mathematics, vol 10. Springer, New York

Okubo A, Levin S (2002) Diffusion and ecological problems—modern perspectives, 2nd edn. Springer, New York. https://doi.org/10.1007/978-1-4757-4978-6

Osio IG, Amon CH (1996) An engineering design methodology with multistage Bayesian surrogates and optimal sampling. Res Eng Des 8:189–206

Peterson RO, Vucetich JA, Page RE, Chouinard A et al (2003) Temporal and spatial aspects of predator–prey dynamics. Alces 39:215–232. https://doi.org/10.1098/rspb.2015.0973

Protter MH, Weinberger HF (1967) Maximum principles in differential equations. Prentice-Hall, Englewood Cliffs. https://doi.org/10.1007/978-1-4612-5282-5

Purcell A (1977) Cold therapy of pierce's disease of grapevines. Plant Dis Rep 61:514–518

Purcell A et al (1980) Environmental therapy for pierce's disease of grapevines. Plant Dis 64:388–390

Reise K, Olenin S, Thieltges DW (2006) Are aliens threatening european aquatic coastal ecosystems? Helgol Mar Res 60:77. https://doi.org/10.1007/s10152-006-0024-9

Renshaw E (1993) Modelling biological populations in space and time, vol 11. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511624094

Richardson DM, Bond WJ (1991) Determinants of plant distribution: evidence from pine invasions. Am Nat 137:639–668

Roberts GO, Gelman A, Gilks WR (1997) Weak convergence and optimal scaling of random walk metropolis algorithms. Ann Appl Probab 7:110–120

Roques L, Soubeyrand S, Rousselet J (2011) A statistical-reaction–diffusion approach for analyzing expansion processes. J Theor Biol 274:43–51. https://doi.org/10.1016/j.jtbi.2011.01.006

Roques L, Walker E, Franck P, Soubeyrand S, Klein E (2016) Using genetic data to estimate diffusion rates in heterogeneous landscapes. J Math Biol 73:397–422. https://doi.org/10.1007/s00285-015-0954-4

Schwarz G et al (1978) Estimating the dimension of a model. Ann Stat 6:461–464. https://doi.org/10.1214/aos/1176344136

Shigesada N, Kawasaki K (1997) Biological invasions: theory and practice, 1st edn. Oxford series in ecology and evolution. Oxford University Press, Oxford

Shigesada N, Kawasaki K, Takeda Y (1995) Modeling stratified diffusion in biological invasions. Am Nat 146:229–251

Simberloff D (1989) Which insect introductions succeed and which fail?, vol 37. Wiley, Chichester, pp 61–75

Simpson TW, Poplinski J, Koch PN, Allen JK (2001) Metamodels for computer-based engineering design: survey and recommendations. Eng Comput 17:129–150. https://doi.org/10.1007/PL00007198

Skellam JG (1951) Random dispersal in theoretical populations. Biometrika 38:196–218. https://doi.org/10.2307/2332328

Soubeyrand S, Roques L (2014) Parameter estimation for reaction-diffusion models of biological invasions. Popul Ecol 56:427–434. https://doi.org/10.1007/s10144-013-0415-0

Soubeyrand S, Laine AL, Hanski I, Penttinen A (2009a) Spatio-temporal structure of host-pathogen interactions in a metapopulation. Am Nat 174:308–320. https://doi.org/10.1086/603624

Soubeyrand S, Neuvonen S, Penttinen A (2009b) Mechanical-statistical modeling in ecology: from outbreak detections to pest dynamics. Bull Math Biol 71:318–338. https://doi.org/10.1007/s11538-008-9363-9

Soubeyrand S, de Jerphanion P, Martin O, Saussac M, Manceau C, Hendrikx P, Lannou C (2018) What dynamics underly temporal observations? Application to the emergence of *Xylella fastidiosa* in France: probably not a recent story. New Phytol. https://doi.org/10.1111/nph.15177

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Methodol) 64:583–639. https://doi.org/10.1111/1467-9868.00353

Strona G, Carstens CJ, Beck PS (2017) Network analysis reveals why *Xylella fastidiosa* will persist in Europe. Sci Rep 7:71. https://doi.org/10.1038/s41598-017-00077-z

Turchin P (1998) Quantitative analysis of movement: measuring and modeling population redistribution in plants and animals. Sinauer, Sunderland

Verhulst PF (1838) Notice sur la loi que la population suit dans son accroissement. In: Mathématique & sciences humaines, vol 167, Quetelet, pp 51–81

Vermeij GJ (1996) An agenda for invasion biology. Biol Conserv 78:3–9

Weinberger H (1978) Asymptotic behavior of a model in population genetics. In: Chadam JM (ed) Nonlinear partial differential equations and applications. Springer, Berlin, pp 47–96

White SM, Bullock JM, Hooftman DAP, Chapman DS (2017) Modelling the spread and control of *Xylella fastidiosa* in the early stages of invasion in Apulia, Italy. Biol Invasions 19:1825–1837. https://doi.org/10.1007/s10530-017-1393-5

Wikle CK (2003a) Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394

Wikle CK (2003b) Hierarchical models in environmental science. Int Stat Rev 71:181–199

Affiliations

Candy Abboud[1] · **Olivier Bonnefon**[1] · **Eric Parent**[2,3] · **Samuel Soubeyrand**[1]

[1] BioSP, INRA, 84914 Avignon, France

[2] UMR 518 Math. Info. Appli., AgroParisTech, Paris, France

[3] UMR 518 Math. Info. Appli., INRA, Paris, France

3.3. Key Perspective Points of Chapter 4

- ❖ An important milestone towards an accurate inference of the PDE parameters is to accurately solve this PDE, which requires a non negligible computational time. Fortunately AMIS algorithm is easily parallelized. An additional advantage of the AMIS is that its tuning parameters are adapted across the algorithm iterations, contrary to MCMC and the ML frequently used in the framework of the mechanistic-statistical approach.

- ❖ We discussed the pros and cons of replacing our deterministic model by a stochastic version or refining it to account for relevant additional epidemiological and environmental drivers of the dynamics, such as long distance dispersal events, seasonality, supplementary climatic variables, and the spatio-temporal distribution of the vector transmitting the pathogen.

- ❖ Two possible ideas were proposed to obtain a more accurate estimation of the introduction date: coupling the analysis of spatio-temporal surveillance data and genetic data, and gaining knowledge about parameters D, b, and α whose estimation is correlated with the introduction date.

4. Bayesian *model-averaging* for predicting species dynamics

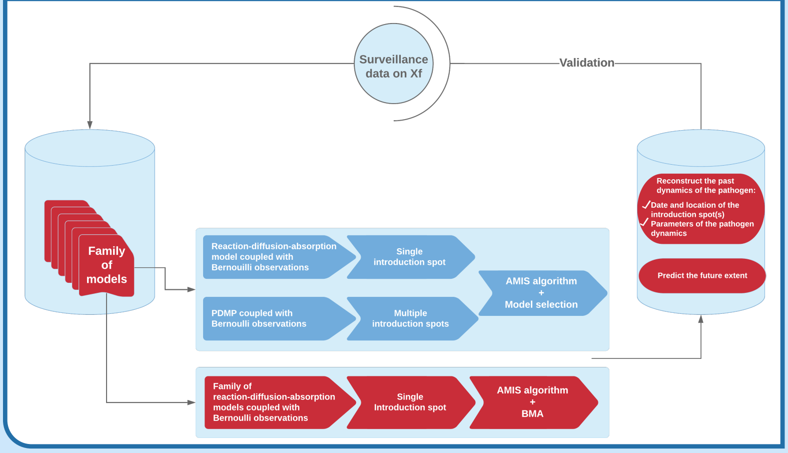
This chapter introduces an article[1] in progress:

Table of contents

[1] C. Abboud, O. Bonnefon, E. Parent, and S. Soubeyrand. Model&data-based prediction of pathogen dynamics. in progress, 2019b

4.1. Graphical Summary

Major Components of The Chapter


Surveillance data on Xf
Validation
Family of models
Reaction-diffusion-absorption model coupled with Bernouilli observations
Single introduction spot
PDMP coupled with Bernoulli observations
Multiple introduction spots
AMIS algorithm + Model selection
Reconstruct the past dynamics of the pathogen:
✓Date and location of the introduction spot(s)
✓Parameters of the pathogen dynamics
Predict the future extent
Family of reaction-diffusion-absorption models coupled with Bernoulli observations
Single Introduction spot
AMIS algorithm + BMA

How to ···

I- Biological Questions

- Get more insights on the pathogen epidemiology.
- Design eradication and containment strategies.
- Assess the potential efficiency of these strategies.

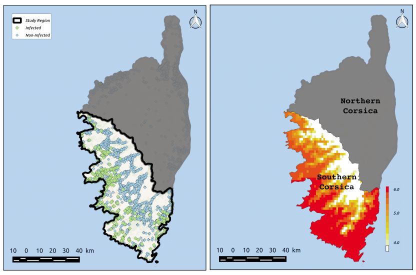
II- Methodological Questions

- Avoid to draw a prediction relying on a single PDE-based model that would be prone to errors.
- Make best use of all the various predictions that can be drawn from a family of models.
- Handle parameter and model uncertainties.
- Investigate the application of the BMA approach in the context of pathogen-dynamics prediction using PDE models and data on Xf.

Methodological Ingredients

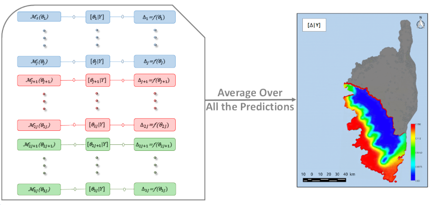
Surveillance Data on Xf with Binary Records

- $\sim$**9000** plants sampled **since 2015** of which **900** have been diagnosed as **infected** (real-time PCR).
- For those $\sim$ 9000 plants, **geographic coordinates** and **sampling dates** are available.
- T : average of the minimum daily temperature over January and February b/n 1995 and 2003 (Map of T with 1 km grid resolution on the right).



BMA Approach

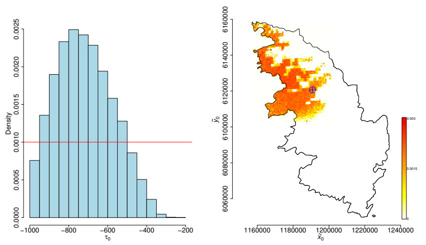
- Aggregate competing models grounded on a family of reaction-diffusion equations with spatially heterogeneous diffusion and reproduction terms.
- Use the AMIS algorithm for parameter estimation of each model.
- Compare different existing approaches based on information criteria and harmonic mean estimators to evaluate the posterior probabilities of models.


$\mathcal{M}_1(\theta_1)$
$[\theta_1|Y]$
$\Delta_1=f(\theta_1)$
$\mathcal{M}_J(\theta_J)$
$[\theta_J|Y]$
$\Delta_J=f(\theta_J)$
$\mathcal{M}_{J+1}(\theta_{J+1})$
$[\theta_{J+1}|Y]$
$\Delta_{J+1}=f(\theta_{J+1})$
$\mathcal{M}_{2J}(\theta_{2J})$
$[\theta_{2J}|Y]$
$\Delta_{2J}=f(\theta_{2J})$
$\mathcal{M}_{2J+1}(\theta_{2J+1})$
$[\theta_{2J+1}|Y]$
$\Delta_{2J+1}=f(\theta_{2J+1})$
$\mathcal{M}_{3J}(\theta_{3J})$
$[\theta_{3J}|Y]$
$\Delta_{3J}=f(\theta_{3J})$
Average Over All the Predictions
$[\Delta|Y]$
10 0 10 20 30 40 km

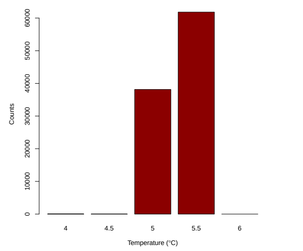
Main Results

Inference of Models Shared Parameters

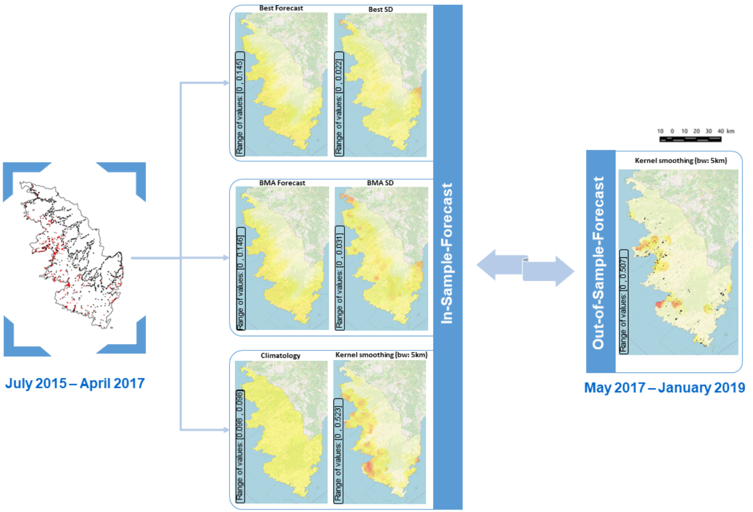
Estimation of the BMA marginal distribution of the date and site of the introduction point.



Estimation of the BMA marginal distribution of the threshold of temperature.



Out-of-sample Predictive Behaviour


July 2015 – April 2017
Best Forecast
Best SD
Range of values: [0 , 0.145]
Range of values: [0 , 0.022]
BMA Forecast
BMA SD
Range of values: [0 , 0.146]
Range of values: [0 , 0.031]
Climatology
Kernel smoothing (bw: 5km)
Range of values: [0.098 , 0.098]
Range of values: [0 , 0.523]
In-Sample-Forecast
Out-of-Sample-Forecast
Kernel smoothing (bw: 5km)
Range of values: [0 , 0.507]
10 0 10 20 30 40 km
May 2017 – January 2019

Conclusion & Perspectives

Conclusion

- The conducted analysis for inferring shared parameters tend to show that the introduction of Xf in South Corsica has probably occurred near Bastellica, to the east of Ajaccio around 1952, long time before its first detection. Results obtained for the introduction date are consistent with the precedent results.
- This approach allows to open up to smoothed threshold of temperature and to account for uncertainties about model forms.
- In our particular case study, the BMA does not seem to outperform the "best" model. However, it succeeded to better reflect the uncertainty about model predictions, avoiding misleading decision making obtained from a single "best" model.
- This approach was tested on simulated data and predictive behaviour was assessed relatively to other reference forecasts.

Perspectives

- Aggregate different structures of the process model by replacing per example the deterministic models by a stochastic version or a version that accounts for additional drivers of the dynamics.
- Refine the definition of the spatial partition by using additional relevant environmental variables, other than the daily winter temperature.
- Incorporate into the models the possibility of multiple introductions of the pathogen.

4.2. Article

ORIGINAL ARTICLE

Model&data-based Prediction of Pathogen Dynamics

Candy Abboud[1*] | Olivier Bonnefon[1] | Eric Parent[2†] | Samuel Soubeyrand[1]

[1]BioSP, INRA, Avignon, 84914, France

[2]AgroParisTech, UMR 518 Math. Info. Appli., Paris, France

Correspondence
Candy Abboud MD, BioSP, INRA, Avignon, 84914, France
Email: candy.abboud@inra.fr

Present address
†INRA, UMR 518 Math. Info. Appli., Paris, France

Funding information
PhD grant INRA-Region PACA (Emplois Jeunes Doctorants 2016-2019); the HORIZON 2020 XF-ACTORS Project SFS-09-201 6; INRA-DGAL project 21000679.

Prediction of invasive-pathogen dynamics is an essential step towards the assessment of eradication and containment strategies. Such predictions are performed using surveillance data and models grounded on partial differential equations (PDE), which form a framework often exploited to design invasion models. The framework allows the construction of phenomenological but concise models relying on mechanistic hypotheses, however, it may lead to models with overly rigid behaviour, in particular for describing phenomena in population biology. Hence, to avoid to draw a prediction relying on a single PDE-based model that would be prone to errors because of potential data-model mismatch, we propose to apply Bayesian model-averaging (BMA) for handling parameter and model uncertainties. In this setting, we use adaptive importance sampling for parameter estimation, and compare different existing approaches based on information criteria and harmonic mean estimators to evaluate the posterior probabilities of models. This approach is applied to predict the future extent of Xf, a phytopathogenic bacterium *in situ* detected in Southern Corsica, France, in 2015.

KEYWORDS
Bayesian model-averaging, Importance sampling, Partial differential equations, Predictive epidemiology, Spatio-temporal model, *Xylella fastidiosa*

1 | INTRODUCTION

The emergence of exogenous pathogens in new territories may induce severe sanitary and socio-economical crises. Such crises are reinforced by the eventually long delay between the establishment of the pathogen in a new territory and its detection (Jones and Baker, 2004; Faria et al., 2014; Soubeyrand et al., 2018), because the cost for pathogen eradication or containment generally increases with this delay, and by the potential for expansion of the pathogen. Hence, reconstructing the past dynamics of the pathogen (Boys et al., 2008; Roques et al., 2016; Soubeyrand and Roques, 2014) and predicting its future extent (Chapman et al., 2015; Peterson et al., 2003) are key steps for understanding the pathogen epidemiology, designing eradication or containment strategies and assessing their potential efficiency.

Partial differential equations (PDE) have been extensively used for modelling spatio-temporal population dynamics (Skellam, 1951; Okubo, 1980; Shigesada et al., 1995; Gatenby and Gawlinski, 1996; Shigesada and Kawasaki, 1997; Turchin, 1998; Okubo and Levin, 2002). PDE can precisely be used for past dynamics reconstruction and future extent prediction, by exploiting their ability (i) to represent dynamics in a phenomenological and concise way, and (ii) to be fitted to data by attaching a probabilistic model of observations within a state-space modelling framework (Berliner, 2003; Roques et al., 2011; Soubeyrand and Roques, 2014; Abboud et al., 2018). However, these equations, mostly relatively simple, are not proficient in describing all the processes and sources of variability involved in an epidemiological dynamics. In addition, various structures of PDE are likely to be considered as candidate models for a given epidemics. When the goal of the study is to draw predictions, the use of one single model is prone to prediction error because this model may not have taken into account crucial drivers of the dynamics. This limitation can be circumvented by considering a set of candidate models and combine them, either by applying a model selection strategy (Burnham et al., 1995) or a model aggregation strategy (Hoeting et al., 1999).

As part of the aggregation strategy, the Bayesian model-averaging (BMA) approach has been proposed by Leamer (1978) to reduce and account for parameter and model uncertainties. This approach consists in averaging over all candidate models in a Bayesian way for weighting models (Raftery, 1996; Volinsky et al., 1997), combining multiple predictions and combining estimations to infer shared parameters (Roberts, 1965; Madigan and Raftery, 1994; Wintle et al., 2003). Theoretically, BMA provides better average predictive ability, as measured by a logarithmic scoring rule, than using any single model (Madigan and Raftery, 1994). The BMA efficiency has been largely explored, in particular with respect to its theoretical properties (Rubin and Schenker, 1986; Madigan and Raftery, 1994; Raftery and Zheng, 2003), leave-one-out predictive performance (Madigan et al., 1995; Lamon and Clyde, 2000; Fernández et al., 2002) and numerical performance (George and McCulloch, 1993; Clyde and George; Viallefont et al., 2001). While BMA is an intuitively attractive solution to the problem of accounting for model uncertainty, it presents several difficulties related to its numerical implementation (Hoeting et al., 1999). By dint of some pioneering work implementing BMA (Madigan and Raftery, 1994; Raftery, 1996; Volinsky et al., 1997), BMA has been applied in numerous study domains such as medicine (Oehler et al., 2009; Yin and Yuan, 2009), ecology (Boone et al., 2005, 2008; Wintle et al., 2003), meteorology (Raftery et al., 2005), genetics (Yeung et al., 2005), economical and political sciences (Eicher et al., 2011; Sidman et al., 2008), engineering and physical sciences (Raftery et al., 2010; Parkinson and Liddle, 2013) and epidemiology (Viallefont et al., 2001). Despite ample literature on BMA and its usefulness, it has been marginally applied in the context of predictive epidemiology.

In this article, we investigate the application of BMA in the context of pathogen-dynamics prediction using PDE-based models and we want to test its efficiency on a real case study. The models are grounded on a family of reaction-diffusion equations with eventual spatially heterogeneous diffusion and reproduction terms. Our aim is to compute, from post-introduction data, the BMA posterior distribution of a certain quantity of interest Δ, which is typically the introduction time or location of the pathogen or its future spatial extent. Following Abboud et al. (2019), we apply to each model the Adaptive Multiple Importance Sampling algorithm (AMIS; Cornuet et al., 2012) for providing an empirical approximation, obtained via a weighted sample $\{\Delta_n, w_n\}_{n=1}^{N}$ of size N, of the posterior distribution of Δ given the specified model. Then, for drawing BMA posterior samples of Δ, we compute posterior probabilities of models using different approximations of the integrated likelihood that have been proposed in the literature. Namely, we compare estimators of the integrated likelihood that use information criteria: WAIC (Watanabe, 2013), BIC (Schwarz et al., 1978), DIC in its two versions (Spiegelhalter et al., 2002; Gelman et al., 2003) and IC (Ando, 2011), as well as harmonic mean estimators (HME; Raftery, 1996; Gelfand and Dey, 1994).

This approach is first tested on simulated data and then applied to make predictions concerning the dynamics of the phytopathogenic bacterium *Xylella fastidiosa* (Xf) in Southern Corsica, France. For this real case study, abundant spatio-temporal and binary post-introduction surveillance data have been collected from an intensive surveillance plan implemented by governmental agencies after the first *in situ* detection of Xf in 2015 in Corsica. This quarantine pathogen in Europe has significantly impacted olive production in Puglia, Italy, and presents a drastic risk of environmental degradation due to its ability to reach a large variety of plant species. It is currently present in a large part of Corsica island and more marginally in Southeastern mainland France (Denancé et al., 2017a; Soubeyrand et al., 2018; Martinetti and Soubeyrand, 2018). Xf is susceptible to cause a major sanitary crisis in France, as the one caused in Italy since 2013 where the socio-economical impacts are considerable due to the grabbing-up and death of a significant proportion of olive trees.

The paper is organised as follows: Data are briefly described in Section 2. The competing models coupling a partial differential equation and a Bernoulli observation process are presented in Section 3. The Bayesian model-averaging technique is described in Section 4. The simulation study is presented in Section 5. Results obtained from surveillance data for Xf in Southern Corsica are detailed in Section 6; we specifically focus on model comparison, parameter inference and out-of sample predictive performance. Finally, Section 7 provides a conclusion and a discussion of perspectives.

2 | SURVEILLANCE DATA WITH BINARY RECORDS

In this article, we analyse spatio-temporal binary post-introduction records informing about the presence/absence of Xf in Southern Corsica, France. Data have been collected since the first detection of the bacterium in the study region in July 2015. Between July 2015 and August 2018, approximately 9000 plants were sampled, among which 900 have been diagnosed as infected with a real-time polymerase chain reaction (real-time PCR) technique (Denancé et al., 2017b). GPS coordinates, sampling dates and sanitary statuses (healthy/infected) are available for all the sampled plants. Spatial locations and sanitary statuses at the sampling times are shown in Figure 1.

As for other bacteria, the growth and mortality of Xf are affected by various environmental variables such as habitability of the environment, nutrients, climatic conditions and availability of dissemination means (e.g. insect vectors). In this study, to account for spatial heterogeneity in the diffusion and the reproduction regimes of the epidemics, we use temperature data to divide the spatial domain denoted by Ω into two sub-domains, say Ω_1 and Ω_2, such that $\Omega = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, and different diffusion and growth terms are applied to Ω_1 and Ω_2. We exploit a freely available database (PVGIS © European Communities, 2001–2008) providing, in particular, monthly averages of the daily minimum temperature reconstructed over a grid with 1×1km spatial resolution (Huld et al., 2006); these monthly averages correspond to the period 1995-2003, but are used as references over the period covered by our study. We use these data to build the average of the daily minimum temperature over January and February, say $T(\mathbf{x})$ for any location $\mathbf{x}$; see Figure 1. Average daily minimum temperature in Winter is known to be a crucial factor for the presence or abundance of Xf (Anas et al., 2008; Martinetti and Soubeyrand, 2018).

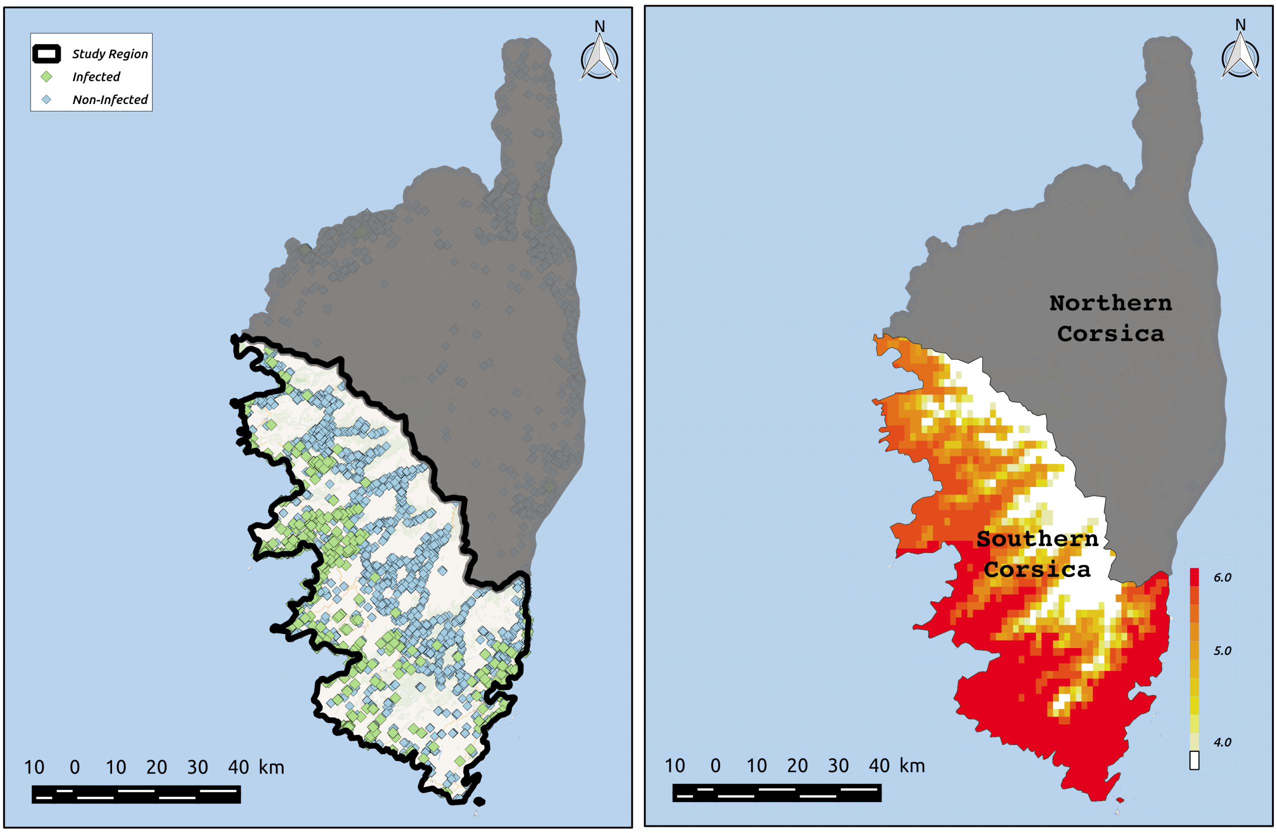


FIGURE 1 Locations of plants (left panel), sampled from July 2015 to August 2018, that have been detected as positive (green dots) or negative (blue dots) to Xf in Southern Corsica, France, and map of the average of the daily minimum temperature (right panel) in Celsius degrees) over January and February reconstructed over a grid with 1×1km spatial resolution.

3 | COMPETING MODELS

Here, a set of models based on parabolic partial differential equations is used to describe pathogen dynamics at large spatial scales. As explained in the introduction, those models have been extensively used to represent population dynamics in a phenomenological and concise way, and can be fitted to data in a hierarchical modelling setting incorporating a probabilistic model of observations (Berliner, 2003; Roques et al., 2011; Soubeyrand and Roques, 2014). In the current section, we propose a family of mechanistic models and we present the model for the observation process.

3.1 | Family of Mechanistic Models

We introduce a discrete family $\mathcal{M} = \{\mathcal{M}_i(\tilde{T}) : 0 \leq i \leq I; \tilde{T} \in \mathcal{T}\}$ of models governing the probability $u(t, \mathbf{x})$ of a host located at site $\mathbf{x} = (x_1, y_1) \in \Omega$ to be infected at time t, where $I \in \mathbb{N}$ and $\mathcal{T}$ is a finite collection of real values. The label i refers to a model structure, i.e. a specific form for the parabolic PDE. The label $\tilde{T}$ refers in our application to a threshold temperature, which splits the spatial domain into two sub-domains where diffusion and growth terms may be different.

The generic form of models in family $\mathcal{M}$ satisfies:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta(D_{i\tilde{T}}(\mathbf{x})u) + f_{i\tilde{T}}(u), & t \geq \tau_0,\ \mathbf{x} \in \Omega, \\ \nabla(D_{i\tilde{T}}(\mathbf{x})u(t,\mathbf{x})).n(\mathbf{x}) = 0, & t \geq \tau_0,\ \mathbf{x} \in \partial\Omega, \\ u(\tau_0,\mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \end{cases} \tag{1}$$

where the first line is the reaction-diffusion equation, the second line gives boundary conditions, the third line gives initial conditions, $\Delta = \dfrac{\partial^2}{\partial x_1^2} + \dfrac{\partial^2}{\partial x_2^2}$ is the 2-dimensional diffusion operator of Laplace, and $\nabla = \dfrac{\partial}{\partial x_1} + \dfrac{\partial}{\partial x_2}$ is the 2-dimensional gradient operator.

The diffusion coefficient $D_{i\tilde{T}}(\mathbf{x})$ may be spatially heterogeneous and is defined as a spatial regularization of:

$$d_{i\tilde{T}}(\mathbf{x}) = \sum_{k=1}^{2} D_{i\tilde{T}k} 1(\mathbf{x} \in \Omega_{\tilde{T}k}), \ \ \forall i \leq I, \forall \tilde{T} \in \mathcal{T}, \tag{2}$$

where $\mathbf{x} \mapsto 1(\mathbf{x} \in \Omega_{\tilde{T}k})$ is the indicator function taking the value 1 if $\mathbf{x} \in \Omega_{\tilde{T}k}$ and 0 otherwise, and the sub-domains $\Omega_{\tilde{T}1}$ and $\Omega_{\tilde{T}2}$ are defined by thresholding the spatial function T, with the threshold value $\tilde{T}$ such that: $\Omega_{\tilde{T}1} = \{\mathbf{x} \in \Omega : T(\mathbf{x}) > \tilde{T}\}$ and $\Omega_{\tilde{T}2} = \{\mathbf{x} \in \Omega : T(\mathbf{x}) \leq \tilde{T}\}$. In our application, T is a measure of temperature in winter, $\Omega_{\tilde{T}1}$ is the warm region of Ω, and $\Omega_{\tilde{T}2}$ the cold one. If $D_{i\tilde{T}1} = D_{i\tilde{T}2}$, then the diffusion coefficient is spatially homogeneous. The spatial regularization is required for the existence and the uniqueness of a classic solution $u(t,\mathbf{x})$ of Equation (1); see (Roques, 2013). Thus $D_{i\tilde{T}}$ is defined as:

$$D_{i\tilde{T}}(\mathbf{x}) = 1(x \in \Omega) \int_{\mathbb{R}^2} \phi(\mathbf{x} - \mathbf{y}) d_{i\tilde{T}}(\mathbf{y}) d\mathbf{y}, \ \ \forall \mathbf{x} \in \Omega, \tag{3}$$

where ϕ is the normal regularization kernel

$$\phi(\mathbf{x}) = \frac{1}{2\pi\mathcal{V}} e^{-\dfrac{||\mathbf{x}||^2}{2\mathcal{V}}}, \tag{4}$$

and the transition speed $\mathcal{V}$ has to be tuned to approach more or less the piecewise constant function $d_{i\tilde{T}}$.

The reproduction term may also be either spatially heterogeneous or not. In the homogeneous case,

$$f_{i\tilde{T}}(u) = bu\left(1 - \frac{u}{K}\right), \tag{5}$$

and in the heterogeneous case,

$$f_{i\tilde{T}}(u) = bu\left(1 - \frac{u}{K}\right) 1(\mathbf{x} \in \Omega_{\tilde{T}1}) - \alpha u 1(\mathbf{x} \in \Omega_{\tilde{T}2}), \ \forall \tilde{T} \in \mathcal{T}, \tag{6}$$

where b is the intrinsic growth rate of the epidemics; $K \in (0,1]$ is a plateau for the probability of infection (i.e. an analog to the carrying capacity of the environment); α is the decrease rate of the infection in $\Omega_{\tilde{T}2}$ in the heterogeneous case.

In the application, we will consider three model structures:

- $\mathcal{M}_0(\tilde{T})$, under which $D_{i\tilde{T}1} = D_{i\tilde{T}2}$ and $f_{i\tilde{T}}$ satisfies Equation (6), i.e. homogeneous diffusion but heterogeneous growth, like in Abboud et al. (2019);

- $\mathcal{M}_1(\tilde{T})$, under which $D_{i\tilde{T}1} \neq D_{i\tilde{T}2}$ and $f_{i\tilde{T}}$ satisfies Equation (6), i.e. heterogeneous diffusion and growth;
- $\mathcal{M}_2(\tilde{T})$, under which $D_{i\tilde{T}1} \neq D_{i\tilde{T}2}$ and $f_{i\tilde{T}}$ satisfies Equation (5), i.e. heterogeneous diffusion, but homogeneous growth.

The second equation in system (1) corresponds to the homogeneous Neumann condition on the boundary $\partial\Omega$ of Ω (i.e. with reflection on the boundary). This condition is formalized by setting that the gradient of the spatial function $x \mapsto D_{i\tilde{T}}(\mathbf{x})u(t,\mathbf{x})$ is orthogonal to the outward normal vector $n(x)$ at point x on the boundary $\partial\Omega$, for all $t \geq \tau_0$. Thus, physically, there is neither outward nor inward flux from and to Ω.

The spatial function u_0 models the introduction of the pathogen in the study domain at time $\tau_0 \in \mathbb{R}$. Following Abboud et al. (2019), the introduction represents the initial phase of the outbreak corresponding to the arrival of the pathogen and its local establishment. Thus, u_0 is not expressed as a Dirac function but as a kernel function centered around the central point of the introduction $\tilde{\mathbf{x}}_0 = (\tilde{x}_0, \tilde{y}_0) \in \Omega$. More precisely, the probability of a host at $\mathbf{x}$ to be infected at τ_0 satisfies:

$$u_0(\mathbf{x}) = P_0 \exp\left(-\frac{\|\mathbf{x} - \tilde{\mathbf{x}}_0\|^2}{2\sigma^2}\right), \tag{7}$$

where P_0 is the infection probability at $(\tau_0, \tilde{\mathbf{x}}_0)$, $\sigma^2 = \frac{r_0^2}{q}$, q is the 0.95-quantile of the χ^2 distribution with two degrees of freedom, and r_0 is the *radius* of the kernel. Thus, at τ_0, if we neglect border effects, 95% of the infected plants are located within the ball with center $\tilde{\mathbf{x}}_0$ and radius r_0.

With such initial and boundary conditions, the system of equations (1) is well-posed (Evans, 1998). In addition, by constraining P_0 in $[0, K]$, the principle of parabolic comparison (Protter, MH and Weinberger, HF, 1967) implies that the solution of (1) remains in the interval $[0, K]$.

3.2 | Probabilistic Models of Observations

Let $t_j \in \mathbb{R}$ denote the sampling time of host $j \in \{1, \ldots, J\}$, $J \in \mathbb{N}^*$, $\mathbf{x}_j \in \Omega$ its location and $Y_j \in \{0, 1\}$ its sanitary status observed at time t_j (1 for infected, 0 for healthy). Conditionally on u, $\mathcal{M}_i(\tilde{T})$ and $\{(t_j, \mathbf{x}_j) : 1 \leq j \leq J\}$, the sanitary statuses $Y_j, j \in \{1, \ldots, J\}$, are assumed to be independent random variables following Bernoulli distributions with success probability $u(t_j, \mathbf{x}_j)$:

$$Y_j \mid u, \mathcal{M}_i(\tilde{T}), \{(t_j, \mathbf{x}_j) : 1 \leq j \leq J\} \underset{\text{indep.}}{\sim} \text{Bernoulli}(u(t_j, \mathbf{x}_j)), \tag{8}$$

where u depends on the model $\mathcal{M}_i(\tilde{T})$ and its vector of parameters $\Theta_{i\tilde{T}}$.

Remark. This data model was proposed in Abboud et al. (2019) for its simplicity. It could be refined to account for sampling errors classically encountered in epidemiology, e.g. false-positive and false-negative observations, as well as spatial and temporal dependencies not accounted for in the process model.

4 | BAYESIAN MODEL-AVERAGING

4.1 | Principle

Briefly, the BMA consists in estimating the expectation of the posterior distributions of the variable of interest Δ provided under all the competing models and weighted by the posterior model probabilities (Raftery, 1996; Hoeting et al., 1999). In the modeling setting introduced above, Δ is typically a vector of shared parameters such as the introduction point ($\tilde{\mathbf{x}}_0$,τ_0), the temperature threshold $\tilde{T}$ or the spatial probability of infection u over a future period. Using Gelfand's bracket notation for probability distributions (Gelfand and Smith, 1990), the BMA posterior distribution of Δ given training data Y satisfies:

$$[\Delta|Y] = \sum_{0\le i\le \mathcal{I}, \tilde{T}\in\mathcal{T}} [\Delta|Y, \mathcal{M}_i(\tilde{T})] \times [\mathcal{M}_i(\tilde{T})|Y]. \tag{9}$$

The posterior model probability of $\mathcal{M}_i(\tilde{T})$ is:

$$[\mathcal{M}_i(\tilde{T})|Y] = \frac{[Y|\mathcal{M}_i(\tilde{T})] \times [\mathcal{M}_i(\tilde{T})]}{\sum_{0\le i'\le \mathcal{I}, \tilde{T}'\in\mathcal{T}} [Y|\mathcal{M}_{i'}(\tilde{T}')] \times [\mathcal{M}_{i'}(\tilde{T}')]}. \tag{10}$$

The integrated likelihood $[Y|\mathcal{M}_i(\tilde{T})]$ of $\mathcal{M}_i(\tilde{T})$, which may be a complex integral depending on the dimension of the unknowns and eventual dependencies, satisfies:

$$[Y|\mathcal{M}_i(\tilde{T})] = \int [Y|\Theta, \mathcal{M}_i(\tilde{T})] \times [\Theta|\mathcal{M}_i(\tilde{T})] d\Theta. \tag{11}$$

where Θ is the vector of parameters of $\mathcal{M}_i(\tilde{T})$; $[Y|\Theta, \mathcal{M}_i(\tilde{T})]$ is the likelihood under $\mathcal{M}_i(\tilde{T})$; $[\Theta|\mathcal{M}_i(\tilde{T})]$ is the prior distribution of Θ under $\mathcal{M}_i(\tilde{T})$; and $[\mathcal{M}_i(\tilde{T})]$ is the prior probability of $\mathcal{M}_i(\tilde{T})$. The posterior mean of Δ is a weighted average of the posterior means under the competing models:

$$\mathbb{E}[\Delta|Y] = \sum_{0\le i\le \mathcal{I}, \tilde{T}\in\mathcal{T}} [\mathcal{M}_i(\tilde{T})|Y] \times \mathbb{E}[\Delta|\mathcal{M}_i(\tilde{T}), Y]. \tag{12}$$

The posterior variance is expressed as follows:

$$\mathbb{V}[\Delta|Y] = \sum_{0\le i\le \mathcal{I}, \tilde{T}\in\mathcal{T}} [\mathcal{M}_i(\tilde{T})|Y] \times \left(\mathbb{V}[\Delta|\mathcal{M}_i(\tilde{T}), Y] + \left(\mathbb{E}[\Delta|\mathcal{M}_i(\tilde{T}), Y] - \mathbb{E}[\Delta|Y]\right)^2\right). \tag{13}$$

4.2 | Implementation

To compute the BMA-posterior distribution of Δ, we undertake a two-step process: In the first step, we compute the posterior distribution of Δ given model $\mathcal{M}_i(\tilde{T})$ and training data Y; The second step consists in the computation of the posterior model probability.

4.2.1 | Empirical Approximation of the Posterior Distribution of Δ Given a Model $\mathcal{M}_i(\tilde{T})$

For the mechanistic-statistical models defined in section 3, the likelihood $[Y|\Delta, \mathcal{M}_{i\tilde{T}}]$ cannot be expressed analytically because it is a function of u that cannot be written in a closed form, thus the need to approximate it using an adequate algorithm for Bayesian inference. Following Abboud et al. (2019), we use the adaptive multiple importance sampling (AMIS; Cornuet et al., 2012) algorithm, that consists of iteratively generating parameter vectors under an adaptive proposal distribution and assigning / updating weights to the parameter vectors. To design efficient importance sampling algorithms, the auxiliary proposal distribution should be chosen as close as possible to the posterior distribution. However, the posterior distribution being unknown, the crucial choice of the proposal is a difficult task (Gelman et al., 1996; Roberts et al., 1997). The main aim of the AMIS algorithm is to overcome this difficulty by tuning the coefficients of the proposal distribution picked from a parametric family of distributions, generally the Gaussian one, at the end of each iteration. In this framework, at each iteration, new coefficient values for the proposal distribution are determined using the current weighted posterior sample (Bugallo et al., 2015), then the posterior sample is augmented by generating new replicates from the newly tuned proposal distribution and the weights of the cumulated posterior sample are recomputed. The AMIS algorithm provides a weighted posterior sample $\{(\Delta_m^l, w_m^l) : 1 \le m \le M, 1 \le l \le L\}$ of size ML, which forms an empirical approximation of the posterior distribution $[\Delta | Y, \mathcal{M}_i(\tilde{T})]$ (m stands for the iteration; l stands for the replicate parallelly generated at iteration m). Conditions leading to the convergence in probability of the posterior mean of any function (integrable with respect to the posterior distribution) of the parameters are described in Cornuet et al. (2012) and are satisfied in our case (Abboud et al., 2019).

We implemented the AMIS algorithm in the R statistical software, with calls to the software `Freefem++` for solving the PDE and, then, computing the likelihood. Parallel computation was performed in order to reduce the non-negligible time of PDE numerical resolution. With $(M, L) = (50, 10^4)$ and the use of 100 cluster cores (the cluster being composed of 40-cores nodes Xeon(R) 2.2 Ghz, 228 Go RAM), the estimation procedure for a fixed value minimum winter temperature $\tilde{T} \in \mathcal{T}$ takes approximately 1.75 days. Conversely to the MCMC algorithm that is often used in the mechanistic-statistical framework (Soubeyrand and Roques, 2014; Lanzarone et al., 2017), AMIS, as a purely Monte Carlo algorithm, can be easily parallelized, and its tuning parameters are, in addition, automatically adapted at each iteration. The AMIS algorithm provides at each iteration an assessment of the posterior distribution of parameters, which is expected to converge to the true posterior distribution and to be stable after a so-called period of burn-in.

4.2.2 | Computation of the integrated likelihood

The evaluation of the integrated likelihood is required to compute the posterior model probability as shown in equation (10). Hereafter, we present various methods to compute the integrated likelihood (that is not analytically tractable in our case) in order to assess their impact on BMA predictions. These methods are either based on information criteria or harmonic mean estimators.

Estimators grounded on information criteria

The following information criteria are generally used for model comparison but can also be exploited to assess the integrated likelihood in BMA as presented in McElreath (2018), and in equation (24) below.

The BIC (Bayesian information criterion; Schwarz et al., 1978) satisfies for model $\mathcal{M}_i(\tilde{T})$:

$$\mathrm{BIC}_{i\tilde{T}} = -2\log[Y | \hat{\Theta}, \mathcal{M}_i(\tilde{T})] + \xi_{i\tilde{T}} \log J, \tag{14}$$

where J is the sample size, $\xi_{i\tilde{T}}$ is the number of model parameters, and $\hat{\Theta}$ is the maximum likelihood estimate of the parameter vector Θ of model $\mathcal{M}_i(\tilde{T})$ in the support $\mathcal{S}_{i\tilde{T}}$ of Θ:

$$\hat{\Theta} = \underset{\Theta \in \mathcal{S}_{i\tilde{T}}}{\mathrm{argmax}} [Y | \Theta, \mathcal{M}_i(\tilde{T})].$$

The DIC (deviance information criterion; Gelman et al., 2014) satisfies for model $\mathcal{M}_i(\tilde{T})$:

$$\mathrm{DIC}_{i\tilde{T}} = \bar{\mathcal{D}} + p_{\mathrm{eff}}, \tag{15}$$

where $\bar{\mathcal{D}}$ is the posterior mean of the deviance $\mathcal{D}(\Theta) = -2\log[Y | \Theta, \mathcal{M}_i(\tilde{T})] + C$ (C is a constant that cancels out when one compares different models) and p_{eff} is the effective number of parameters of the model. The difference in the two versions of the DIC considered here lies in the calculation of p_{eff}. In the first version proposed by Spiegelhalter et al. (2002),

$$p_{\text{eff}} = p_{\mathcal{D}} = \bar{\mathcal{D}} - \mathcal{D}(\bar{\Theta}), \tag{16}$$

where $\bar{\Theta}$ is the posterior mean of Θ under the model $\mathcal{M}_i(\tilde{T})$: $\bar{\Theta} = \mathbb{E}[\Theta|Y, \mathcal{M}_i(\tilde{T})]$. In the second version proposed by Gelman et al. (2003),

$$p_{\text{eff}} = \frac{1}{2}\mathbb{V}(\mathcal{D}(\Theta)|Y, \mathcal{M}_i(\tilde{T})), \tag{17}$$

where $\mathbb{V}(\mathcal{D}(\Theta)|Y, \mathcal{M}_i(\tilde{T}))$ is the posterior variance of $\mathcal{D}(\Theta)$.

The IC (information criterion) of Ando (2011) is supposed to solve over-fitting issues in comparison with DIC:

$$\text{IC}_{i\tilde{T}} = \bar{\mathcal{D}} + 2p_{\mathcal{D}} := 3\bar{\mathcal{D}} - 2\mathcal{D}(\bar{\Theta}). \tag{18}$$

In practice, the different terms appearing in the four criteria, namely $\hat{\Theta}$, $\bar{\Theta}$, $\bar{\mathcal{D}}$ and $\mathbb{V}(\mathcal{D}(\Theta)|Y, \mathcal{M}_i(\tilde{T}))$, are replaced by their empirical values using the weighted posterior sample $\{(\Delta_m^l, w_m^l) : 1 \leq m \leq M, 1 \leq l \leq L\}$ provided by the AMIS algorithm applied to the model $\mathcal{M}_i(\tilde{T})$.

The WAIC (Watanabe–Akaike information criterion; Watanabe, 2010), defined in the deviance scale for allowing comparison to DIC, satisfies for model $\mathcal{M}_i(\tilde{T})$:

$$\text{WAIC}_{i\tilde{T}} = -2(\text{lppd}_{i\tilde{T}} - \text{p}_{\text{WAIC}_{i\tilde{T}}}), \tag{19}$$

where $\text{lppd}_{i\tilde{T}}$ is the log pointwise predictive density and $p_{\text{WAIC}_{i\tilde{T}}}$ is an estimate of the effective number of free parameters in the model. The term $\text{lppd}_{i\tilde{T}}$ is the logarithm of the predictive density integrated over the posterior distribution of the model parameters summed over all observations:

$$\text{lppd}_{i\tilde{T}} = \sum_{j=1}^{J} \log \int [Y_j|\Theta, \mathcal{M}_i(\tilde{T})] \times [\Theta|Y, \mathcal{M}_i(\tilde{T})] d\Theta. \tag{20}$$

In practice, $\text{lppd}_{i\tilde{T}}$ is replaced in the WAIC formula by its empirical analogue (see Gelman et al., 2014) using the weighted posterior sample:

$$\sum_{j=1}^{J} \log \left(\sum_{m=1}^{M} \sum_{l=1}^{L} [Y_j|\Theta_m^l, \mathcal{M}_i(\tilde{T})] w_m^l \right). \tag{21}$$

Following the initial proposal by Watanabe (2010), the term $p_{\text{WAIC}_{i\tilde{T}}}$ can be expressed as:

$$\text{p}_{\text{WAIC}_{i\tilde{T}}} = 2\sum_{j=1}^{J} \left(\log \left(\mathbb{E}\left[Y_j|\Theta, \mathcal{M}_i(\tilde{T})\right] \right) - \mathbb{E}\Big(\log \left[Y_j|\Theta, \mathcal{M}_i(\tilde{T})\right] \Big) \right). \tag{22}$$

We will also consider the following expression for $p_{\text{WAIC}_{i\tilde{T}}}$ advocated by Gelman et al. (2014):

$$\text{p}_{\text{WAIC}_{i\tilde{T}}} = \sum_{j=1}^{J} \mathbb{V}\Big(\log[Y_j|\Theta, \mathcal{M}_i(\tilde{T})]\Big), \tag{23}$$

WAIC has been used to approach the integrated likelihood $[Y|\mathcal{M}_i(\tilde{T})]$ by the so-called Akaike weight (McElreath, 2018):

$$\widehat{[Y|\mathcal{M}_i(\tilde{T})]} = \frac{\exp(-\frac{1}{2}\text{dWAIC}_{i\tilde{T}})}{\sum\limits_{0\le i\le \mathcal{I},\tilde{T}\in\mathcal{T}} \exp(-\frac{1}{2}d\text{WAIC}_{i\tilde{T}})}, \tag{24}$$

where $\text{dWAIC}_{i\tilde{T}} = \text{WAIC}_{i\tilde{T}} - \min\limits_{i,\tilde{T}}\{\text{WAIC}_{i\tilde{T}}\}$. If prior probabilities of models are the same, then Akaike weights and posterior model probabilities coincide. McElreath (2018) also suggested to replace WAIC by other information criteria in (24) and we will test this proposal below by using BIC, DIC and IC.

Harmonic Mean Estimators

For approaching the integrated likelihood, Newton and Raftery (1994) noticed, from Bayes theorem,

$$[Y|\mathcal{M}_i(\tilde{T})]^{-1} = \int [\Theta|Y, \mathcal{M}_i(\tilde{T})][Y|\Theta, \mathcal{M}_i(\tilde{T})]^{-1} d\Theta = \mathbb{E}([Y|\Theta, \mathcal{M}_i(\tilde{T})]^{-1}|Y),$$

and proposed the sample harmonic mean of the likelihood as an estimator of $[Y|\mathcal{M}_i(\tilde{T})]$. Thus, using the weighted posterior sample $\{(\Delta_m^l, w_m^l) : 1 \le m \le M, 1 \le l \le L\}$, the harmonic mean estimator satisfies:

$$\widehat{[Y|\mathcal{M}_i(\tilde{T})]} = \left(\sum_{m=1}^{M}\sum_{l=1}^{L} \frac{1}{[Y|\Theta_m^l, \mathcal{M}_i(\tilde{T})]} w_m^l\right)^{-1}. \tag{25}$$

This estimator is easily computed and is consistent if the sample size tends to infinity but is rather unstable in practice and in theory (the variance of the weights is not finite).

Gelfand and Dey (1994) proposed a generalized version of equation (25), namely,

$$\widehat{[Y|\mathcal{M}_i(\tilde{T})]} = \left(\sum_{m=1}^{M}\sum_{l=1}^{L} \frac{f(\Theta_m^l)}{[Y|\Theta_m^l, \mathcal{M}_i(\tilde{T})] \times [\Theta_m^l|\mathcal{M}_i(\tilde{T})]} w_m^l\right)^{-1}, \tag{26}$$

where $f(.)$ is an importance probability density function. Equation (26) gives an unbiased and consistent estimator of the integrated likelihood if $\int f(\Theta)[Y|\Theta, \mathcal{M}_i(\tilde{T})]^{-1}[\Theta|\mathcal{M}_i(\tilde{T})]^{-1} d\Theta < \infty$. We choose for f an approximation of the posterior distribution of Θ, namely the multivariate normal distribution with mean vector and covariance matrix estimated from the weighted posterior sample $\{(\Theta_m^l, w_m^l) : 1 \le m \le M, 1 \le l \le L\}$. This is supposed to give a stable version of the harmonic mean estimator (Raftery et al., 2006).

4.2.3 | Priors and posterior samples

For the applications, we assume as a prior knowledge that the models are equally weighted. Because several model structures with different sets of parameters were considered, the prior distribution of Θ partly depends on the model structure. These distributions combine vague uniform and Dirac distributions (Dirac distributions are considered for r_0 and p_0 for identifiability issues) and are provided in ESM 1. AMIS is then applied to obtain a weighted posterior sample of size $ML = 5 \times 10^5$ for each candidate model; see Section 4.2.1. Posterior model probabilities, empirical approximations of BMA posterior distributions and other posterior quantities (including predicted infection maps) were calculated by sampling with replacement 10^4 (models×)parameters with respect to parameter weights and, if necessary, with respect to model weights.

5 | APPLICATION TO SIMULATED DATA

A simulation study is carried out by generating three different datasets $\{O^{(g)} : g = 1, 2, 3\} = \{(t_j, \mathbf{x}_j, Y_j^{(g)}) : 1 \leq j \leq J\}$ from two different generative models. We wish to assess which approach for the computation of the posterior model probabilities is reliable. To obtain an assessment relevant for the real case study tackled in the next section, we use the same spatial domain Ω, observations locations $\mathbf{x}_j$ and observation times t_j than in the real data set and, for most of the parameters, we use values close to parameter estimates obtained in the next section. The two models that we used were $\mathcal{M}_0(5.5)$, in which the diffusion is homogeneous ($D_1 = D_2$), and $\mathcal{M}_1(5.5)$, in which the diffusion is heterogeneous. For the latter model, we considered two cases: $D_2 = 0.9D_1$ and $D_2 = 0.1D_1$. Table 1 summarizes model specifications and provides parameter values.

TABLE 1 Specifications of models $\mathcal{M}_0(5.5)$ and $\mathcal{M}_1(5.5)$ from which datasets were generated for the simulation study.

Dataset	Model	Diffusion	Parameter values
O_1	$\mathcal{M}_0(5.5)$	$D_1 = D_2 = D$	$D = 3.2\text{e+}05$, $b = 0.049$, $K = 0.15$,
			$\alpha = 0.26$, $\tilde{\mathbf{x}}_0 = (1.176023\text{e+}06, 6.108375\text{e+}06)$, $\tau_0 = -335$
O_2	$\mathcal{M}_1(5.5)$	$D_1 = 0.9D_2$	$D_1 = 2.0\text{e+}05$, $D_2 = 1.8\text{e+}05$, $b = 0.025$, $K = 0.14$,
			$\alpha = 0.50$, $\tilde{\mathbf{x}}_0 = (1.178962\text{e+}06, 6.113653\text{e+}06)$, $\tau_0 = -650$
O_3	$\mathcal{M}_1(5.5)$	$D_1 = 0.1D_2$	$D_1 = 2.0\text{e+}05$, $D_2 = 2.0\text{e+}04$, $b = 0.025$, $K = 0.14$,
			$\alpha = 0.50$, $\tilde{\mathbf{x}}_0 = (1.178962\text{e+}06, 6.113653\text{e+}06)$, $\tau_0 = -650$

The reliability of each approach for computing posterior model probabilities was assessed by fitting the two models to the three generated datasets and by checking, for each dataset, whether the true model has the largest probability.

Figures 2–4 report the posterior map of the introduction location and the histograms of the marginal posterior distributions of the other parameters when the true model is fitted to data. We can observe that the 'True' parameter values are all in the 95% credible interval.

Afterward, in order to get the BMA posterior distribution, we first consider as a prior knowledge that both candidate models $\mathcal{M}_0(5.5)$ and $\mathcal{M}_1(5.5)$ are equally weighted. Then, we compute posterior model probabilities using evidence-based and Bayesian predictive estimators (see Table 2). While DIC$_2$ (Equations (15) and (16)) was found to correctly identify the 'True' model when applied to the three case studies, both BIC (Equation (14)) and DIC$_1$ (Equations (15) and (17)) led to incorrect model choice. In particular, DIC$_1$ gives nonsensical results, which is probably due to the fact that the posterior distribution is not well summarized by its mean (Gelman et al., 2014). Eventhough the IC (Equation (18)) is a Bayesian predictive information criteria, it does not seem to accurately compute posterior model probability in the case where data was generated from model $\mathcal{M}_1(5.5)$. We notice that the harmonic mean estimators HME$_1$ (Equation (26)) and HME$_2$ (Equation (25)) and the WAIC (Equation (19)) in its two versions WAIC$_1$ (Equation (22)) and WAIC$_2$ (Equation (23)) are able to correctly identify the 'True' model. However, it has been shown that the WAIC relying on data partition can cause difficulties in the case of spatial data (Gelman et al., 2014). In the application to real data, we will favor the harmonic mean estimator of Gelfand and Dey (HME$_1$) because it is the stabilized version of the harmonic mean estimator of Raftery (HME$_2$) and it is based on a well-founded theory.

TABLE 2 Posterior model probabilities obtained for each simulated dataset using different methods to assess the integrated likelihood, namely, those based on BIC, DIC$_1$ and DIC$_2$ (evidence-based estimators) and those based on HME$_1$, HME$_2$, WAIC$_1$, WAIC$_2$ and IC (predictive estimators).

Type of approach	Method	Model	Diffusion	Posterior Model Probability	
				$\mathcal{M}_0(5.5)$	$\mathcal{M}_1(5.5)$
Evidence-based	BIC	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	1.00	0.00
			$D_2 = 0.1D_1$	1.00	0.00
	DIC$_1$	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	<0.01	>0.99
			$D_2 = 0.1D_1$	1.00	0.00
	DIC$_2$	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	0.57	0.43
			$D_2 = 0.1D_1$	0.30	0.70
Predictive	HME$_1$	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	0.12	0.88
			$D_2 = 0.1D_1$	0.08	0.92
	HME$_2$	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	0.80	0.20
			$D_2 = 0.1D_1$	0.23	0.77
	WAIC$_1$	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	0.40	0.60
			$D_2 = 0.1D_1$	0.41	0.59
	WAIC$_2$	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	0.40	0.60
			$D_2 = 0.1D_1$	0.40	0.60
	IC	$\mathcal{M}_0(5.5)$	$D_2 = D_1$	1.00	0.00
		$\mathcal{M}_1(5.5)$	$D_2 = 0.9D_1$	0.00	1.00
			$D_2 = 0.1D_1$	1.00	0.00

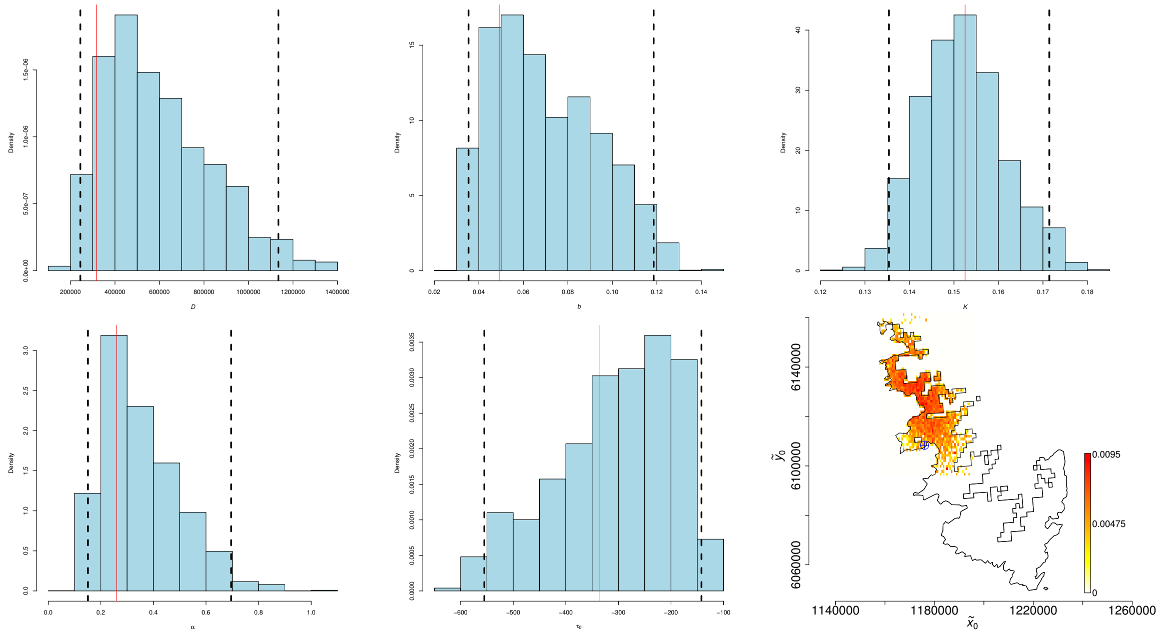


FIGURE 2 Posterior map of the introduction location and histograms of marginal posterior distributions for all the other parameters of model $\mathcal{M}_0(5.5)$ when it is fitted to data $\mathcal{O}_1$. Red line: true value. Dashed black lines 0.025 and 0.975 posterior quantiles. Blue cross on the map: true introduction location. Color palette: posterior probability of the introduction location.

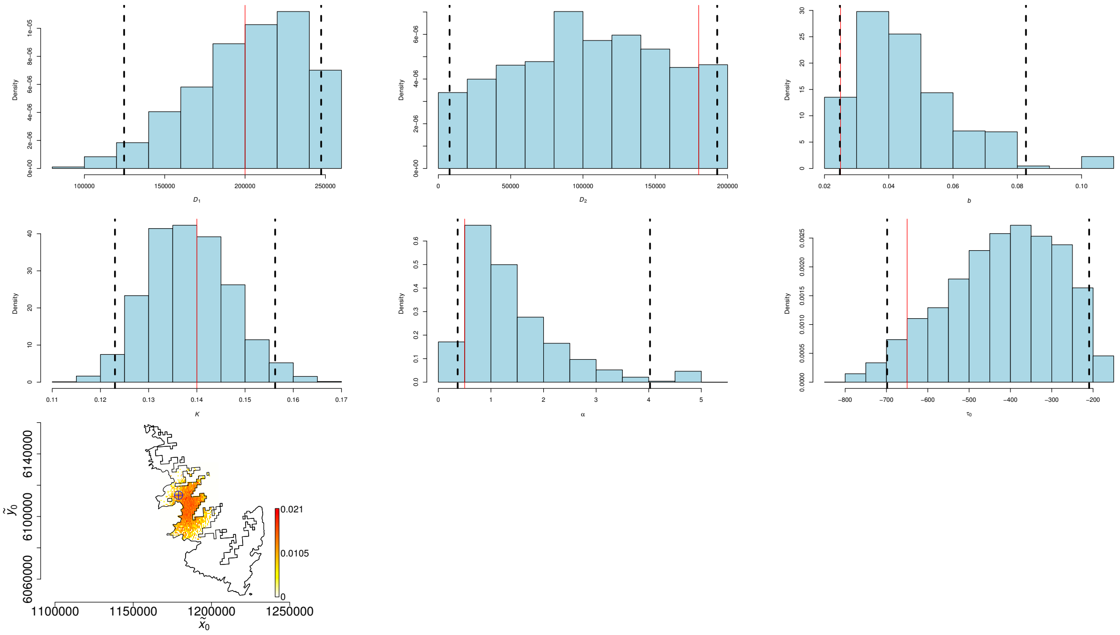


FIGURE 3 Posterior map of the introduction location and histograms of marginal posterior distributions for all the other parameters of model $\mathcal{M}_1(5.5)$ when it is fitted to data $\mathcal{O}_2$. Red line: true value. Dashed black lines 0.025 and 0.975 posterior quantiles. Blue cross on the map: true introduction location. Color palette: posterior probability of the introduction location.

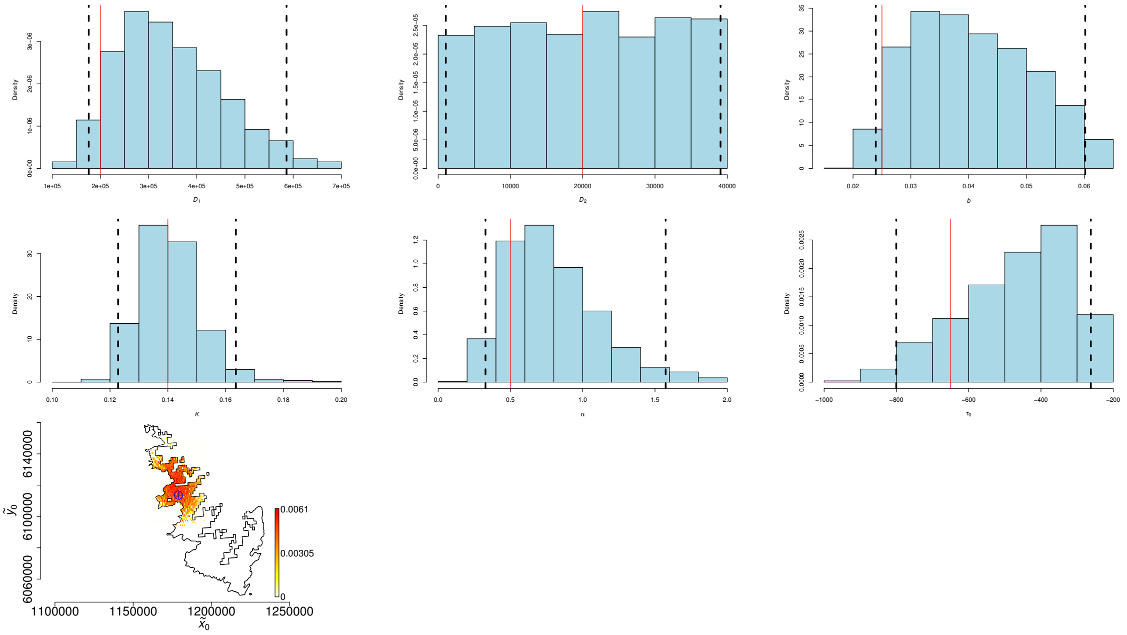


FIGURE 4 Posterior map of the introduction location and histograms of marginal posterior distributions for all the other parameters of model $\mathcal{M}_1(5.5)$ when it is fitted to data O_3. Red line: true value. Dashed black lines 0.025 and 0.975 posterior quantiles. Blue cross on the map: true introduction location. Color palette: posterior probability of the introduction location.

6 | APPLICATION TO *XYLELLA FASTIDIOSA* DATA

6.1 | Model Comparison

In the analysis of the Xf data in Southern Corsica, we first compare BMA methods on the set of 15 models presented in Section 3 for model comparison purposes. For each model, vague and uniform priors were considered for the model parameters. All models were assumed equally likely a priori. The Xf dataset is split into a training dataset that contains all the points sampled between July 2015 and April 2017 and a validation dataset that contains all the points sampled between May 2017 and January 2019. We begin by estimating the joint posterior distribution of each model parameter vector conditionally on the training data, to allow us to later validate our approach by creating forecasts starting in May 2017. AMIS algorithm was applied for this purpose as explained in Section 4.2.1. Convergence diagnostic and goodness-of-fit tests (not shown) were carried out like in Abboud et al. (2019) for each model and yield satisfying results (at least concerning the stabilization of the algorithm). For computing the empirical posterior model probabilities we use various evidence-based and predictive estimators: Table 3 compares posterior model probabilities obtained via BIC (Equation (14)), DIC$_1$ (Equations (15) and (16)), DIC$_2$ (Equations (15) and (17)), IC (Equation (18)), HME$_1$ (Equation (26)), HME$_2$ (Equation (25)), WAIC$_1$ (Equations (19) and (22)) and WAIC$_2$ (Equations (19) and (23)). Most of the used estimators agreed in the selection of model $\mathcal{M}_0(5.5)$ as the best model, associated the lowest posterior model probabilities to the same models, but differed in the subsequent ranking. Furthermore, the DIC$_1$ and IC ranked the model $\mathcal{M}_0(5)$ as the best model. However, the last estimators led to wrong model choice when applied to simulated data. It is also clear from the table that some significant posterior model probabilities were associated to other models such as, $\mathcal{M}_0(5)$, $\mathcal{M}_1(5)$ and $\mathcal{M}_1(5.5)$. The two highest posterior model probabilities obtained with HME$_1$ goes for $\mathcal{M}_0(5.5)$ with $[\mathcal{M}_0(5.5)|Y] = 0.61$ and $\mathcal{M}_0(5.5)$ with $[\mathcal{M}_0(5)|Y] = 0.38$, while the lowest posterior model probabilities are associated to the models in the set $\mathcal{M}_2(\tilde{T})$ and the models with $\tilde{T} = 6$.

TABLE 3 Posterior model probabilities approximated using: BIC, DIC$_1$ (Spiegelhalter et al., 2002), DIC$_2$ (Gelman et al., 2003), IC (Ando, 2011), HME$_1$ (Gelfand and Dey, 1994), HME$_2$ (Raftery, 1996), WAIC$_1$ (Watanabe, 2010) and WAIC$_2$ (Gelman et al., 2014). The highest posterior model probability value obtained with each method is highlighted in yellow.

		Posterior Model Probability							
$\mathcal{M}_i(\tilde{T})$		BIC	DIC$_1$	DIC$_2$	IC	HME$_1$	HME$_2$	WAIC$_1$	WAIC$_2$
$\tilde{T}$ = 4 °C									
	$\mathcal{M}_0$	2.745×10^{-3}	1.032×10^{-8}	2.360×10^{-4}	5.019×10^{-14}	3.809×10^{-4}	2.509×10^{-5}	1.522×10^{-4}	1.684×10^{-4}
	$\mathcal{M}_1$	2.737×10^{-14}	7.353×10^{-17}	4.367×10^{-12}	6.953×10^{-23}	4.816×10^{-15}	4.932×10^{-14}	1.045×10^{-11}	1.152×10^{-11}
	$\mathcal{M}_2$	1.195×10^{-62}	1.916×10^{-67}	1.694×10^{-65}	1.068×10^{-71}	1.123×10^{-64}	3.278×10^{-67}	3.119×10^{-64}	3.465×10^{-64}
$\tilde{T}$ = 4.5 °C									
	$\mathcal{M}_0$	1.231×10^{-1}	1.537×10^{-7}	1.977×10^{-3}	5.560×10^{-13}	2.594×10^{-5}	2.786×10^{-4}	3.292×10^{-3}	3.201×10^{-3}
	$\mathcal{M}_1$	7.885×10^{-11}	2.487×10^{-14}	6.821×10^{-9}	1.676×10^{-20}	8.199×10^{-12}	1.703×10^{-10}	2.681×10^{-9}	3.194×10^{-9}
	$\mathcal{M}_2$	7.430×10^{-63}	8.880×10^{-69}	6.582×10^{-64}	4.602×10^{-75}	2.784×10^{-65}	1.986×10^{-64}	1.210×10^{-63}	1.343×10^{-63}
$\tilde{T}$ = 5 °C									
	$\mathcal{M}_0$	1.412×10^{-1}	8.413×10^{-1}	2.313×10^{-2}	9.513×10^{-1}	3.814×10^{-1}	7.465×10^{-3}	3.393×10^{-2}	3.342×10^{-2}
	$\mathcal{M}_1$	5.818×10^{-7}	5.884×10^{-3}	3.619×10^{-5}	4.558×10^{-2}	1.763×10^{-3}	6.800×10^{-5}	2.222×10^{-5}	2.521×10^{-5}
	$\mathcal{M}_2$	3.697×10^{-58}	2.831×10^{-63}	2.708×10^{-60}	9.547×10^{-68}	4.027×10^{-60}	5.850×10^{-61}	8.180×10^{-60}	8.842×10^{-60}
$\tilde{T}$ = 5.5 °C									
	$\mathcal{M}_0$	7.328×10^{-1}	1.439×10^{-1}	9.723×10^{-1}	1.258×10^{-3}	6.126×10^{-1}	9.913×10^{-1}	9.604×10^{-1}	9.610×10^{-1}
	$\mathcal{M}_1$	1.071×10^{-4}	8.979×10^{-3}	2.285×10^{-3}	1.843×10^{-3}	3.787×10^{-3}	9.099×10^{-4}	2.183×10^{-3}	2.148×10^{-3}
	$\mathcal{M}_2$	4.898×10^{-62}	1.690×10^{-68}	2.503×10^{-64}	2.034×10^{-74}	1.856×10^{-64}	3.896×10^{-65}	8.396×10^{-64}	8.891×10^{-64}
$\tilde{T}$ = 6 °C									
	$\mathcal{M}_0$	8.100×10^{-73}	2.855×10^{-79}	6.007×10^{-73}	4.477×10^{-86}	3.852×10^{-81}	6.049×10^{-76}	3.120×10^{-73}	3.250×10^{-73}
	$\mathcal{M}_1$	4.864×10^{-69}	1.919×10^{-73}	1.720×10^{-67}	4.635×10^{-80}	2.712×10^{-76}	2.385×10^{-70}	6.280×10^{-68}	6.163×10^{-68}
	$\mathcal{M}_2$	3.054×10^{-62}	8.454×10^{-69}	2.693×10^{-64}	5.788×10^{-75}	3.253×10^{-64}	2.692×10^{-65}	1.056×10^{-63}	1.158×10^{-63}

6.2 | Inference

Motivated by the results obtained on simulated data, we use the HME$_1$ for computing the posterior model probabilities, and then inferring shared parameters (the introduction point and the threshold of temperature). BMA marginal and 2D posterior distributions of shared parameters are displayed in Figures and 5 and 6. Figure 5 illustrates the advantage of the BMA since one obtains a posterior distribution of the threshold of temperature instead of a unique selected value as in Abboud et al. (2019). The introduction of Xf tends to be relatively ancient (posterior median: −763.7 months before July 2015, i.e. introduction around 1952; posterior mean −748.7 months) but also relatively uncertain (posterior standard deviation: 156 months). In comparison, the inference of the introduction time provided by the best model is: posterior median −791 months before July 2015, i.e. introduction around 1951; posterior mean −779.02 months; posterior standard deviation: 133 months. We notice that the posterior standard deviation with BMA is larger than the one with the best model, certainly better reflecting the estimation uncertainty.

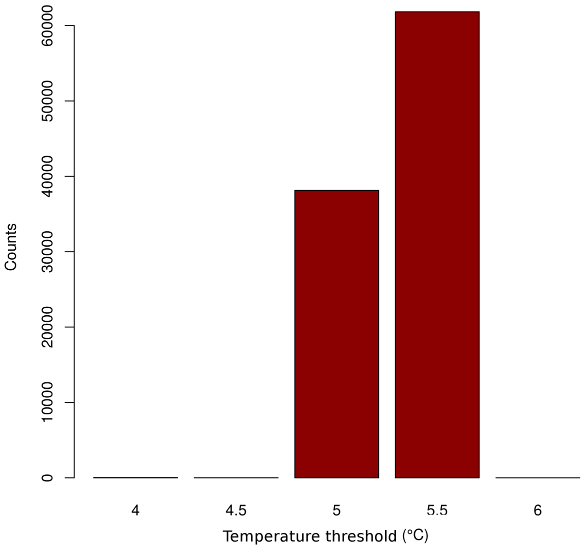
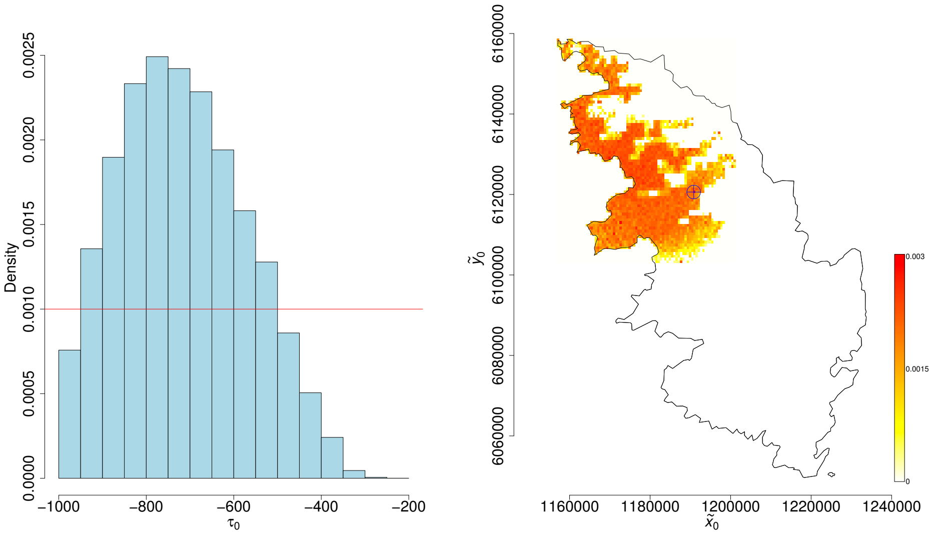


FIGURE 5 BMA marginal posterior distribution of the threshold of temperature $\tilde{T}$.

FIGURE 6 BMA Posterior distributions of the introduction time τ_0 (histogram) and the introduction point $\tilde{\mathbf{x}}_0$ (color palette). The prior for τ_0 was uniform over $[-1000, 0]$ (red line). The value of $\tilde{\mathbf{x}}_0$ having the largest weight in the BMA posterior sample of size 10^5 is indicated by a blue cross. The prior for $\tilde{\mathbf{x}}_0$ was uniform over $\Omega_{\tilde{T}1}$ for model $\mathcal{M}_i(\tilde{T})$, $0 \leq i \leq I$.



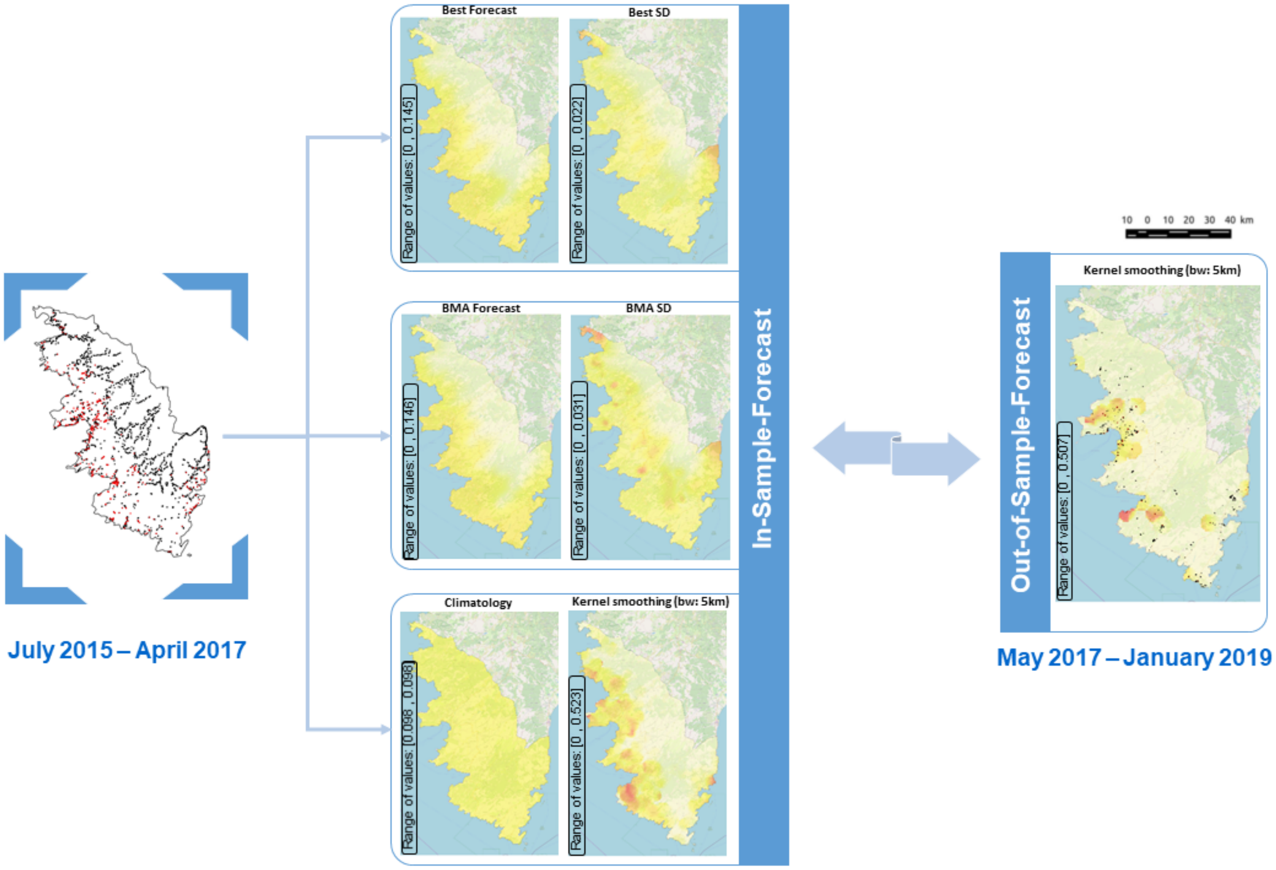

Best Forecast
Best SD
Range of values: [0 , 0.145]
Range of values: [0 , 0.022]
BMA Forecast
BMA SD
Range of values: [0 , 0.146]
Range of values: [0 , 0.031]
Climatology
Kernel smoothing (bw: 5km)
Range of values: [0.098 , 0.098]
Range of values: [0 , 0.523]
In-Sample-Forecast
July 2015 – April 2017
10 0 10 20 30 40 km
Kernel smoothing (bw: 5km)
Out-of-Sample-Forecast
Range of values: [0 , 0.507]
May 2017 – January 2019

FIGURE 7 Comparison between the in-sample forecasts obtained from training data namely, the posterior mean provided by the best model (top), the posterior mean provided by BMA (middle), climatology and kernel smoothing reference forecasts (bottom), and the out-of-sample forecast obtained using a bandwidth of 5 kilometers.

6.3 | Out-of-Sample Predictive behaviour

Figures 7–9 show several in-sample forecasts obtained from training data and out-of sample forecasts obtained from validation data. The out-of-sample forecast is computed as a spatial kernel smoother of validation data, i.e. $\{(\mathbf{x}_j, Y_j) : t_j \in [\text{May 2017–January 2019}]\}$, using the Epanechnikov kernel, which is proportional to $d \mapsto (1 - d^2)1(|d| \leq 1)$ and is applied in the kernel smoother to the geographical distance scaled by different bandwidth values. Thus, in Figure 7, the right panel shows the resulting map for a 5km bandwidth, and this is considered as the *true* time-averaged infection probability between May 2017 and January 2019 that we want to predict. Varying the bandwidth allows us to explore different visions of the *truth*. We compare the out-of-sample forecast for a given smoothing level to four in-sample forecasts computed from training data between July 2015 and April 2017, namely:

- the posterior mean of u computed and averaged over the period May 2017–January 2019, which is provided by the best model selected from and fitted to training data (i.e., the best model is the model with the highest posterior model probability);
- the posterior mean of u computed and averaged over the period May 2017–January 2019, which is provided by BMA applied to training data;
- the so-called climatology forecast (Mason, 2004), which is the mean of $\{Y_j : t_j \in [\text{July 2015–April 2017}]\}$;

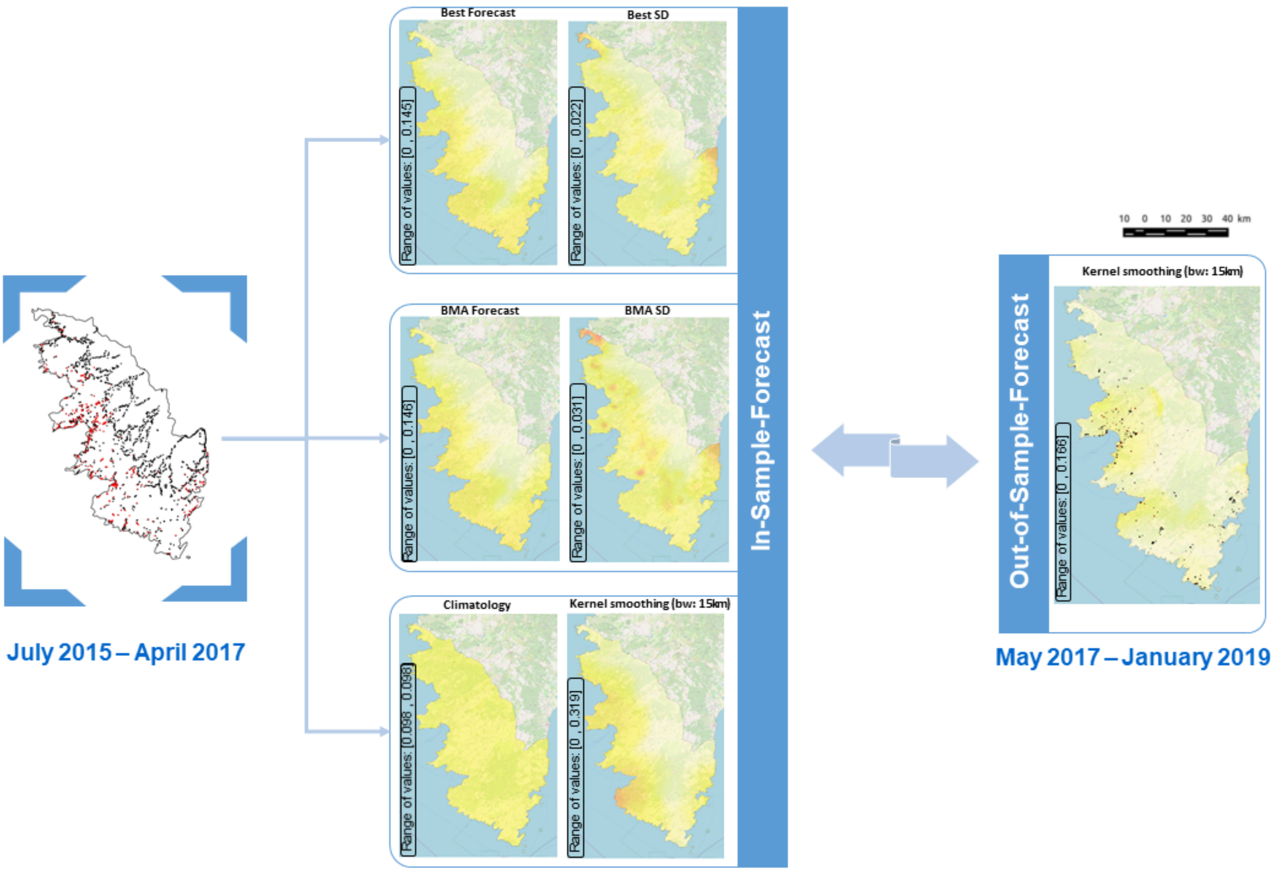

Best Forecast
Best SD
Range of values: [0 , 0.145]
Range of values: [0 , 0.022]
BMA Forecast
BMA SD
Range of values: [0 , 0.146]
Range of values: [0 , 0.031]
Climatology
Kernel smoothing (bw: 15km)
Range of values: [0.098 , 0.098]
Range of values: [0 , 0.319]
In-Sample-Forecast
July 2015 – April 2017
10 0 10 20 30 40 km
Kernel smoothing (bw: 15km)
Out-of-Sample-Forecast
Range of values: [0 , 0.166]
May 2017 – January 2019

FIGURE 8 Comparison between the in-sample forecasts obtained from training data namely, the posterior mean provided by the best model (top), the posterior mean provided by BMA (middle), climatology and kernel smoothing reference forecasts (bottom), and the out-of-sample forecast obtained using a bandwidth of 15 kilometers.

- the spatial kernel smoother of $\{(\mathbf{x}_j, Y_j) : t_j \in [\text{July 2015–April 2017}]\}$, using the Epanechnikov kernel and the same bandwidth than the one used for the out-of-sample forecast.

The root-mean-squared error (RMSE) was computed to measure the predictive behaviour of the BMA, the best model, the climatology and the kernel smoothing over May 2017–January 2019, i.e. how close are these forecasts compared to the "expected truth". This quantity was calculated over a regular square grid covering Ω and with 1km×1km cell size:

$$\text{RMSE} = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(\hat{u}_h - \bar{u}_h^{(b)})},$$

where $\hat{u}_h$ is the average (in time and space) prediction of u in grid cell h over May 2017–January 2019 provided by one of the predictors; $\bar{u}_h^{(b)}$ is the average (in time and space) of u in grid cell h provided by the spatial kernel smoother with bandwidth $b > 0$ applied to validation data; and H is the number of grid cells.

Thus, the RMSE was computed for bandwidths ranging between 2.5 and 25 kilometers. Figure 10 presents the RMSE curves for different bandwidths of kernel smoothing and for all the in-sample forecasts. The BMA and the best model almost have the same accuracy with slightly lower RMSE values for the BMA model. This is also observed from Figures 7–9 indicating that the posterior mean of the BMA forecast and the posterior mean of the best model forecast

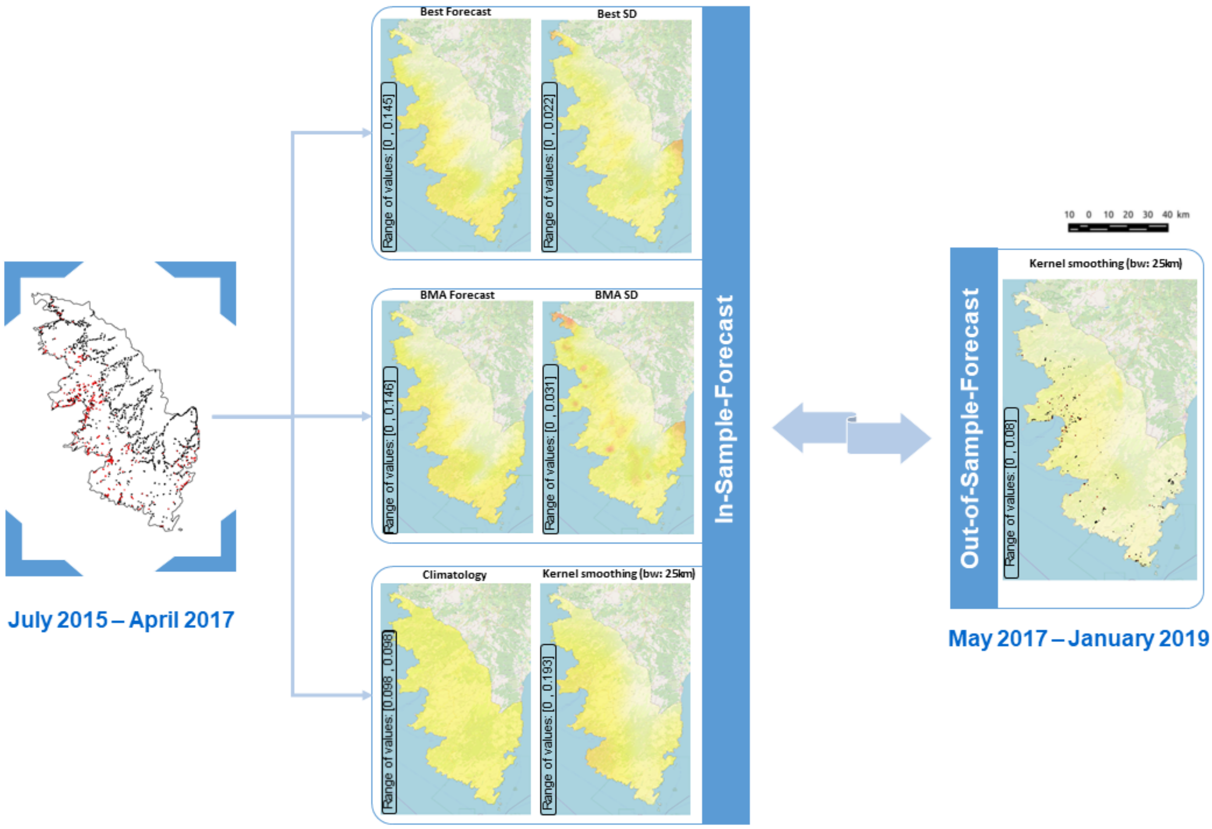
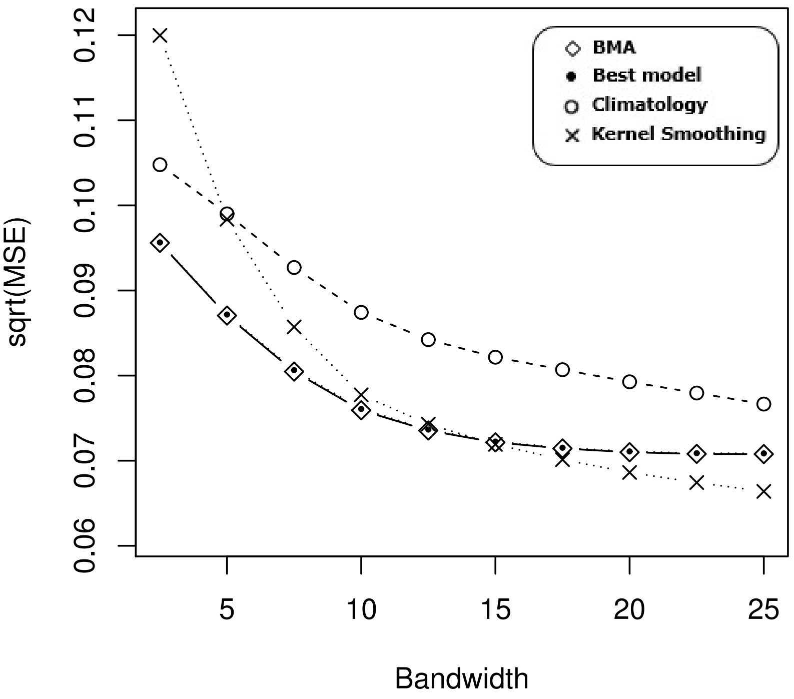

Best Forecast
Best SD
Range of values: [0 , 0.145]
Range of values: [0 , 0.022]
BMA Forecast
BMA SD
Range of values: [0 , 0.146]
Range of values: [0 , 0.031]
Climatology
Kernel smoothing (bw: 25km)
Range of values: [0.098 , 0.098]
Range of values: [0 , 0.193]
In-Sample-Forecast
10 0 10 20 30 40 km
Kernel smoothing (bw: 25km)
Out-of-Sample-Forecast
Range of values: [0 , 0.08]
July 2015 – April 2017
May 2017 – January 2019

FIGURE 9 Comparison between the in-sample forecasts obtained from training data namely, the posterior mean provided by the best model (top), the posterior mean provided by BMA (middle), climatology and kernel smoothing reference forecasts (bottom), and the out-of-sample forecast obtained using a bandwidth of 25 kilometers.

are very similar. However, the standard deviations for BMA are larger than those provided by the best model, certainly better reflecting the uncertainty about the predictions. In addition, BMA better performs than kernel smoothing and climatology for bandwidths up to 15Km. However, when the bandwidth is larger, BMA leads to larger RMSE values than kernel smoothing but still lower values than climatology. We can interpret these results as follows: 1) The climatology, which predicts the same infection probability everywhere, obviously does not account for the major spatially-structured effect of cold temperatures in winter on Xf reproduction and/or propagation. 2) When the bandwidth b is large, the out-of-sample forecast tends to a very smooth function that even yields significantly positive infection probabilities in regions where Xf reproduction and/or propagation is hampered (i.e. cold regions in winter). This bias in what we called the *expected truth* is also encountered in the in-sample forecast based on kernel smoothing and partly explains the advantage of this predictor for large bandwidths. It could also explains why the climatology predictor could improve its performance at even larger bandwidths (for extremely large bandwidths, the kernel smoother and the climatology coincide). 3) For small bandwidths, the high probability areas identified by the kernel smoother applied to training data and to validation data are spatially close but do not exactly coincide. In contrast, the quite smooth mechanistic predictions based on BMA do not predict peaks of infection as observed in the out-of-sample forecast but correctly delineate regions where these peaks can arise.

FIGURE 10 RMSE curves showing the out-of-sample Predictive Behaviour of the BMA compared to the best model, the climatology and the kernel smoothing.



7 | CONCLUSION

In summary, we have presented the use of a PDE-based Bayesian *model-averaging* approach aiming to infer and predict invasive species dynamics using multiple competing models, with application to Xf. In addition to combining model inferences and predictions, the BMA can be used for a direct model comparison by computing model posterior probabilities, which are equal to model weights. To compute these weights we proceed trying various evidence-based and predictive methods proposed in the literature. Most of the used estimators agreed in the selection of model $\mathcal{M}_0(5.5)$ as the best model, and associated the lowest weights to the same models, but differed in the detailed ranking. Motivated by the results obtained on simulated data and based on the well founded theory of the Gelfand and Dey's method, we have shown the results obtained when applying this method to real data in order to average over all model predictions and to infer shaared parameters such as the introduction point and the threshold of temperature.

Our analyses show that the introduction of Xf in Southern Corsica occurred probably near Bastellica, to the east of Ajaccio around 1952 (95%-posterior interval: $[1933, 1979]$), long time before its first detection. The estimation of the introduction site is relatively consistent with the results obtained by Abboud et al. (2019) in the sense that the posterior distributions obtained from both analyses overlap. It is also compatible with the result of Soubeyrand et al. (2018) and Denancé et al. (2017a), who estimated that the date of introduction of this bacterium is relatively ancient. Our BMA approach is expected to0 better reflect the uncertainty about the respective inferences as assessed by credibility intervals showed in the result section. Moreover, we compare the BMA approach to the approach in Abboud et al. (2019) where first the posterior distribution of model's parameter vector was inferred given the threshold of temperature and then, this threshold was chosen using selection criteria. The PDE-based BMA approach presented in this paper, allows to open up to smoothed threshold of temperature by means of the weighted BMA posterior sample of model parameter vectors (i.e. temperature-dependent), which informs us on the temperature threshold by providing an empirical approximation of its marginal BMA posterior distribution.

Based on the ample literature on *model-averaging*, we were expecting this technique would provide ameliorated predictions and more realistic estimate of the uncertainty associated with model predictions than any single model (Hoeting et al., 1999; Draper, 1995; Wintle et al., 2003). However, our application shows that the BMA does not seem to outperform the best model. In this case study, as discussed by Wintle et al. (2003), this may be due in part to the dominance of one or two predictors in all of the models or the lack of complete independence between the training and testing data. Likewise, in our application, the credibility intervals provided with BMA were larger than the ones provided by the best model. This could reflect the fact that BMA has succeeded in better assessing the uncertainty about model predictions, avoiding overconfidence about predictions and misleading decision making obtained when using a single best model. However, to firmly confirm this result, complementary studies should be conducted to calibrate the credibility intervals. Obviously, the deterministic candidate models (Equation (1–7)) that we proposed to describe the dynamics of the pathogen succeed to fairly describe the discrepancies between low and high probability of infection but fails when it comes to the details, e.g., the spatio-temporal disease clusters that can be observed using kernel smoothing with small bandwidth. These details could be implicitly taken into account by coupling the partial differential equation to stochastic or observed covariables that would result in more flexible realizations. Although the population size for the case study of Xf is here relatively large, to damper strong variations from the mean effect, stochastic population-dynamic models (Renshaw, 1993; Kareiva and Shigesada, 1983; Soubeyrand et al., 2009) could allow to relax the deterministic behaviour for the dynamics. In contrast, our parsimonious choice of candidate models that only incorporate the main epidemiological and environmental drivers, provides a concise description of the dynamics of the pathogen, and can be fitted to data in a reasonable computer time. The advantage of this approach is that it can be rapidly applied for endorsing a fast reaction after the detection of a new invasive pathogen as suggested in Abboud et al. (2019). Since the detection of Xf in Europe, several modeling approaches have been implemented to provide more insights on the spread of this invasive pathogen in European environments (Strona et al., 2017; White et al., 2017; Bosso et al., 2016; Godefroid et al., 2018; Soubeyrand et al., 2018; Martinetti and Soubeyrand, 2018). An interesting perspective of our work would be to consider all the predictions obtained from these various models in the case study of Xf in Southern Corsica within the BMA framework. BMA is indeed an approach for taking advantage of different model structures and our work could be extended to more diverse mathematical representations of the infection probability, including stochastic representations.

ACKNOWLEDGEMENTS

This research was funded by a PhD grant INRA-Région PACA (Emplois Jeunes Doctorants 2016-2019), the HORIZON 2020 XF-ACTORS Project SFS-09-2016 and the INRA-DGAL project 21000679. We thank DGAL, Anses, SRAL, FREDON, LNR-LSV and certified laboratories for data collection and data availability. We thank Afidol for their endorsement in the PhD grant.

REFERENCES

Abboud, C., Bonnefon, O., Parent, E. and Soubeyrand, S. (2019) Dating and localizing an invasion from post-introduction data and a coupled reaction–diffusion–absorption model. *Journal of Mathematical Biology*, **79**, 765–789.

Abboud, C., Senoussi, R. and Soubeyrand, S. (2018) Piecewise-deterministic Markov Processes for Spatio-temporal Population Dynamics. In *Statistical Inference for Piecewise-deterministic Markov Processes* (ed. Azaïs, Romain and Bouguet, Florian). ISTE Editions/Wiley.

Anas, O., Harrison, U. J., Brannen, P. M. and Sutton, T. B. (2008) The effect of warming winter temperature on the severity of pierce's disease in the appalachian mountains and piedmont of the southeastern united states. *Plant Health Progress.*, **101094**, 450–459.

Ando, T. (2011) Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, **31**, 13–38.

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

Boone, E. L., Simmons, S. J., Bao, H. and Stapleton, A. E. (2008) Bayesian hierarchical regression models for detecting qtls in plant experiments. *Journal of Applied Statistics*, **35**, 799–808.

Boone, E. L., Ye, K. and Smith, E. P. (2005) Assessment of two approximation methods for computing posterior model probabilities. *Computational Statistics & Data Analysis*, **48**, 221 – 234.

Bosso, L., Russo, D., Febbraro, M. D., Cristinzio, G. and Zoina, A. (2016) Potential distribution of *Xylella fastidiosa* in Italy: a maximum entropy model. *Phytopathologia Mediterranea*, **55**, 62–72.

Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. L. (2008) Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, **18**, 125–135.

Bugallo, M. F., Martino, L. and Corander, J. (2015) Adaptive importance sampling in signal processing. *Digital Signal Processing*, **47**, 36–49.

Burnham, K. P., White, G. C. and Anderson, D. R. (1995) Model selection strategy in the analysis of capture-recapture data. *Biometrics*, **51**, 888–898.

Chapman, D. S., White, S. M., Hooftman, D. A. and Bullock, J. M. (2015) Inventory and review of quantitative models for spread of plant pests for use in pest risk assessment for the EU territory. *EFSA Supporting Publications*, **12**.

Clyde, M. and George, E. I. () Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 681–698.

Cornuet, J., Marin, J.-M., Mira, A. and Robert, C. P. (2012) Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, **39**, 798–812.

Denancé, N., Cesbron, S., Briand, M., Rieux, A. and Jacques, M.-A. (2017a) Is *Xylella fastidiosa* really emerging in France? . In *1st Annual Conference of the EuroXanth - COST Action Integrating Science on Xanthomonadaceae for integrated plant disease management in Europe* (eds. J. Costa and R. Koebnik), vol. 7. Coimbra, Portugal: EuroXanth.

Denancé, N., Legendre, B., Briand, M., Olivier, V., Boisseson, C., Poliakoff, F. and Jacques, M.-A. (2017b) Several subspecies and sequence types are associated with the emergence of *Xylella fastidiosa* in natural settings in France. *Plant Pathology*, **66**, 1054–1064.

Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 45–70.

Eicher, T. S., Papageorgiou, C. and Raftery, A. E. (2011) Default priors and predictive performance in bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, **26**, 30–55.

Evans, L. C. (1998) *Partial differential equations*, vol. 19 of *Graduate studies in mathematics*. Providence, Rhode Island: American Mathematical Society.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G. and Lemey, P. (2014) The early spread and epidemic ignition of hiv-1 in human populations. *Science*, **346**, 56–61.

Fernández, C., Ley, E. and Steel, M. F. J. (2002) Bayesian modelling of catch in a north-west atlantic fishery. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 257–280.

Gatenby, R. A. and Gawlinski, E. T. (1996) A reaction-diffusion model of cancer invasion. *Cancer research*, **56**, 5745–5753.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**, 501–514.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian data analysis*. Texts in statistical science series. New York: Chapman & Hall/CRC, 2 edn.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for bayesian models. *Statistics and computing*, **24**, 997–1016.

Gelman, A., Roberts, G. O., Gilks, W. R. et al. (1996) Efficient metropolis jumping rules. *Bayesian statistics*, **5**, 599–608.

George, E. I. and McCulloch, R. E. (1993) Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Godefroid, M., Cruaud, A., Streito, J.-C., Rasplus, J.-Y. and Rossi, J.-P. (2018) Climate change and the potential distribution of *Xylella fastidiosa* in Europe. *bioRxiv*.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors. *Statist. Sci.*, **14**, 382–417.

Huld, T. A., Šúri, M., Dunlop, E. D. and Micale, F. (2006) Estimating average daytime and daily temperature profiles within europe. *Environmental Modelling & Software*, **21**, 1650–1661.

Jones, D. R. and Baker, R. H. A. (2004) Introductions of non-native plant pathogens into great britain, 1970–2004. *Plant Pathology*, **56**, 891–910.

Kareiva, P. and Shigesada, N. (1983) Analyzing insect movement as a correlated random walk. *Oecologia*, **56**, 234–238.

Lamon, E. C. and Clyde, M. A. (2000) Accounting for model uncertainty in prediction of chlorophyll a in lake okeechobee. *Journal of Agricultural, Biological, and Environmental Statistics*, **5**, 297–322.

Lanzarone, E., Pasquali, S., Gilioli, G. and Marchesini, E. (2017) A Bayesian estimation approach for the mortality in a stage-structured demographic model. *Journal of mathematical biology*, **75**, 759–779.

Leamer, E. (1978) *Specification searches: Ad hoc inference with nonexperimental data*, vol. 53. John Wiley & Sons Incorporated.

Madigan, D., Gavrin, J. and Raftery, A. E. (1995) Enhancing the predictive performance of bayesian graphical models. *Communications in Statistics-Theory and Methods*, **24**, 2271–2292.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Martinetti, D. and Soubeyrand, S. (2018) Identifying lookouts for epidemio-surveillance: application to the emergence of *Xylella fastidiosa* in France. Submitted.

Mason, S. J. (2004) On using "climatology" as a reference strategy in the brier and ranked probability skill scores. *Monthly Weather Review*, **132**, 1891–1895. URL: `https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2`.

McElreath, R. (2018) Overfitting, regularization, and information criteria. In *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Newton, M. A. and Raftery, A. E. (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**, 3–26.

Oehler, V. G., Yeung, K. Y., Choi, Y. E., Bumgarner, R. E., Raftery, A. E. and Radich, J. P. (2009) The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data. *Blood*, **114**, 3292–3298.

Okubo, A. (1980) *Diffusion and ecological problems: mathematical models*, vol. 10 of *Interdisciplinary Applied Mathematics*. New York: Springer-Verlag.

Okubo, A. and Levin, S. (2002) *Diffusion and ecological problems - Modern Perspectives*. New York: Springer-Verlag, 2 edn.

Parkinson, D. and Liddle, A. R. (2013) Bayesian model averaging in astrophysics: a review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **6**, 3–14.

Peterson, R. O., Vucetich, J. A., Page, R. E., Chouinard, A. et al. (2003) Temporal and spatial aspects of predator–prey dynamics. *Alces*, **39**, 215–232.

Protter, MH and Weinberger, HF (1967) *Maximum Principles in Differential Equations*. New Jersey: Prentice-Hall, Englewood Cliffs.

Raftery, A. E. (1996) Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Raftery, A. E., Kárný, M. and Ettler, P. (2010) Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, **52**, 52–66.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. (2006) Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics 8*, 1–45.

Raftery, A. E. and Zheng, Y. (2003) Discussion. *Journal of the American Statistical Association*, **98**, 931–938.

Renshaw, E. (1993) *Modelling biological populations in space and time*, vol. 11. Cambridge: Cambridge University Press.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, **7**, 110–120.

Roberts, H. V. (1965) Probabilistic prediction. *Journal of the American Statistical Association*, **60**, 50–62.

Roques, L. (2013) *Modèles de réaction-diffusion pour l'écologie spatiale: Avec exercices dirigés*. Editions Quae.

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. (2016) Using genetic data to estimate diffusion rates in heterogeneous landscapes. *Journal of mathematical biology*, **73**, 397–422.

Rubin, D. B. and Schenker, N. (1986) Efficiently simulating the coverage properties of interval estimates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **35**, 159–167.

Schwarz, G. et al. (1978) Estimating the dimension of a model. *The annals of statistics*, **6**, 461–464.

Shigesada, N. and Kawasaki, K. (1997) *Biological invasions: theory and practice*. Oxford series in ecology and evolution. Oxford, New York: Oxford University Press, 1 edn.

Shigesada, N., Kawasaki, K. and Takeda, Y. (1995) Modeling Stratified Diffusion in Biological Invasions. *The American Naturalist*, **146**, 229–251.

Sidman, A. H., Mak, M. and Lebo, M. J. (2008) Forecasting non-incumbent presidential elections: Lessons learned from the 2000 election. *International Journal of Forecasting*, **24**, 237 – 258.

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

Soubeyrand, S., de Jerphanion, P., Martin, O., Saussac, M., Manceau, C., Hendrikx, P. and Lannou, C. (2018) What dynamics underly temporal observations? Application to the emergence of *Xylella fastidiosa* in France: probably not a recent story. *New Phytologist*.

Soubeyrand, S., Neuvonen, S. and Penttinen, A. (2009) Mechanical-statistical modeling 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**, 427–434.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Strona, G., Carstens, C. J. and Beck, P. S. (2017) Network analysis reveals why *Xylella fastidiosa* will persist in Europe. *Scientific Reports*, **7**, 71.

Turchin, P. (1998) *Quantitative Analysis of Movement: measuring and modeling population redistribution in plants and animals*. Sunderland, Massachusetts: Sinauer.

Viallefont, V., Raftery, A. E. and Richardson, S. (2001) Variable selection and bayesian model averaging in case-control studies. *Statistics in Medicine*, **20**, 3215–3230.

Volinsky, C. T., Madigan, D., Raftery, A. E. and Kronmal, R. A. (1997) Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **46**, 433–448.

Watanabe, S. (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

— (2013) A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, **14**, 867–897.

White, S. M., Bullock, J. M., Hooftman, D. A. P. and Chapman, D. S. (2017) Modelling the spread and control of *xylella fastidiosa* in the early stages of invasion in apulia, italy. *Biological Invasions*, **19**, 1825–1837.

Wintle, B. A., McCarthy, M. A., Volinsky, C. T. and Kavanagh, R. P. (2003) The use of bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, **17**, 1579–1590.

Yeung, K. Y., Bumgarner, R. E. and Raftery, A. E. (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, **21**, 2394–2402.

Yin, G. and Yuan, Y. (2009) Bayesian model averaging continual reassessment method in phase i clinical trials. *Journal of the American Statistical Association*, **104**, 954–968.

4.3. Key Points of Chapter 4

- The advantage of the BMA approach compared to the approach presented in Chapter 3 is that it can better reflect the uncertainty about model predictions. BMA helps avoiding overconfidence predictions and misleading decision making obtained when using a single model.

- We discuss the results obtained about the performance of the BMA predictions compared to the predictions obtained from the "best" model. We were surprised that the difference was so insignificant for our case study, with regards to theoretical considerations in favor of BMA.

- To submit this article, I still have to evaluate the integrated likelihood using the non-biased estimator mentioned in Section 2.2.2.2.4. This estimator is naturally provided in a classical importance sampling by averaging over computed unnormalized weights. Its application in the framework of AMIS has first to be investigated, and then compared to the HME_1 to identify potential discrepancies, if any, with respect to the distribution of posterior model probabilities.

5. Piecewise-deterministic Markov Processes a way to gain more realism

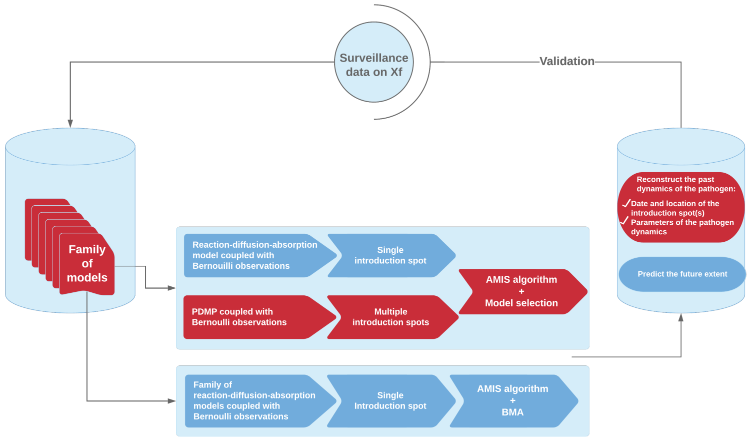
The chapter is divided into two parts: The first part introduces a part of a book chapter[1] published by Wiley on the 6^{th} of August 2018, and cited hereafter. The second part introduces the first analysis obtained on a work in progress, and discusses the results and the difficulties.

Table of contents

[1] C. Abboud, R. Senoussi, and S. Soubeyrand. Piecewise-deterministic Markov Processes for Spatio-temporal Population Dynamics. In Azaïs, Romain and Bouguet, Florian, editor, *Statistical Inference for Piecewise-deterministic Markov Processes*. ISTE Editions/Wiley, 2018

5.1. Graphical Summary

Major Components of The Chapter


Surveillance data on Xf
Validation
Family of models
Reaction-diffusion-absorption model coupled with Bernouilli observations
Single introduction spot
PDMP coupled with Bernoulli observations
Multiple introduction spots
AMIS algorithm + Model selection
Reconstruct the past dynamics of the pathogen:
✓Date and location of the introduction spot(s)
✓Parameters of the pathogen dynamics
Predict the future extent
Family of reaction-diffusion-absorption models coupled with Bernoulli observations
Single Introduction spot
AMIS algorithm + BMA

How to ⋯

I- Biological Questions

- Know when and where the strains of Xf that triggered the epidemic observed since 2015 in Corsica were introduced in this region.
- Get more insights on the pathogen epidemiology in order to adapt surveillance strategies.

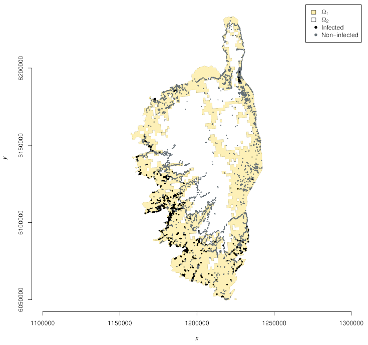
II- Methodological Questions

- Extend the modeling and inferring framework of the last two chapters, when multiple introductions potentially occurs.
- Adapt this framework to allow inferring model parameters in a reasonable time span.

Methodological Ingredients

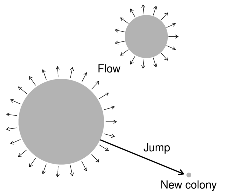
Surveillance Data on Xf with Binary Records

- ~**17000** plants sampled **since 2015** of which **1000** have been diagnosed as **infected** (real-time PCR).
- For those ~ 17000 plants, **geographic coordinates** and **sampling dates** are available.
- T : average of the minimum daily temperature over January and February b/n 1995 and 2003 (Map of T with 1 km grid resolution on the right) thresholded by $\bar{T} = 5$℃.



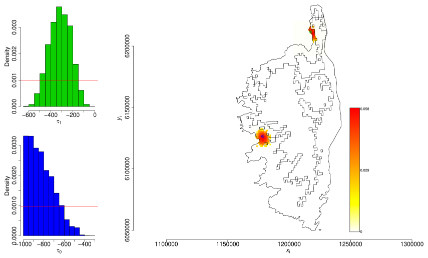
PDMP with Multiple Introductions

- Aggregate multiple PDMP with multiple introductions based on a coupled reaction-diffusion-absorption equation.
- Use the AMIS algorithm for model's parameter estimation in the framework of the mechanistic-statistical approach.
- Jointly infer initial conditions and parameters of the dynamics.

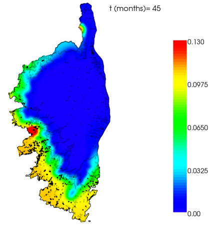


Main Results

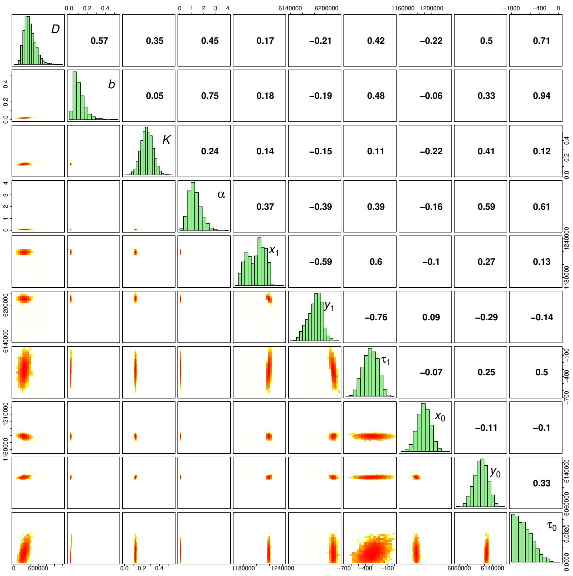
Estimation of the marginal distributions of two introduction dates and the 2D posterior distribution of introduction locations.



Numerical Solution of the model with two introductions for values of Θ associated with the highest posterior weight.



Estimation of the marginal posterior distribution of parameters (panels in the diagonal) and 2D posterior distributions of parameters (lower triangle panels). Correlation cœfficients are provided in the upper triangle panels.



Conclusion & Perspectives

Conclusion

- In this Chapter, we show that the *coalescing colony* model can be formulated as a spatio-temporal PDMP, and then propose a PDE-based PDMP to incorporate into the models of Chapters 3 and 4 the possibility of multiple introductions.
- We do not infer the PDMP as such, but we inferred in a Bayesian framework, Initial conditions and parameters related to diffusion, reproduction, and mortality, under a simplified problem and implementation.
- Lower standard deviation values where obtained for model parameters compared to the values obtained in Chapter 3.
- Our preliminary results raises questions about whether the AMIS algorithm will still be tractable in a high dimensional problem such as a spatio-temporal PDMP with multiple introductions.

Perspectives

- Infer the PDMP as a whole and alternative approaches.
- Apportion the different sources of uncertainty.
- Use additional information brought by genetic data.
- Adapt AMIS to reduce computational time and reach the stable state faster.

5.2. Book Chapter

7

Piecewise-deterministic Markov Processes for Spatio-temporal Population Dynamics

7.1. Introduction

7.1.1. *Models for Population Dynamics*

Population dynamics is the study of the structure, the pattern and the biological and environmental drivers of populations. Studies of population dynamics are carried out at various scales, from the microscopic scale to the global scale, and are particularly relevant in ecology and epidemiology.

Numerous and diverse modeling approaches have been proposed to mathematically represent population dynamics. These modeling approaches are based on diverse mathematical tools adapted to (i) different resolutions at which the population dynamics are considered (e.g. individuals, groups, presence in quadrats, and numbers of individuals in districts), and (ii) different levels of perceptions (e.g. the population itself, its averaged characteristics, or more generally aggregated functions of the population patterns). For instance, ODEs were used to describe the average growth of populations [TUR 03, chap.3], branching processes were used to model the growth and adaptation of populations [MÉL 11], PDEs and integrodifferential equations were used to represent the spatio-temporal intensity of populations with local and non-local dispersal capacities [ROQ 10, ALF 13], SDEs were used to model trajectories of individuals [GLO 15], temporal point processes were used to build birth-death models [CHA 06], spatio-temporal point processes were used to model the temporal evolution of the spatial pattern of individuals forming a population [SOU 11], stochastic Markovian areal processes were used to model large-scale dynamics [SOU 09b], regressions (eventually including

Chapter written by Candy ABBOUD, Rachid SENOUSSI and Samuel SOUBEYRAND.

auto-regressive components) were used to take into account the effect of environmental variables on population characteristics [BOR 17].

Suppose that we are interested in fitting a spatio-temporal population dynamic model to data. There is, like in many other application fields, a trade-off between model realism and estimation complexity. For example, fitting a population dynamic model essentially constructed from a partial differential equation containing a few parameters [SOU 14b] is generally easier than fitting a (more flexible and realistic) hierarchical stochastic spatio-temporal Markovian model including a few parameters but numerous latent variables [SOU 09b]. In this example, two extreme cases are considered: (i) a model with a deterministic behavior and a few degrees of freedom, which may yield poor goodness-of-fit, and (ii) a model with a stochastic behavior and lots of degrees of freedom, which may induce identifiability issues. Intermediate models are required to achieve rapid, realistic and consistent inferences. Spatio-temporal PDMPs can play this role.

7.1.2. *Spatio-temporal PDMP for Population Dynamics*

Spatio-temporal PDMPs can be occasionally encountered in the theoretical and quantitative population dynamic literature, but these models are generally not called PDMPs. Here, we give three examples of spatio-temporal PDMPs built at three different levels: the population, the metapopulation (which is a set of populations) and the individual. These processes are illustrated in Figure 7.1.

The coalescing colony model [SHI 95], which was developed to represent stratified diffusion in biological invasions, is a PDMP. Stratified diffusion typically consists of two components: neighborhood diffusion and long-distance dispersal. The former component is modeled in the coalescing colony model by a deterministic expansion of colonies (this is the *flow*); The latter component is modeled by the random Markovian generation of new colonies away from the existing colonies (this is the *jump* process). The coalescing colony model was developed to investigate the impact of stratified dispersal on the rate of expansion of populations with several propagation modes.

The metapopulation epidemic model proposed in [SOU 09a] is another example of spatio-temporal PDMP representing a population dynamic. Here, the population of interest is a pathogen population whose hosts are spread in a set of disconnected areas, called host patches. In this model, host patches can be either healthy or infected by the pathogen; When a host patch is infected, the local pathogen population grows in a deterministic way (this is the *flow*); Infected patches can infect distant healthy patches in a stochastic manner (this is the *jump* process; the pathogen *jumps* from infected patches to healthy patches). The metapopulation epidemic model was fitted to presence/absence data of the pathogen in host patches at the end of successive epidemic seasons.

PDMPs can also provide concise mathematical descriptions of trajectories of individuals. Examples of such models are given in [CAI 17, chap. 1] under the term velocity-jump models. These models were used to carry out a statistical analysis of the expansion of the cane toad using data obtained by monitoring successive daily locations of a sample of toads. In the simplest model, each individual randomly alternates between encamped and running modes, whose durations are independently and exponentially distributed (this is the *jump* process). When an individual *jumps* towards a new running mode, the direction is randomly drawn in a specified distribution. When an individual is running, its movement is deterministic and linear given the random direction of the movement (this is the *flow*).

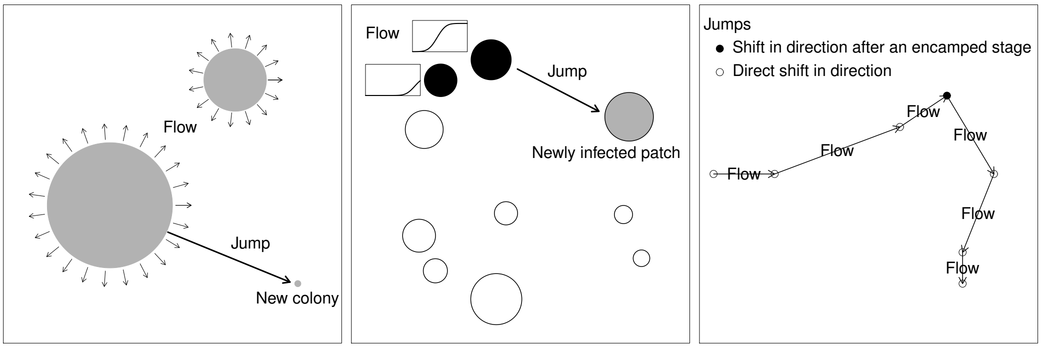


Figure 7.1: Illustrations of the flows and jumps for the coalescing colony model (left), the metapopulation epidemic model (center) and the simple velocity-jump model.

7.1.3. *Chapter Contents*

In the following, we describe three contexts where PDMPs arise for describing population dynamics at the population level, the metapopulation level and the individual level, respectively. Section 7.2 shows how the coalescing colony model was built and how it can be formulated as a PDMP. It also introduces a spatio-temporal PDMP based on a reaction-diffusion equation that could be used to model the dynamic of an invading pathogen (e.g. *Xylella fastidiosa* in Corsica) that might have been introduced at multiple points in space and time. Section 7.3 presents the metapopulation epidemic model mentioned above and gives details about how it was fitted to data. Section 7.4 describes a theoretical framework for building trajectory models with jumps, including PDMPs.

7.2. Stratified Dispersal Models

In this section, we briefly review some mathematical models describing spatio-temporal dynamics of populations. We are especially interested in the dispersal modes incorporated in these models. Thus, we will consider some reaction-diffusion models including only short-distance dispersal processes, and coalescing colony models including both short-distance and long-distance dispersal processes. We will show how the latter model can be formulated as a PDMP. Finally we will present an original reaction-diffusion-based PDMP describing invasion dynamics with multiple introductions.

7.2.1. *Reaction-diffusion Equations for Modeling Short-distance Dispersal*

There are typically three stages arising successively during a biological invasion process: (1) establishment where a few individuals arrive and succeed to settle, (2) linear expansion when the invasion occurs by neighborhood diffusion as in this section or biphasic expansion when the invasion is driven by stratified diffusion (see Section 7.2.2), and (3) concentration of the invasive species in the area of invasion until saturation [COL 04, RIC 00]. When one aims to model dispersal phenomena such as biological invasions, reaction-diffusion equations are frequently used and have been exploited in many domains, especially in medicine, ecology and epidemiology [GAT 96, ROQ 13, MUR 96]. Reaction-diffusion equations are partial differential equations of parabolic type [EVA 98]. Here, we describe some reaction-diffusion equations, in which dispersal is considered as a random diffusion process.

Random diffusion at the population level can be derived from random walks at the individual level. Random walks are often used to describe invasions by species that move via short-distance dispersal. Basic random walk models describe the path of an individual moving in a spatial domain via a succession of random steps. Typically, in a unidimensional space, as illustrated in Figure 7.2, the individual located at x can move to the left and reach $x - d$ with probability $\mathbb{P}_L$, move to the right and reach $x + d$ with probability $\mathbb{P}_R$ or stay at the same place with probability $\mathbb{P}_S = 1 - \mathbb{P}_L - \mathbb{P}_R$. Such a microscopic and individual-based description of movements can be used to obtain diffusion equations at the population level [ROQ 13, SHI 97, SKE 51]. In particular, the 1D random walk without directional bias and with constant and non-persistant increments leads to the following form of diffusion equation: $\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}$, where u is the density of population.

In 1937, Fisher analyzed the rate of advance of advantageous genes with a PDE [FIS 37], which has been generalized into:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + \underbrace{u(r - bu)}_{f(u)}, \quad t \geq 0 \qquad [7.1]$$

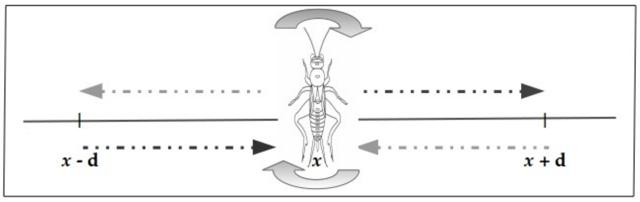


Figure 7.2: Unidimensional random walk model.

where $u = u(t, x)$ is the frequency of the advantageous gene at time t and spatial location x in a unidimensional space; $D > 0$ is the coefficient measuring the rate of dispersal; r stands for the intrinsic growth rate of the species; and b corresponds to the coefficient measuring the effect of intra-specific competition; $f(u)$ is the population growth term.

In the line with Fisher's work, Skellam [SKE 51] proposed two-dimensional PDEs for describing population dynamics. The so-called Skellam model, in particular, allowed him to theoretically study population spread with Malthusian growth. This model incorporates two terms, namely the population dispersal term and the population growth term, and assumes that there is no intra-specific competition:

$$\frac{\partial u}{\partial t} = D\Delta u + ur, \quad t \geq 0 \qquad [7.2]$$

Figure 7.3 presents the solution of Equation [7.2] in a two-dimensional space, for specific values of parameters, initial conditions and boundary conditions.

Positive wavefront type-solutions exist for Equation [7.2]. One simplified form of a traveling wave (in a unidimensional space) is a function of the form:

$$u(t, x) = U(x - ct)$$

where $c \in \mathbb{R}$ is the speed of the front $U \in \mathcal{C}^2(\mathbb{R})$. Note that a traveling wavefront can be defined not only when $t > 0$ but also for any $t \in \mathbb{R}$.

Skellam showed that the rate of spread at the front of the population range asymptotically approaches $c_0 = 2\sqrt{rD}$ when a small population is initially introduced at the origin. Furthermore, Luther [LUT 06] and Kolmogorov et al. [KOL 37] were the first to prove the existence of wavefront type-solutions for a diffusion equation with a logistic growth term $f(u) = ru(1 - u)$ (Fisher-KPP). Kolmogorov et al. showed that some initial distributions converge asymptotically to a traveling wave propagating to the right with a well defined, constant speed $c = 2\sqrt{rD}$. When the growth term includes an Allee effect as follows: $f(u) = ru(1 - u)(u - \theta)$, where $\theta \in]0, 1[$ is the Allee effect parameter, then there

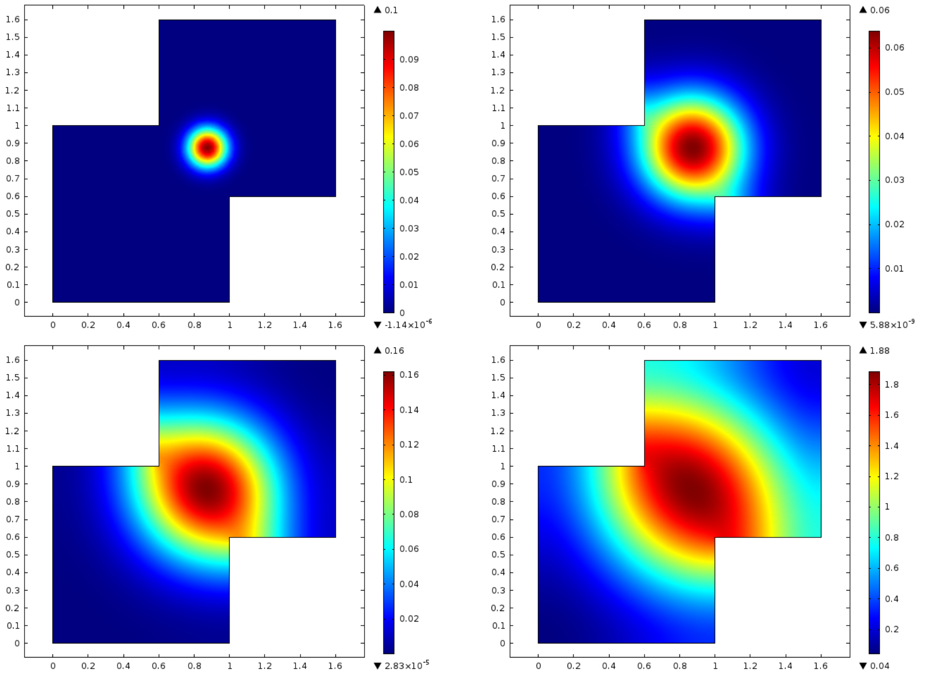


Figure 7.3: Numerical solution $u(t, \mathbf{x})$ of Skellam model [7.2] in a bi-dimensional space (where $\mathbf{x} = (x, y)$) with Neumann boundary conditions, at time 0 (top left), 3 (top right), 6 (bottom left) and 12 (bottom right). The dispersal coefficient and the intrinsic growth rate were fixed at $(D, r) = (5 \times 10^{-3}, 0.5)$. The initial condition was $u(0, \mathbf{x}) = 0.1 \exp(-(10\|\mathbf{x} - \tilde{\mathbf{x}}_0\|)^2)$, where $\tilde{\mathbf{x}}_0 = (\tilde{x}_0, \tilde{y}_0) = (0.8, 0.8)$.

exists a unique positive wavefront-type solution with $\lim_{x \longrightarrow -\infty} U = 1$, $\lim_{x \longrightarrow +\infty} U = 0$. In addition, the speed of the front is [HAD 75, ROT 81, LEW 93]:

$$c = \sqrt{2rD}(\frac{1}{2} - \theta) \qquad [7.3]$$

7.2.2. *Stratified Diffusion*

The models introduced above are generally not adapted to describe the dynamics of populations that expand their range not only by neighborhood dispersal but also by long-distance dispersal, which can corresponds to rare but significant events. The term *stratified diffusion* was used to describe this twofold dispersal process [HEN 89].

Shigesada et al. [SHI 95] proposed stratified diffusion models (derived from Skellam's equation for neighborhood dispersal) and studied their properties. In theses models, the population of interest is in a homogeneous environment and expands its range continuously in time for neighborhood dispersal and at discrete random times for long-distance dispersal (i.e. colonization events). Two frameworks were considered: (i) the nuclei of colonization created by long-distance migrants are located far enough to assume that their ranges do not overlap, mutually and with the mother colony, for a long time; (ii) the nuclei of colonization created by long-distance migrants merge with the mother colony as soon as they *touch* the mother colony (because of their own expansion and the expansion of the mother colony), but the merging of two nuclei of colonization is neglected. Framework (ii) led to the *coalescing colony model* [SHI 95] that we revisit in the next section by incorporating an Allee effect.

7.2.3. *Coalescing Colony Model with Allee effect*

Model description and properties

Suppose that a few individuals invade a given location of the 2D Euclidean space at $t = 0$, succeed to settle, and form a so-called *mother colony* with a disk shape whose radius increases at a constant rate c by neighborhood diffusion (the establishment phase is neglected). By setting $c = \sqrt{2rD}(\frac{1}{2} - \theta)$, the expansion of the mother colony is an approximation of the population expansion governed by the following PDE incorporating an Allee effect (see Equation [7.3]):

$$\frac{\partial u}{\partial t} = D\Delta u + ru(1-u)(u-\theta),$$

given adequate initial conditions.

The expansion of the mother colony is augmented by long-distance dispersal events generating child colonies. More precisely, the mother colony releases long-distance dispersers that settle at a distance $L > 0$ of the border of the mother colony and produce child colonies. The rate of generation of child colonies, say $\tilde{\lambda}$, is assumed to depend on the current radius z of the mother colony. Typically, $\tilde{\lambda}$ is a non-decreasing function of z. Shigesada et al. considered three cases:

- $\tilde{\lambda}(z) = \lambda_0$, i.e. the mother colony produces long-distance migrants at a constant rate;
- $\tilde{\lambda}(z) = \lambda_1 z$, i.e. the mother colony produces long-distance migrants at a time-varying rate proportional to its perimeter;
- $\tilde{\lambda}(z) = \lambda_2 z^2$, i.e. the mother colony produces long-distance migrants at a time-varying rate proportional to its area.

Additionally, every child colony expands its range circularly at the constant rate c, like the mother colony, but do not release long-distance migrants. When the mother colony and a child colony collide, the area covered by the child colony is instantaneously assigned to the mother colony, which remains a disk with same center but with a larger radius. Collisions between child colonies are neglected. An illustration of this process is provided in Figure 7.4.

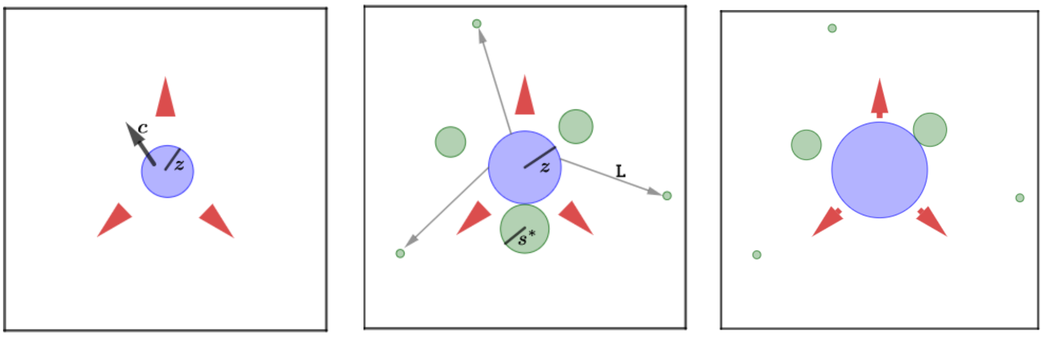


Figure 7.4: Illustration for the coalescing colony model. First, from $t = 0$, the range of the mother colony (disks) expands by short-distance dispersal with a constant rate c (left). Then, the mother colony generates long-distance dispersers to the distance L from its border at the rate $\tilde{\lambda}(z(t))$. The child colonies (circles) expands their range at the rate c until they collides with the mother colony after a period of duration $\frac{L}{2c}$. Finally (right), at the time of coalescence, the range of the blue including the green colony is immediately reshaped into a circular pattern while the total area of both colonies remains the same.

The coalescing colony model is characterized by the following properties [SHI 95]. The expectation of the number of child colonies having radius s at time t, say $n(s,t)$, satisfies the following von Foerster equation and initial / boundary conditions:

$$\begin{cases} \frac{\partial n}{\partial t}(s,t) + c\frac{\partial n}{\partial s}(s,t) = 0 & \text{for } s \in (0, s^*(t)) \\ n(s,0) = 0 \\ cn(0,t) = \tilde{\lambda}(z(t)), \end{cases} \qquad [7.4]$$

where $z(t)$ is the radius of the mother colony at time t and $s^*(t)$ is the radius of the first child colony coalescing with the mother colony immediately before the collision. Equation [7.4] has an explicit solution:

$$n(s,t) = \frac{1}{c}\tilde{\lambda}\Big(z\big(t - \frac{s}{c}\big)\Big)\mathbb{1}_{\{ct \geq s > 0\}}(s,t).$$

The area $\pi z(t)^2$ of the mother colony satisfies, before and after collision with a child colony:

$$\frac{d}{dt}\pi z^2 = \begin{cases} 2\pi zc & \text{for } t \in (0, t_1) \\ 2\pi zc + \pi {s^*}^2 n(s^*, t)(c - \frac{ds^*}{dt}) & \text{for } t \geq t_1, \end{cases}$$

where $t_1 = \frac{L}{2c}$ is the time when the first mother-child collision occurs.

Finally, $z(t)$ and $s^*(t)$ are linked by the following equation when $t \geq t_1$:

$$L = z(t) - z\Big(t - \frac{s^*(t)}{c}\Big) + s^*(t),$$

where $t - \frac{s^*(t)}{c}$ is the time when the collided child colony was at a distance L of the mother colony (for further details see Shigesada et al. [SHI 95]).

PDMP Formulation of the Coalescing Colony Model with Allee Effect:

The coalescing colony model can be seen as a precursory example of PDMPs modeling spatio-temporal population dynamics. In this case, the PDMP is the Boolean process formed by the union of the mother and child colonies:

$$X_t = \mathcal{B}(O, z(t)) \cup \Big(\bigcup_{i=1}^{m(t)} A_i(t)\Big)$$

$$A_i(t) = \begin{cases} \mathcal{B}(O_i, s_i(t)) & \text{if } d(O, O_i) > z(t) + s_i(t) \\ \emptyset & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(O, z(t))$ is the ball with center O and radius $z(t)$ covered by the mother colony, $m(t)$ is the number of child colonies generated until time t, and $\mathcal{B}(O_i, s_i(t))$ is the ball with center O_i and radius $s_i(t)$ covered by child colony i until its collision with the mother colony, that is to say while $z(t) + s_i(t) < d(O, O_i)$, and $d(\cdot, \cdot)$ is the inter-point Euclidean distance. Between collision times (thereafter called *jump times*), the radii of the mother and child colonies grow at the constant speed c given by Equation [7.3]. We remind, in addition, that the coalescence of two child colonies and the generation of grandchild colonies by child colonies (i.e. secondary colonizations) are neglected.

Let T_j be the j-th jump time corresponding to the time of generation of child colony j. Let τ_j be the time of collision between the mother colony and child colony j. Over $[T_j, T_{j+1})$, $m(t) = j$, eventual collisions following the expansion of colonies occur in a deterministic way and describing the dynamic of X_t is equivalent to describing the dynamics of the radii $z(t)$ and $s_i(t)$, $i = 1, \ldots, j$, because the centers O and O_i are fixed. For $t \in [0, T_1)$,

$$z(t) = ct$$

and for $t \in [T_j, T_{j+1})$, $j \geq 1$, the radii of the mother and child colonies satisfy:

$$\begin{aligned} z(t) =& z(T_j) + c(t - T_j) \\ &+ \sum_{i=1}^{j} \left[\left(s_i(\tau_i^-)^2 + z(\tau_i^-)^2 \right)^{1/2} - z(\tau_i^-) \right] \mathbb{1}(t \geq \tau_i > T_j) \\ s_i(t) =& \{ s_i(T_j) + c(t - T_j) \} \mathbb{1}(t < \tau_i), \quad \forall i = 1, \cdots, j, \end{aligned}$$

where

$$\begin{aligned} s_i(\tau_i^-) =& s_i(T_j) + c(\tau_i - T_j) \\ z(\tau_i^-) =& z(\max\{\tau_{i-1}, T_j\}) + c(\tau_i - \max\{\tau_{i-1}, T_j\}) \end{aligned}$$

with the conventions $\tau_0 = 0$ and $s_i(t) = 0$ when child colony i has merged with the mother colony. We now give the expression of τ_i for i such that $T_j < \tau_i < T_{j+1}$. Let $t_0 = \max\{\tau_{i-1}, T_j\}$ be the time of the event (i.e. a collision or the generation of a child colony) preceeding τ_i. If a collision occured at t_0 and if the resulting instantaneous growth of the mother colony led the mother colony to touch or overlap colony i, then $\tau_i = t_0$ (i.e. multiple instantaneous collisions occur). Otherwise, τ_i satisfies the following equation:

$$d(O, O_i) = L + z(T_i) = z(t_0) + c(\tau_i - t_0) + s_i(t_0) + c(\tau_i - t_0),$$

whose solution is:

$$\tau_i = t_0 + \frac{d(O, O_i) - z(t_0) - s_i(t_0)}{2c}. \qquad [7.5]$$

In the case of instantaneous collisions, the fraction in Equation [7.5] is non-positive (since the sum of radii $z(t_0) + s_i(t_0)$ is larger than or equal to $d(O, O_i)$). Thus, whatever the event at t_0,

$$\begin{cases} \tau_i = t_0 + \max\left\{0, \dfrac{d(O, O_i) - z(t_0) - s_i(t_0)}{2c}\right\} \\ t_0 = \max\{\tau_{i-1}, T_j\}. \end{cases}$$

Therefore, τ_i can be recursively defined as a function of radii and center locations at time T_j, which are functions of X_{T_j}

To demonstrate that X_t can be viewed as a PDMP, we will now give the expression of the *flow function* Φ, the *jump rate* λ and the *jump kernel* Q. Let

$$\mathbf{x} = \mathcal{B}(O_{\mathbf{x}}, z_{\mathbf{x}}) \cup \Big(\bigcup_{k=1}^{K_{\mathbf{x}}} \mathcal{B}(O_{\mathbf{x}k}, s_{\mathbf{x}k}) \Big)$$

be in the set $\mathcal{X}$ of unions of disjoint balls included in $\mathbb{R}^2$ and suppose that k is ordered such as the sequence of $d(O_{\mathbf{x}}, O_{\mathbf{x}k})$ increases with k. Note that knowing $\mathbf{x}$ is equivalent to knowing $\{O_{\mathbf{x}}, z_{\mathbf{x}}, O_{\mathbf{x}k}, s_{\mathbf{x}k}; k = 1, \ldots, K_{\mathbf{x}}\}$. Define Φ over $\mathcal{X} \times \mathbb{R}_+$ as follows:

$$\Phi(\mathbf{x}, t) = \mathcal{B}(O_{\mathbf{x}}, \phi_1(\mathbf{x}, t)) \cup \left(\bigcup_{k=1}^{K_{\mathbf{x}}} \mathcal{B}(O_{\mathbf{x}k}, \phi_2(\mathbf{x}, t, k)) \right),$$

with the convention $\mathcal{B}(O_{\mathbf{x}k}, 0) = \emptyset$ and

$$\begin{aligned}
\phi_1(\mathbf{x}, t) &= z_{\mathbf{x}} + ct \\
&\quad + \sum_{k=1}^{K_{\mathbf{x}}} \Bigg[\left((s_{\mathbf{x}k} + ct)^2 + (\phi_1(\mathbf{x}, \tau_{\mathbf{x},k-1}) + c(t - \tau_{\mathbf{x},k-1}))^2 \right)^{1/2} \\
&\qquad - (\phi_1(\mathbf{x}, \tau_{\mathbf{x},k-1}) + c(t - \tau_{\mathbf{x},k-1})) \Bigg] \mathbb{1}(t \geq \tau_{\mathbf{x}k}) \\
\phi_2(\mathbf{x}, t, k) &= (s_{\mathbf{x}k} + ct) \mathbb{1}(t < \tau_{\mathbf{x}k}), \quad \forall k = 1, \cdots, K_{\mathbf{x}} \\
\tau_{\mathbf{x}0} &= 0 \\
\tau_{\mathbf{x}k} &= \tau_{\mathbf{x},k-1} + \max \left\{ 0, \frac{d(O_{\mathbf{x}}, O_{\mathbf{x}k}) - \phi_1(\mathbf{x}, \tau_{\mathbf{x},k-1}) - \phi_2(\mathbf{x}, \tau_{\mathbf{x},k-1}, k)}{2c} \right\} \\
&\quad \forall k = 1, \ldots, K_{\mathbf{x}}.
\end{aligned}$$

Thus, X_t is a PDMP with flow function Φ:

$$X_t = \begin{cases} \Phi(X_{T_j}, t) & \text{if } t \in [T_j, T_{j+1}) \\ U_{j+1} & \text{if } t = T_{j+1}, \end{cases}$$

where the inter-jump duration $S_{j+1} = T_{j+1} - T_j$ (with $j \geq 1$ and the convention $T_0 = 0$) has a survival function satisfying:

$$P(S_{j+1} \geq t) = \exp\left(- \int_0^t \lambda(\Phi(X_{T_j}, v)) dv \right);$$

the rate function $\lambda : \mathcal{X} \to \mathbb{R}_+$ satisfies:

$$\lambda(\mathbf{x}) = \tilde{\lambda}(z_{\mathbf{x}}),$$

with $\tilde{\lambda}(z_{\mathbf{x}}) = \lambda_1 z_{\mathbf{x}}$ for example as proposed in Section 7.2.3; and U_{j+1} is drawn from the jump kernel $Q(\Phi(X_{T_j}, S_{j+1}), \cdot)$ such that:

$$U_{j+1} = \Phi(X_{T_j}, S_{j+1}) \cup \mathcal{B}(O_{\text{new}}, 0)$$

with O_{new} uniformly drawn on the circle centered around O and radius $z(T_{j+1}) + L$.

7.2.4. *A PDMP Based on Reaction-Diffusion for Modeling Invasions with Multiple Introductions*

Section 7.2.1 presented the use of reaction-diffusion equations for modeling population dynamics with short-distance dispersal and Section 7.2.2 presented the combination of a jumping process and an approximation of a reaction-diffusion equation to obtain a model with both short and long-distance dispersal. The latter model was shown to be a PDMP. Here, we introduce another spatio-temporal PDMP based on reaction-diffusion for modeling dynamics with short-distance dispersal only but with multiple introductions of the species of interest. In this model, the flow represented by a reaction-diffusion equation with an Allee effect will be stochastically disrupted at random times to mimic introductions having a limited extent in space. This model will be used in a future study to describe the dynamics of the plant-pathogenic bacterium *Xylella fastidiosa* (Xf) in Corsica. Figure 7.5 shows the pattern of plants which have been detected as infected by Xf in Corsica between August 2015 and May 2017. This map displays several clusters of infected plants with different sizes, which may have been induced by several introductions of the pathogen in different areas of Corsica and at different times.

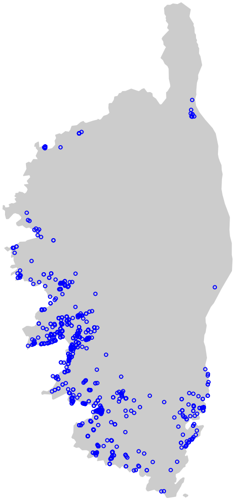


Figure 7.5: Pattern of plants which have been detected as infected by *Xylella fastidiosa* in Corsica between August 2015 and May 2017.

In what follows, we introduce a candidate model for describing the invasion of Corsica by Xf and lay some track to estimate the unknown parameters and latent variables of the model. Assume that $u(t, \mathbf{x})$, which will be used to model the probability that a plant located at $\mathbf{x} \in \Omega \subset \mathbb{R}^2$ is infected at time t, satisfies between two introductions of the invading species:

$$\begin{cases} \frac{\partial u}{\partial t} = D\Delta u + bu(u-\theta)(1-u) & \text{in } \Omega \\ \nabla u.n = 0 & \text{on } \partial\Omega, \end{cases} \qquad [7.6]$$

where D is the dispersal rate, b the intrinsic growth rate of Xf, and $\theta \in]0; \frac{1}{2}[$ the reaction threshold in Ω which induces an Allee effect (Ω, in the Xf application, will be the area covered by the Corsican territory).

The progression of u will be interrupted at each introduction time and *re-initialized*. At the first introduction time, i.e. $t = \tau_0$, u is initialized as follows:

$$u_0(\mathbf{x}) = u(\tau_0, \mathbf{x}) = f(\mathbf{x}, \mathbf{x}_0) \quad \text{in } \Omega,$$

where $f : \Omega \mapsto [0,1]$ is a continuous function, which is typically decreasing with the distance from $\mathbf{x}_0$ to $\mathbf{x}$ (like a kernel function). Thus, the invading species is first introduced around $\mathbf{x}_0$ at τ_0.

The subsequent introductions (i.e. the jumps) are assumed to be governed by a spatio-temporal homogeneous Poisson point process Ψ with constant intensity λ over $\Omega \times (\tau_0, \tau_{\text{end}})$. Let $\{\psi_0^1, \cdots, \psi_0^N\}$ be a realization of Ψ where $\psi_0^i = (\mathbf{x}_0^i, T_i)$, and set $(\mathbf{x}_0^0, T_0) = (\mathbf{x}_0, \tau_0)$ and $T_{N+1} = \tau_{\text{end}}$. We define the spatio-temporal PDMP $\{X_t\}_{\tau_0 \le t < \tau_{\text{end}}}$ by:

$$X_t(\mathbf{x}) = \begin{cases} f(\mathbf{x}, \mathbf{x}_0^0) & \text{if } t = T_0 = \tau_0 \\ u(t, \mathbf{x}) & \text{if } t \in (T_i, T_{i+1}),\ i = 0, \dots, N \\ u(t, \mathbf{x}) + f(\mathbf{x}, \mathbf{x}_0^i) & \text{if } t = T_i,\ i = 1, \dots, N \end{cases}$$

where u is governed by Equation [7.6] over $(T_i, T_{i+1}]$ with initial state at T_i being X_{T_i}. Then, $\min\{1, \max\{0, X_t(\mathbf{x})\}\}$ is viewed as the probability that a plant located at $\mathbf{x} \in \Omega$ is infected at time t. The $\min - \max$ operator is used because X_t may sporadically go out of $[0,1]$.

In the application of interest, namely the invading dynamic of Xf in Corsica, the estimation of model parameters (D, b, θ, λ and eventual parameters arising in f) and latent variables (jump times T_i and introduction locations $\mathbf{x}_0^i$) will be carried out in a mechanistic-statistical framework, which can cope with various types of data [ROQ 11, SOU 09a, SOU 09b, WIK 03a, WIK 03b]. Consider, for instance, that data collection consists of independently sampling plants in $\Omega \times (\tau_0, \tau_{\text{end}})$ and diagnosing their health statuses. Let $Z(\mathbf{s}_j, t_j) \in \{0,1\}$ be the observed health status of plant j sampled at location $\mathbf{s}_j$ and time t_j, $j = 1, \dots, n$, where 0 stands for the *observed healthy status* and 1 for the *observed infected status*. Let ϵ_{FN} be the probability of diagnosing a plant as healthy whereas it is infected (false-negative rate) and ϵ_{FP} be the probability of diagnosing a plant as infected whereas it is healthy (false-postive rate). Then, $Z(\mathbf{s}_j, t_j)$ can be assumed to be Bernoulli distributed as follows:[1]

$$Z(\mathbf{s}_j, t_j) \mid \{X_t\} \overset{\text{indep.}}{\sim} \text{Bernoulli}\Big(\epsilon_{\text{FN}} \min\{1, \max\{0, X_{t_j}(\mathbf{s}_j)\}\} \\ + \epsilon_{\text{FP}}(1 - \min\{1, \max\{0, X_{t_j}(\mathbf{s}_j)\}\})\Big),$$

and the estimation of model parameters and latent variables can be made, in a frequentist or Bayesian framework, with the resulting likelihood and an appropriate algorithm (an example of Bayesian algorithm will be given in the next section for a different model).

7.3. Metapopulation Epidemic Model

See Appendix C

7.4. Stochastic Approaches for Modeling Spatial Trajectories

See Appendix C

7.5. Conclusion

This chapter gave an introduction to spatio-temporal PDMPs used to model population dynamics. Spatio-temporal PDMPs offer the possibility to build flexible models and achieve relatively realistic and consistent inferences. Thus, we presented three different modeling frameworks corresponding to three resolutions, namely the population, the metapopulation and the individual. We have seen that, depending on the dynamics of interest, the jumps in the PDMP can correspond to long-distance dispersal events, new introductions, or significant shifts in individual behaviors.

In the examples of models presented above, the spatio-temporal dependencies are contained in the flow function, whereas jumps are independent and identically distributed. However, for populations whose individuals can be transported in groups [SOU 11, SOU 14a], jumps should be correlated in space and time. For instance, in the metapopulation model of Section 7.3, a source patch could release a group of spores transported by wind towards a set of nearby patches. Such a process could lead to the simultaneous infection of several aggregated patches. Hence, developing

[1] With respect to its initial version, this Equation has been corrected in this manuscript.

PDMPs with dependent random jumps would be interesting for better taking into account specificities of some population dynamics. Moreover, it would be also challenging from the perspective of model construction, simulation and inference.

Acknowledgements

This work was partially funded by the HORIZON 2020 XF-ACTORS Project SFS-09-2016 and the INRA-DGAL Xylella Project n° 21000679.

7.6. Bibliography

[ALF 13] ALFARO M., COVILLE J., RAOUL G., "Travelling waves in a nonlocal reaction-diffusion equation as a model for a population structured by a space variable and a phenotypic trait", *Communications in Partial Differential Equations*, vol. 38, p. 2126–2154, Taylor & Francis, 2013.

[BOR 17] BORDIER C., DECHATRE H., SUCHAIL S., PERUZZI M., SOUBEYRAND S., PIOZ M., PÉLISSIER M., CRAUSER D., LE CONTE Y., ALAUX C., "Colony adaptive response to simulated heat waves and consequences at the individual level in honeybees (*Apis mellifera*)", *Scientific Reports*, vol. 7, Nature Publishing Group, 2017.

[CAI 17] CAILLERIE N., Stochastic and deterministic kinetic equations in the context of mathematics applied to biology, Theses, Université de Lyon, July 2017.

[CHA 06] CHAMPAGNAT N., FERRIÈRE R., MÉLÉARD S., "Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models", *Theoretical population biology*, vol. 69, p. 297–321, Elsevier, 2006.

[COL 04] COLAUTTI R. I., RICCIARDI A., GRIGOROVICH I. A., MACISAAC H. J., "Is invasion success explained by the enemy release hypothesis?", *Ecology Letters*, vol. 7, p. 721–733, 2004.

[EVA 98] EVANS L. C., *Partial differential equations*, Graduate studies in mathematics, American Mathematical Society, Providence, R.I, 1998.

[FIS 37] FISHER R. A., "The wave of advance of advantageous genes", *Annals of Eugenics*, vol. 7, p. 355–369, 1937.

[GAT 96] GATENBY R. A., GAWLINSKI E. T., "A reaction-diffusion model of cancer invasion", *Cancer research*, vol. 56, p. 5745–5753, 1996.

[GLO 15] GLOAGUEN P., ETIENNE M.-P., LE CORFF S., Stochastic differential equation based on a multimodal potential to model movement data in ecology, working paper or preprint, September 2015.

[HAD 75] HADELER K., ROTHE F., "Travelling fronts in nonlinear diffusion equations", *Journal of Mathematical Biology*, vol. 2, p. 251–263, Springer, 1975.

[HEN 89] HENGEVELD R., *Dynamics of biological invasions*, Springer Science & Business Media, 1989.

[KOL 37] KOLMOGOROV I., PETROVSKY A., PISCOUNO N., "Etude de l'équation de la diffusion avec croissance de la quantité de la matière et son application à un problème biologique", *Moscow University Bull. Math*, vol. 1, p. 1–25, 1937.

[LEW 93] LEWIS M., KAREIVA P., "Allee dynamics and the spread of invading organisms", *Theoretical Population Biology*, vol. 43, p. 141–158, Elsevier, 1993.

[LUT 06] LUTHER R., "Räumliche Ausbreitung chemischer Reaktionen", *Zeitschrift für Elektrochemie*, vol. 12, p. 596–600, 1906.

[MÉL 11] MÉLÉARD S., "*Random Modeling of Adaptive Dynamics and Evolutionary Branching*", p. 175–192, Springer Basel, Basel, 2011.

[MUR 96] MURRAY J., KULESA P., "On a dynamic reaction–diffusion mechanism: The spatial patterning of teeth primordia in the alligator", *Journal of the Chemical Society, Faraday Transactions*, vol. 92, p. 2927–2932, The Royal Society of Chemistry, 1996.

[RIC 00] RICHARDSON D. M., PYŠEK P., REJMANEK M., BARBOUR M. G., PANETTA F. D., WEST C. J., "Naturalization and invasion of alien plants: concepts and definitions", *Diversity and distributions*, vol. 6, p. 93–107, 2000.

[ROQ 10] ROQUES L., HAMEL F., FAYARD J., FADY B., KLEIN E., "Recolonisation by diffusion can generate increasing rates of spread", *Theoretical population biology*, vol. 77, p. 205–212, Elsevier, 2010.

[ROQ 11] ROQUES L., SOUBEYRAND S., ROUSSELET J., "A statistical-reaction–diffusion approach for analyzing expansion processes", *Journal of Theoretical Biology*, vol. 274, p. 43–51, Elsevier, 2011.

[ROQ 13] ROQUES L., *Modèles de réaction-diffusion pour l'écologie spatiale: [avec exercices dirigés]*, Éd. Quae, Versailles, 2013, OCLC: 951442261.

[ROT 81] ROTHE F., "Convergence to pushed fronts", *The Rocky Mountain Journal of Mathematics*, vol. 11, p. 617–633, JSTOR, 1981.

[SHI 95] SHIGESADA N., KAWASAKI K., TAKEDA Y., "Modeling stratified diffusion in biological invasions", *The American Naturalist*, vol. 146, p. 229–251, University of Chicago Press, 1995.

[SHI 97] SHIGESADA N., KAWASAKI K., *Biological invasions: theory and practice*, Oxford University Press, UK, 1997.

[SKE 51] SKELLAM J. G., "Random dispersal in theoretical populations", *Biometrika*, vol. 38, p. 196–218, 1951.

[SOU 09a] SOUBEYRAND S., LAINE A., HANSKI I., PENTTINEN A., "Spatio-temporal structure of host-pathogen interactions in a metapopulation", *The American Naturalist*, vol. 174, p. 308-320, 2009.

[SOU 09b] SOUBEYRAND S., NEUVONEN S., PENTTINEN A., "Mechanical-statistical modelling in ecology: from outbreak detections to pest dynamics", *Bulletin of Mathematical Biology*, vol. 71, p. 318-338, 2009.

[SOU 11] SOUBEYRAND S., ROQUES L., COVILLE J., FAYARD J., "Patchy patterns due to group dispersal", *Journal of Theoretical Biology*, vol. 271, p. 87-99, 2011.

[SOU 14a] SOUBEYRAND S., MRKVIČKA T., PENTTINEN A., “A Nonstationary Cylinder–Based Model Describing Group Dispersal in a Fragmented Habitat”, *Stochastic Models*, vol. 30, p. 48–67, Taylor & Francis, 2014.

[SOU 14b] SOUBEYRAND S., ROQUES L., “Parameter estimation for reaction-diffusion models of biological invasions”, *Population ecology*, vol. 56, p. 427–434, Springer, 2014.

[TUR 03] TURCHIN P., *Complex population dynamics: a theoretical/empirical synthesis*, vol. 35, Princeton University Press, 2003.

[WIK 03a] WIKLE C. K., “Hierarchical Bayesian models for predicting the spread of ecological processes”, *Ecology*, vol. 84, p. 1382–1394, Wiley Online Library, 2003.

[WIK 03b] WIKLE C. K., “Hierarchical models in environmental science”, *International Statistical Review*, vol. 71, p. 181–199, Wiley Online Library, 2003.

5.3. Unravelling Multiple Introductions of an Invasive Species

This section is a first step towards inferring a spatio-temporal PDMP designed as the example introduced at the end of Section 5.2. Here, we do not infer a spatio-temporal PDMP as such, but we aim to explore the possible difficulties that we may encounter in future working projects in this direction. Thus, we infer model parameters using post-introduction data of Xf in Corsica (and not only South Corsica as in previous chapters), supposing that only one jump has occurred (i.e., two introductions of the invasive species). We also simplify the problem and its implementation by using one of the models introduced in Chapters 3 and 4, and by fixing the temperature threshold included in the model. At the end of this section we discuss the extension of the approach proposed below to infer the PDMP as a whole and alternative approaches.

5.3.1. Methods

We are interested in the invasion of a pathogen with multiple introductions in a domain Ω included in $\mathbb{R}^2$. This domain is partitioned into two sub-domains to account for spatial heterogeneities in the reproduction regimes of the pathogen. The sub-domains Ω_1 and Ω_2 are defined by thresholding a spatial function, say T, with the threshold value $\tilde{T}$ that is hold fixed, such that: $\Omega = \Omega_1 \cup \Omega_2$; $\Omega_1 = \Omega_1(T, \tilde{T}) = \{\mathbf{x} \in \Omega : T(\mathbf{x}) > \tilde{T}\}$; and $\Omega_2 = \Omega_2(T, \tilde{T}) = \{\mathbf{x} \in \Omega : T(\mathbf{x}) \leq \tilde{T}\}$.

Let us re-call a model used in Chapters 3 and 4, describing the flow $u(t, \mathbf{x})$ between the introductions, where $u(t, \mathbf{x})$ is the probability of a plant located at $\mathbf{x} \in \Omega$ to be infected at time t. The probability $u(t, \mathbf{x})$ satisfies:

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\Delta u + bu\left(1 - \dfrac{u}{K}\right)\mathbb{1}(\mathbf{x} \in \Omega_1) - \alpha u \mathbb{1}(\mathbf{x} \in \Omega_2), & t \geq \tau_0,\ \mathbf{x} \in \Omega, \\ \nabla u(t, \mathbf{x}).n(\mathbf{x}) = 0, & t \geq \tau_0,\ \mathbf{x} \in \partial\Omega, \end{cases} \tag{5.1}$$

where $D > 0$ is the diffusion coefficient; b corresponds to the intrinsic growth rate of the pathogen infection in Ω_1; $K \in (0, 1]$ is a plateau for the probability of infection (i.e., an analogue to the carrying capacity of the environment); α is the decrease rate of the infection in Ω_2; $\Delta = \dfrac{\partial^2}{\partial x_1^2} + \dfrac{\partial^2}{\partial x_2^2}$ is the 2-dimensional diffusion operator of Laplace; $\nabla = \dfrac{\partial}{\partial x_1} + \dfrac{\partial}{\partial x_2}$ is the 2-dimensional gradient operator; $\mathbf{x} \mapsto \mathbb{1}(\mathbf{x} \in \Omega_i)$ is the characteristic function taking the value 1 if $\mathbf{x} \in \Omega_i$ and 0 otherwise; $\tau_0 \in \mathbb{R}$ is the first introduction time of the pathogen.

Homogeneous boundary conditions are considered on the boundary $\partial\Omega$ of Ω, i.e., with reflection on the boundary (second line of Equation (5.1)). Physically, this signifies that there is neither outward nor inward flux from and to Ω.

The progression of u will be interrupted at each introduction time and *re-initialized* conditionally on the state of u right before the introduction time. At the first introduction time τ_0, u is initialized as follows:

$$u_0(\mathbf{x}) = u(\tau_0, \mathbf{x}) = f(\mathbf{x}, \mathbf{x}_i), \tag{5.2}$$

where f is made explicit in Equation (5.4) below.

The subsequent introductions (i.e. the jumps) are assumed to be governed by a spatio-temporal homogeneous Poisson point process Ψ with constant intensity λ over $\Omega \times (\tau_0, \tau_{\text{end}})$. Let $\{\psi_1, \cdots, \psi_N\}$, $N \in \mathbb{N}^*$, be a non-empty realization of Ψ where $\psi_i = (\mathbf{x}_i, \tau_i)$, $i \in \{1, \cdots, N\}$. In addition, let $\tau_{N+1} = \tau_{\text{end}}$. We define the spatio-temporal PDMP $\{X_t\}_{\tau_0 \leq t < \tau_{\text{end}}}$ by:

$$X_t(\mathbf{x}) = \begin{cases} f(\mathbf{x}, \mathbf{x}_0) & \text{if } t = \tau_0, \\ u(t, \mathbf{x}) & \text{if } t \in (\tau_i, \tau_{i+1}),\ i = 0, \dots, N, \\ u(t, \mathbf{x}) + f(\mathbf{x}, \mathbf{x}_i) & \text{if } t = \tau_i,\ i = 1, \dots, N, \end{cases} \tag{5.3}$$

where u is governed by Equation (5.1) over $(\tau_i, \tau_{i+1}]$ with initial state at τ_i being X_{τ_i}. Then, $\min\{1, \max\{0, X_t(\mathbf{x})\}\}$ is viewed as the probability that a plant located at $\mathbf{x} \in \Omega$ is infected at time t. The $\min - \max$ operator is used because X_t may sporadically go out of $[0, 1]$ at jumping times depending on the amplitude of the jumps. Following the way we modeled the introduction in Chapters 3 and 4, f satisfies:

$$f(\mathbf{x}, \mathbf{x}_i) = p_0 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right), \tag{5.4}$$

where p_0 is the *supplementary* infection probability at $(\tau_i, \mathbf{x}_i)$, $\forall i = 0, \cdots, N$; $\|.\|$ is the Euclidean norm; $\sigma^2 = \frac{r_0^2}{q}$, q is the 0.95-quantile of the χ^2 distribution with two degrees of freedom; and r_0 is the radius of the kernel. Thus, at τ_i, if we neglect border effects, 95% of the infected *supplementary* plants are located within the ball with center $\mathbf{x}_i$ and radius r_0.

Remark: To avoid the $\min - \max$ operator, the supplementary infection probability could apply only on the susceptible fraction of the local population. Thus, in Equation (5.3), at $t = \tau_i$, $X_t(\mathbf{x})$ would be equal to $u(t, \mathbf{x}) + (1 - u(t, \mathbf{x}))f(\mathbf{x}, \mathbf{x}_i)$.

In the application of interest, namely the invading dynamics of Xf in Corsica, we dispose of spatio-temporal post-introduction data collected between July 2015 and April 2019. Over this period, approximately 17000 plants were sampled, among which 1000 have been diagnosed as infected using real-time *polymerase chain reaction* (PCR) technique [Denancé et al., 2017b]. Figure 5.1 shows coordinates and health statuses of sampled plants in the island. Moreover, we build the average of the daily minimum temperature to divide the spatial domain into two-subdomains. This average is provided by a freely available temperature dataset [Huld et al., 2006], and is used in Chapters 3 and 4. Here, the threshold of temperature is fixed to $\tilde{T} = 5$ (in Celsius degrees). Model partitioning for this threshold is shown in Figure 5.1.

Due to lack of time, the estimation of model parameters is performed for exploration purposes. For this reason, we mainly focus on inferring parameters of the model embedding only two introductions (i.e., $N = 1$). Precisely, we aim to infer the vector of parameters $\Theta = [D,\ b,\ K,\ \alpha,\ \mathbf{x}_0,\ \tau_0,\ \mathbf{x}_1,\ \tau_1,\ r_0,\ p_0]$. Any other component arising in f is hold fixed. This estimation is carried out in the mechanistic-statistical framework. Consider that data collection consists of independently sampling plants in $\Omega \times (\tau_0, \tau_{\text{end}})$ and diagnosing their health statuses. Let $Y(t_j, \mathbf{s}_j) \in \{0, 1\}$ be the observed health status of plant j sampled at location $\mathbf{s}_j$ and time t_j, $j = 1, \dots, n \in \mathbb{N}*$, where 0 stands for the *o*bserved healthy status and 1 for the *o*bserved infected status. Then, $Y(t_j, \mathbf{s}_j)$ is assumed to be Bernoulli distributed as follows:

$$Y(t_j, \mathbf{s}_j) \mid \{X_t\} \overset{\text{indep.}}{\sim} \text{Bernoulli}\Big(\min\{1, \max\{0, X_{t_j}(\mathbf{s}_j)\}\}\Big). \tag{5.5}$$

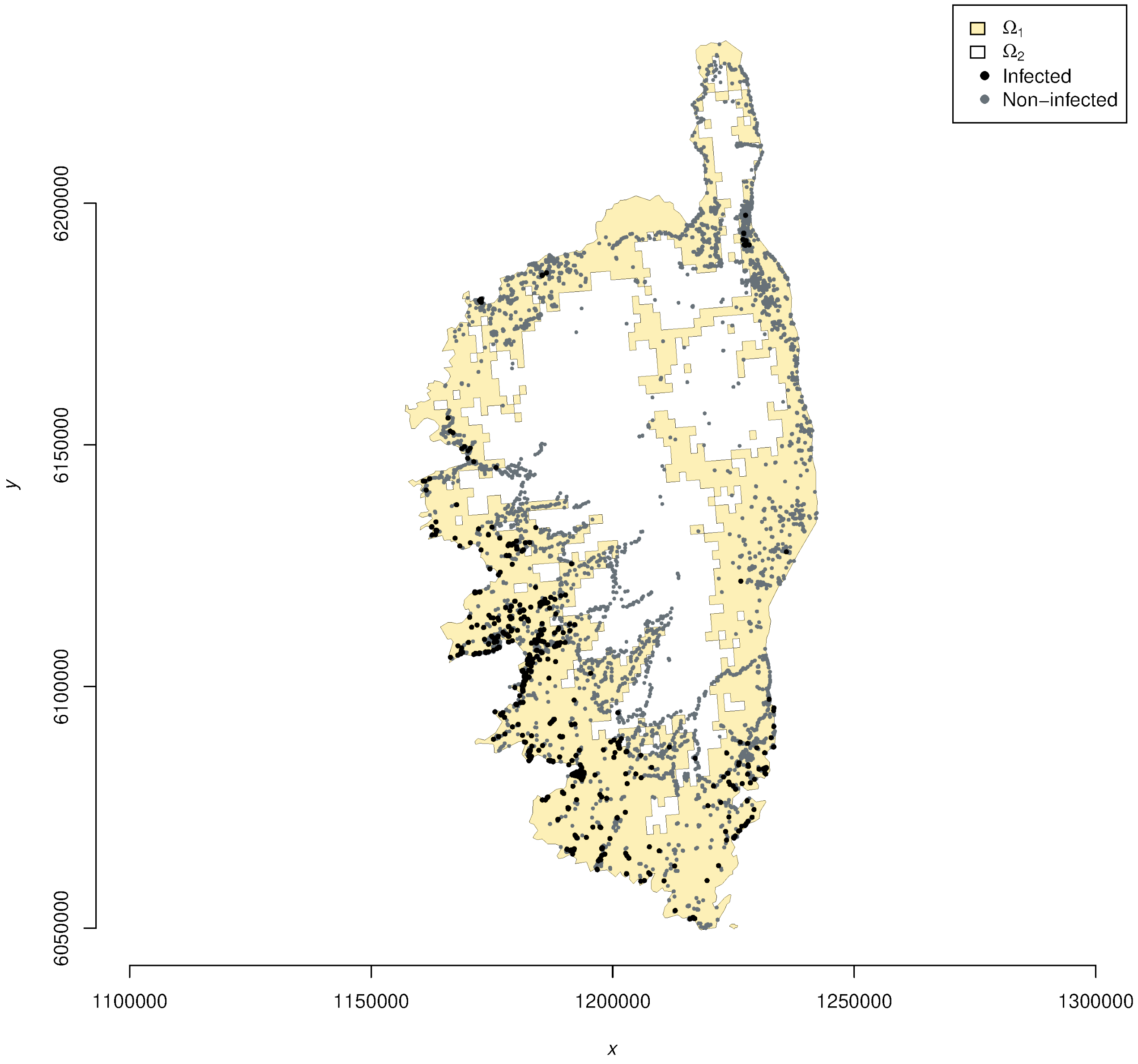


Figure 5.1.: Locations of plants sampled from July 2015 to April 2019, that have been detected as positive (black dots) or negative (grey dots) to Xf in Corsica, France, and partition of the study domain Ω into the sub-domains Ω_1 and Ω_2 for $\tilde{T} = 5°\text{C}$.

The estimation of Θ is made in the Bayesian framework introduced in Chapter 3 by specifying the likelihood function and the prior distribution, and by implementing the AMIS algorithm. The likelihood satisfies, using Equation (5.5):

$$[Y|\Theta] = \prod_{j=1}^{n} u(t_j, \mathbf{s}_j)^{Y(t_j,\mathbf{s}_j)}(1 - u(t_j, \mathbf{s}_j))^{1-Y(\mathbf{s}_j,t_j)}. \tag{5.6}$$

The prior distribution for Θ, which combines vague and uniform distributions and Dirac distributions, satisfies:

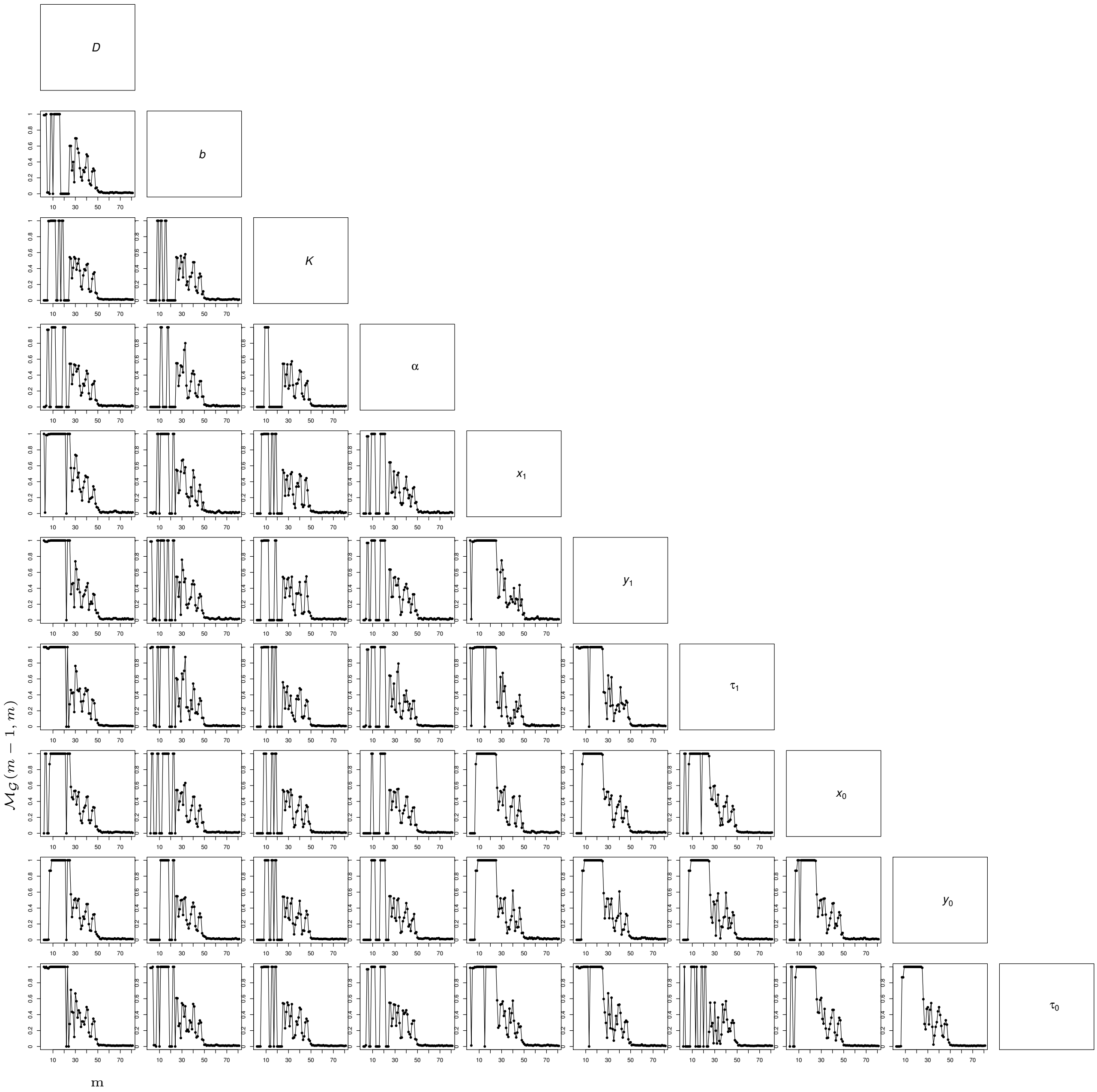


Figure 5.2.: Variation in the deviation measure $\mathcal{M}_{\mathcal{G}}(m-1,m)$ between the assessments of the posterior distribution at iteration $m-1$ and $m>1$ of the AMIS algorithm. $\mathcal{M}_{\mathcal{G}}(m-1,m)$ is plotted for different partitions $\mathcal{G}$ allowing the assessment of the stabilization of all the 2D posterior distributions of parameters D, b, K, α, x_0, y_0, τ_0, x_1, y_1 and τ_1.

$$
\begin{aligned}
[\Theta] = & \frac{1}{(10^8-50)\times 100\times 1\times 100\times 1000\times|\Omega_1|} \\
& \times \mathbb{1}(D\in[50;10^8], b\in[0;100], K\in]0;1], \alpha\in[0;100]) \\
& \times \mathbb{1}(\tau_0\in[-1000;0], \tau_1\in[-1000;40], \mathbf{x}_0\in\Omega_1, \mathbf{x}_1\in\Omega_1) \\
& \times \mathsf{Dirac}_{5000}(r_0)\times\mathsf{Dirac}_{0.1}(p_0),
\end{aligned}
$$

where $|\Omega_1|$ is the area of Ω_1 and $\text{Dirac}_b(B)$ is equal to 1 if $B = b$, and 0 otherwise. The Dirac distribution for $\tilde{T}$ is chosen to deal with implementation issues explained in Chapter 3. We choose Dirac prior distributions for r_0 and p_0 in the aim of precisely defining what is an introduction and to avoid any identifiability issues. For D, b, K and α, we specify vague uniform priors satisfying constraints of positivity. In addition, the plateau K has to be less than 1, as indicated in Section 5.3.1. For the first introduction time τ_0, we choose a uniform distribution between -1000 months and 0 month before the first detection of Xf in Corsica. For the second introduction time τ_1, we choose a uniform distribution between -1000 months and 40 months (i.e., few months before the last observation). Note that, using a temporal model and aggregated data, Soubeyrand et al. [2018] inferred an introduction date around -360 months before the first detection of Xf in South Corsica. Finally, the introduction locations $\mathbf{x}_0$ and $\mathbf{x}_1$ were supposed to be uniformly distributed in Ω_1, the sub-domain where the conditions are favorable for the expansion of Xf.
The AMIS algorithm (see Section 2.2.2.2.5) iteratively generates parameter vectors under the adaptive multinormal proposal distribution $\mathcal{N}(\mu_{m-1}, \Sigma_{m-1})$, $m = 1, \cdots, M$, where M is the number of algorithm iterations. μ_{m-1} and Σ_{m-1} are the mean vector and the covariance matrix of this distribution at iteration m, also called tuning parameters.

5.3.2. Preliminary Results

5.3.2.1. Stabilization of the AMIS Algorithm and Exploration of the Search Space

We assess the stabilization of the AMIS as proposed in Chapter 3. Thus, we evaluate the variation in the following deviation measure between the assessments of the posterior distribution at iteration $m - 1$ and $m > 1$:

$$\mathcal{M}_{\mathcal{G}}(m-1, m) = \max_{c \in \mathcal{G}} |p_m(c) - p_{m-1}(c)|,$$

where $p_m(c)$ denotes the assessment at iteration m of the posterior probability that Θ is in the sub-domain $c \subset \mathbb{R}^{10}$ of the parameter space, i.e.

$$p_m(c) = \sum_{m'=1}^{m} \sum_{l=1}^{L} w_{m'}^l \mathbb{1}(\Theta_{m'}^l \in c),$$

$\mathcal{G}$ is a partition of a sub-space of the parameter space and $\{(\Theta_{m'}^l, w_{m'}^l) : 1 \leq m' \leq m, 1 \leq l \leq L\}$ is the weighted posterior sample provided by AMIS at iteration $m \in \{1, \cdots, M\}$.

Figure 5.2 gives the variation in $\mathcal{M}_{\mathcal{G}}(m-1, m)$ for different partitions $\mathcal{G}$ allowing us to assess the stabilization of all the 2D posterior distributions of parameters). For each pair of parameters, $\mathcal{G}$ was defined as the set of infinite cylinders with rectangular bases whose orthogonal projection in the 2 dimensions of interest forms a 20×20 regular rectangular grid. In each dimension of interest, the endpoints of the grid were set at the minimum and maximum values of the corresponding parameter having a weight w_M^l larger than 10^{-5} (the 2D posterior distributions over these 20×20 grids are displayed in Figure 5.6). Figure 5.2 shows the stabilization of all the 2D posterior distributions after iteration 50.

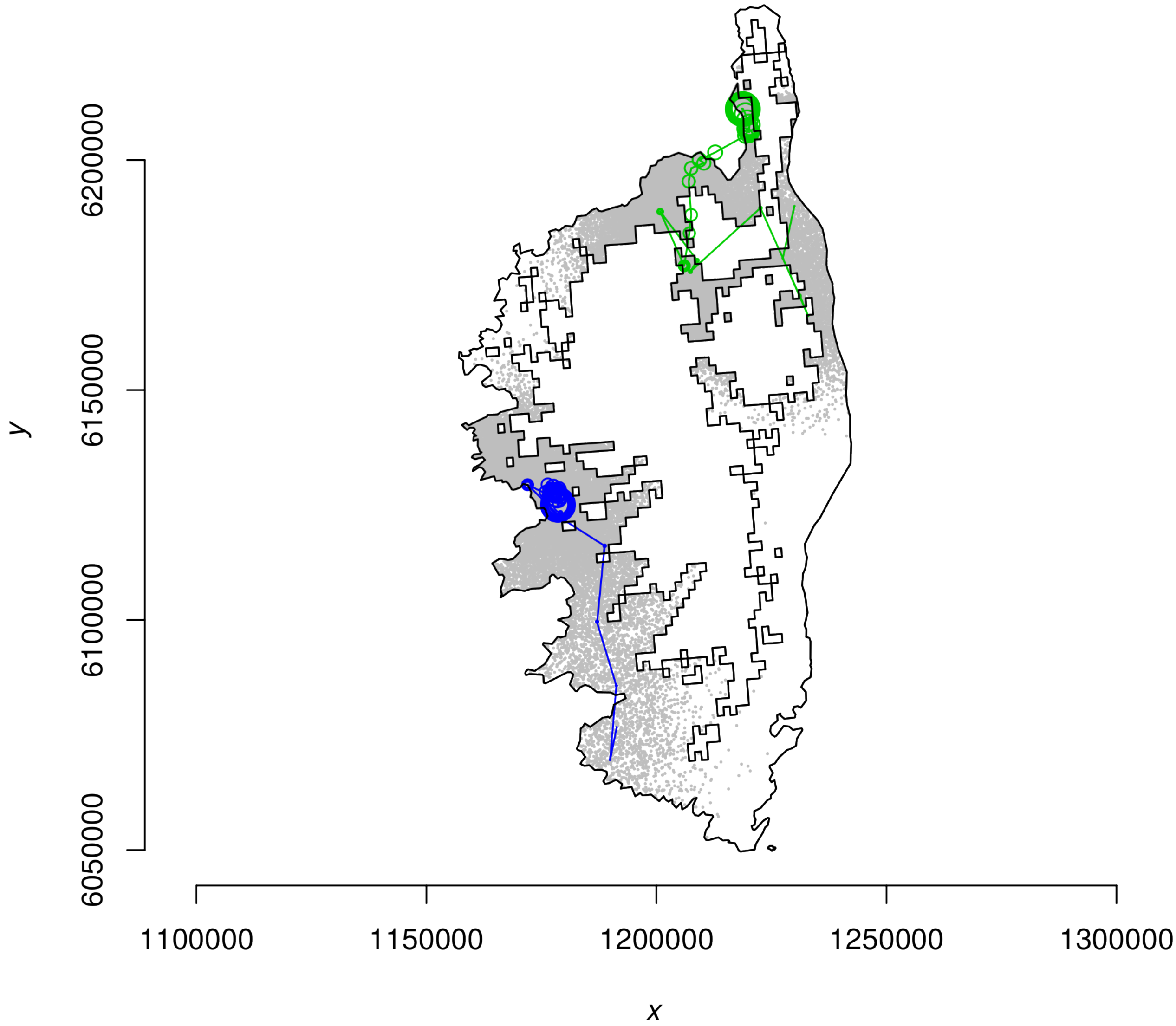


Figure 5.3.: Search space of the parameters $\mathbf{x}_0$ and $\mathbf{x}_1$ showing all the parameters generated from the proposal distributions since the first AMIS iteration (grey dots), and trajectory across AMIS iterations of the components of μ_m (the mean vector of the proposal distribution) corresponding to $\mathbf{x}_0$ (in green) and $\mathbf{x}_1$ (in blue). The larger the circle the higher the iteration.

Figure 5.3 shows, for the introduction locations, the search space exploited by the proposal distributions across AMIS iterations. If the entire domain Ω_1 has not been explored, a large fraction of it did. Unexplored regions could be visited by proposing initial proposal distributions with larger supports with respect to $\mathbf{x}_0$ and $\mathbf{x}_1$. However, unexplored regions may be unlikely regions for the introductions, given the data and the constraints incorporated in the model and the prior, and therefore legitimately unexplored. Figure 5.3 also shows that AMIS rapidly identified an introduction in the West of Corsica corresponding to the introduction identified in Chapters 3 and 4. AMIS was more "hesitating" for the second introduction in the North where Ω_1 is more fragmented and the disease prevalence is lower.

5.3.2.2. Posterior Distributions of the Parameters

Table 5.1.: Posterior medians, means and standard deviations of parameters of the reaction-diffusion-absorption equation obtained in Chapter 3 and in the current study.

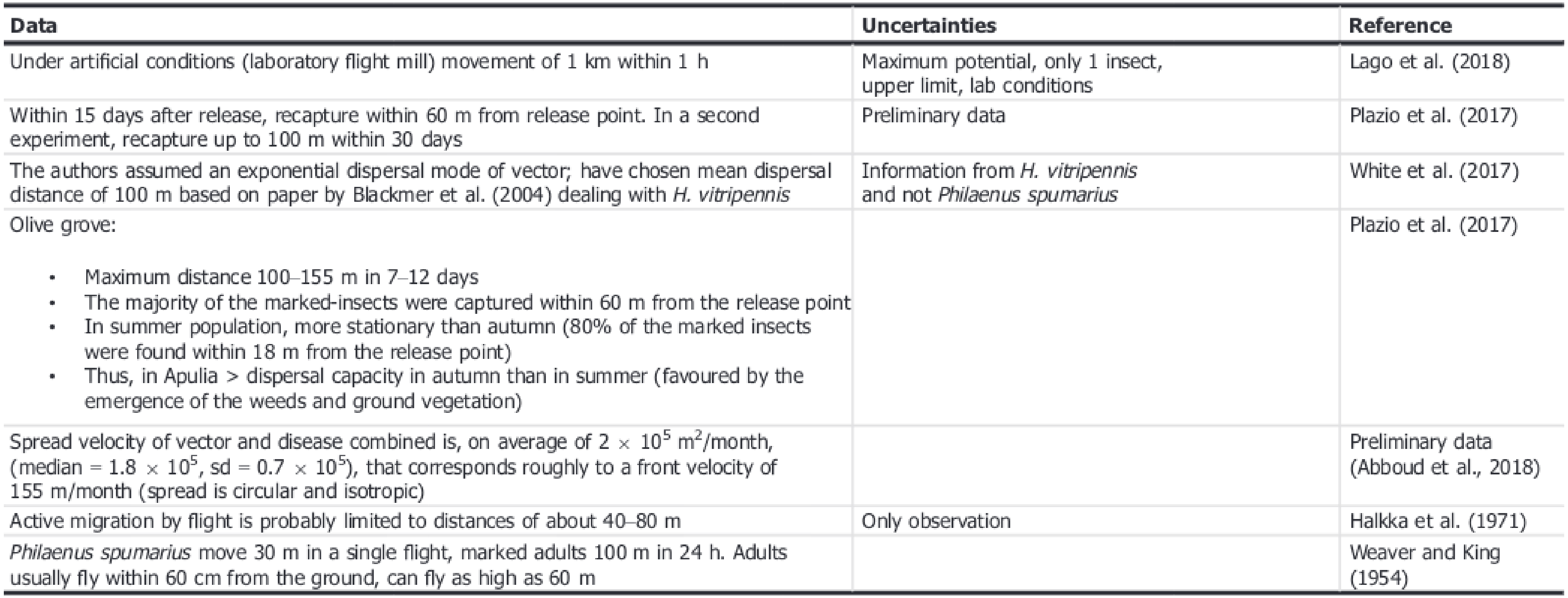
Parameter	Unit	Study	Median	Mean	Standard deviation
D	$m^2 \cdot$ month^{-1}	Current	2.4×10^5	2.4×10^5	0.5×10^5
		Chapter 3	1.8×10^5	2.0×10^5	0.7×10^5
b	month^{-1}	Current	0.019	0.019	0.003
		Chapter 3	0.026	0.027	0.008
K	probability	Current	0.13	0.13	0.01
		Chapter 3	0.15	0.15	0.01
α	month^{-1}	Current	0.07	0.07	0.02
		Chapter 3	0.12	0.13	0.05

Table 5.1 compares posterior statistics of D, b, K and α obtained for this application and for the analogue model with one introduction proposed and fitted to data from South Corsica only in Chapter 3. Standard deviations of parameters tend to be lower in the current study than in Chapter 3. This may be due to the additional data that are considered and that bring additional information. Another possibility is that the parameter space was not fully explored by the inference algorithm, which might got stuck in a local maximum. We observe that the plateau for the probability of infection is around 12%. This value is lower than the one obtained in Chapter 3, and this may be due to the dependence of this parameter from the population of plants that have been sampled, and which is not equal in both studies. Concerning the diffusion parameter D (directly estimated by AMIS; see Table 5.1), as in Chapter 3, the inference allowed us to assess the length of a straight line move of the pathogen during a time unit, namely the month. This length is given by Equation (5.7) [Turchin, 1998, Roques et al., 2016] that we re-call hereafter:

$$D = \frac{(\text{length of a straight line move during one time step})^2}{4 \times \text{duration of the time step}}, \tag{5.7}$$

and has a posterior median equal to 176 meters (posterior mean: 178m; posterior standard deviation: 19m). These figures correspond to the move of the pathogen with different means, in particular via insects and transportation of infected plants, which are both modeled by the diffusion operator in Equation (5.1). The posterior mean length is larger than values inferred or assumed in earlier studies, which generally only take into account movement of insect vectors [Halkka et al., 1971, Lago et al., 2018, Plazio et al., 2017, Weaver and King, 1954, White et al., 2017] (see Figure 5.4).

Marginal and 2D posterior distributions of parameters are displayed in Figures 5.5 and 5.6. The introduction of Xf in Corsica tends to be relatively ancient. The first introduction, τ_0, has a posterior median of -842 (posterior mean: -822 months before July 2015, i.e. introduction around 1945; posterior standard deviation: 124 months). The second introduction τ_1 has a posterior median of -315 (posterior mean: -317 months before July 2015, i.e. introduction around 1988; posterior standard deviation: 96 months).



Data	Uncertainties	Reference
Under artificial conditions (laboratory flight mill) movement of 1 km within 1 h	Maximum potential, only 1 insect, upper limit, lab conditions	Lago et al. (2018)
Within 15 days after release, recapture within 60 m from release point. In a second experiment, recapture up to 100 m within 30 days	Preliminary data	Plazio et al. (2017)
The authors assumed an exponential dispersal mode of vector; have chosen mean dispersal distance of 100 m based on paper by Blackmer et al. (2004) dealing with *H. vitripennis*	Information from *H. vitripennis* and not *Philaenus spumarius*	White et al. (2017)
Olive grove: • Maximum distance 100–155 m in 7–12 days • The majority of the marked-insects were captured within 60 m from the release point • In summer population, more stationary than autumn (80% of the marked insects were found within 18 m from the release point) • Thus, in Apulia > dispersal capacity in autumn than in summer (favoured by the emergence of the weeds and ground vegetation)		Plazio et al. (2017)
Spread velocity of vector and disease combined is, on average of 2×10^5 m^2/month, (median = 1.8×10^5, sd = 0.7×10^5), that corresponds roughly to a front velocity of 155 m/month (spread is circular and isotropic)		Preliminary data (Abboud et al., 2018)
Active migration by flight is probably limited to distances of about 40–80 m	Only observation	Halkka et al. (1971)
Philaenus spumarius move 30 m in a single flight, marked adults 100 m in 24 h. Adults usually fly within 60 cm from the ground, can fly as high as 60 m		Weaver and King (1954)

Figure 5.4.: Vector local spread published in [EFSA Panel on Plant Health (PLH) et al., 2019].

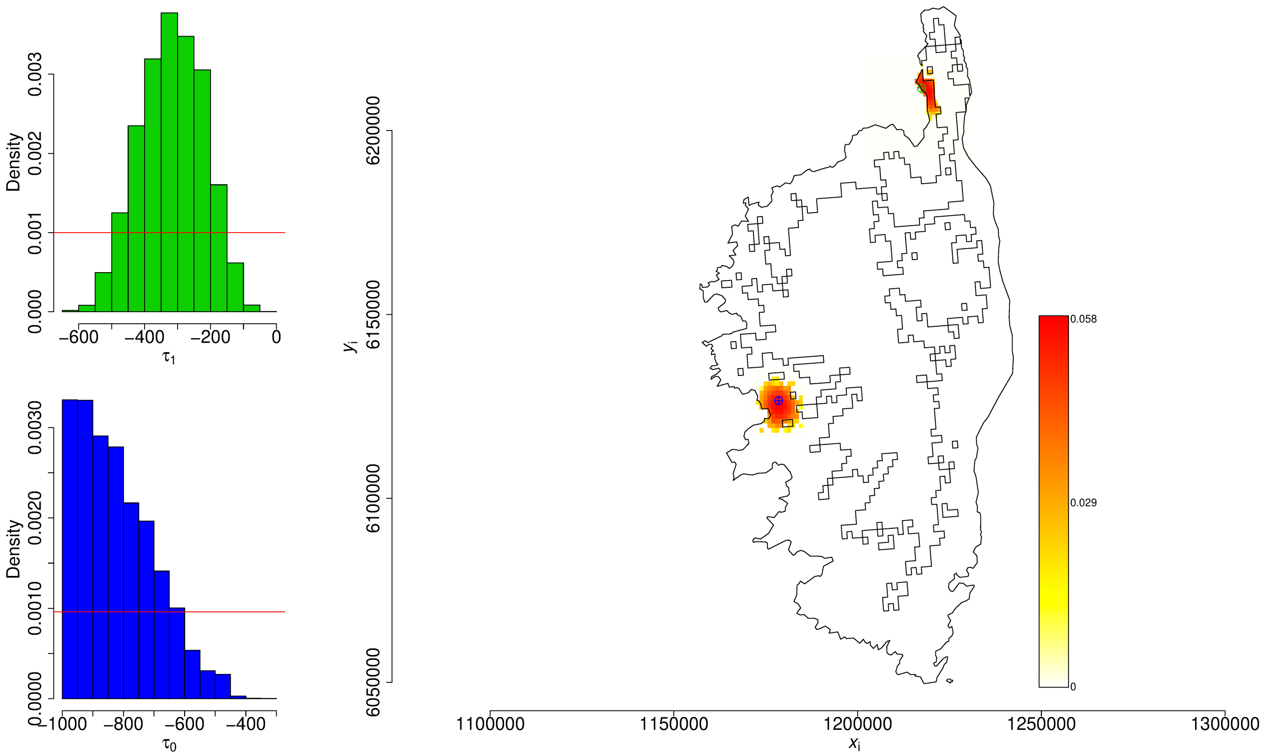


Figure 5.5.: Marginal posterior distributions of the introduction times τ_0 (histogram in blue) and τ_1 (histogram in green), and 2D posterior distributions of the introduction locations $\mathbf{x}_0$ and $\mathbf{x}_1$ (color palette). The prior for τ_0 and τ_1 was uniform (red line). The value of $\mathbf{x}_i$ having the largest weight in AMIS is indicated by a blue cross for $i = 0$ and by a green cross for $i = 1$. The prior for $\mathbf{x}_i$, $i = 1, 2$, was uniform over the space Ω_1 delimited by the contours.

Figure 5.7 shows the numerical solution of the model with two introductions for the values of Θ that has the largest posterior weight. This numerical solution is shown at the times of the introductions τ_0 and τ_1, at the time of the first observation in July 2015, and at the time of the last observation in April 2019. We notice that there are some discrepancies between the spatial distribution of positive cases sampled in the North-East and the East of the island, and the model's solution for the Θ associated to the largest posterior weight. Indeed, positive cases observed on the East coast are in regions with very low probability in April 2019. This could reflect an issue with the model incorporating only two introductions. However, it has to be noted that these positive cases occurred in areas with particularly low observed disease prevalence and are surrounded by lots of negative cases (see Figure 5.1). More investigations should be carried out to check the adequacy between the model and the data, such as goodness-of-fit tests and simulations study, and to assess whether the models with larger numbers of introductions could solve this issue.

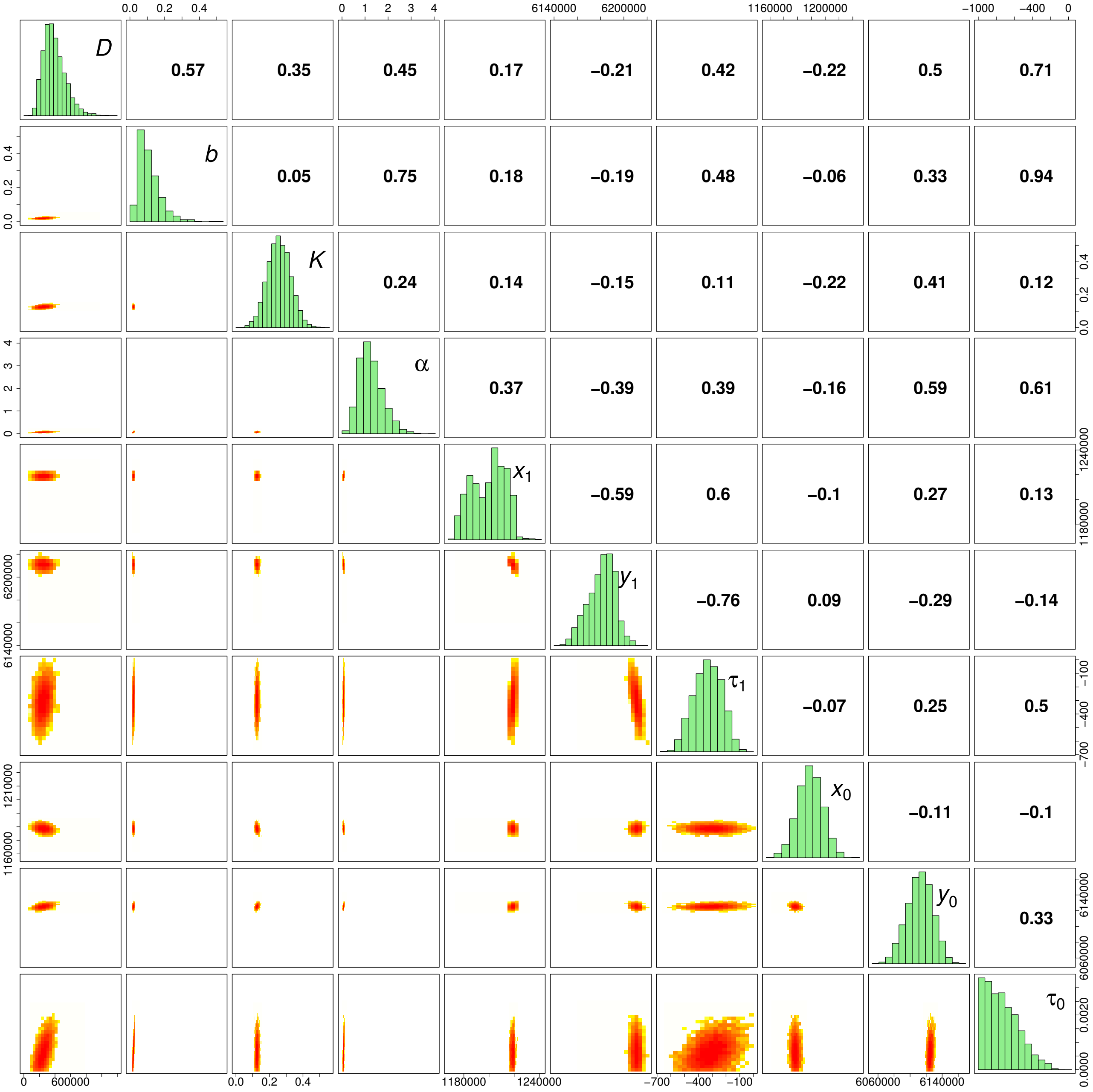


Figure 5.6.: Marginal posterior distributions of parameters (panels in the diagonal) and 2D posterior distributions of parameters over 20×20 grids (panels in the lower triangle). Figures in the upper triangle panels provide correlation coefficients.

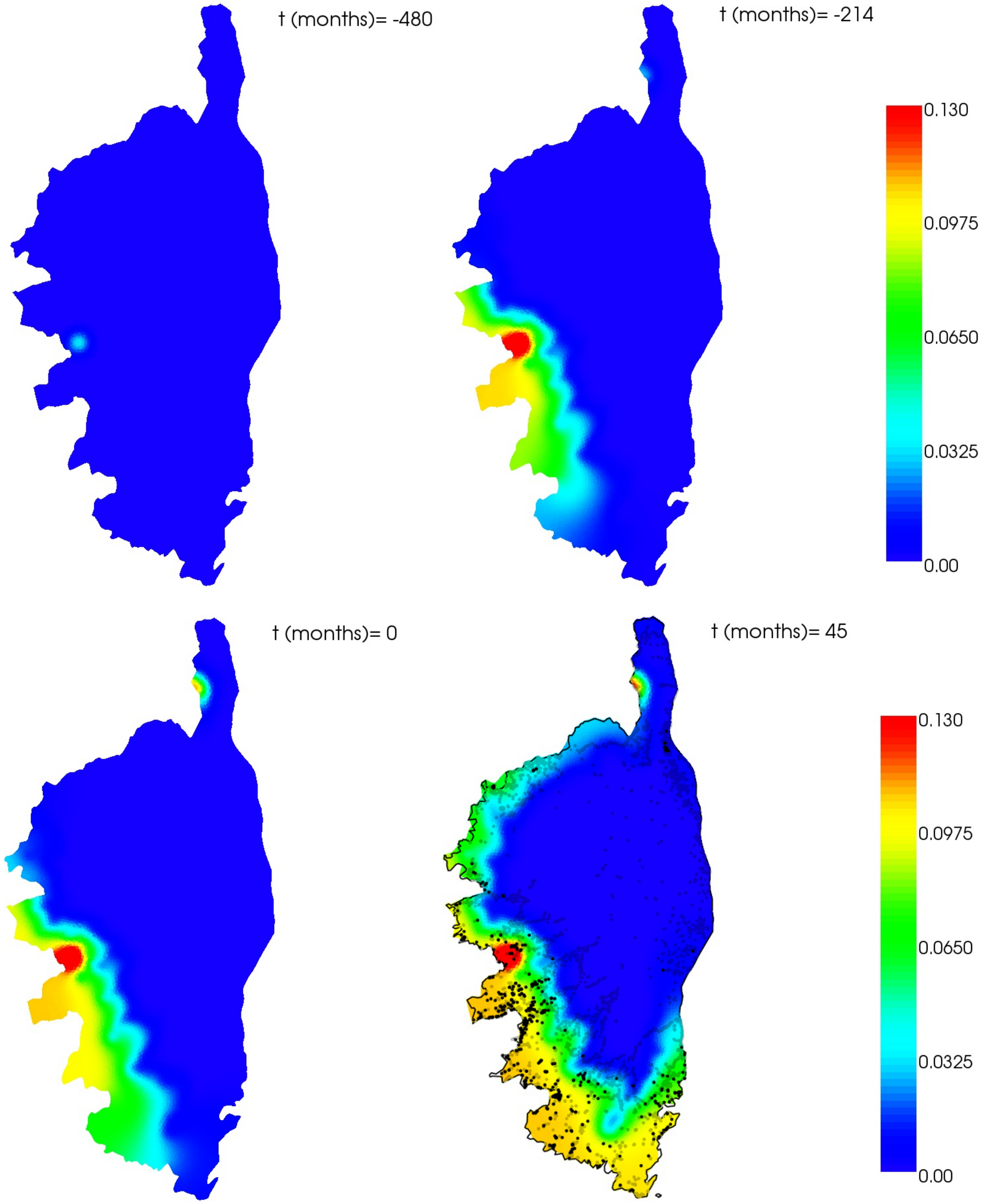


Figure 5.7.: Numerical solution $\{X_t\}_t$ of Equations (5.1)–(5.5) in Corsica with Neumann boundary conditions, at time $t = \tau_0$, $t = \tau_1$, $t = 0$ the time of the first observation in May 2015 and $t = 45$ time of the last observation in April 2019. $(D, b, K, \alpha) = (3.1e + 05, 0.03, 0.1, 0.1)$. The first initial condition was $u(\tau_0, \mathbf{x}_0) = 0.01 \exp(-(5 \times 10^3 \|\mathbf{x} - \mathbf{x}_0\|)^2)$, where $\mathbf{x}_0 = (x_0, y_0) = (1178530, 6.126555)$. The second introduction occurred in $\mathbf{x}_1 = (x_1, y_1) = (1217379, 6211505)$. Locations of Xf-positive (black dots) and negative (grey dots) cases observed between July 2015 and April 2019.

5.4. Raised questions drawn from Chapter 5

- ❖ When superposing the data to the model solution (see Figure 5.7) we notice some obvious discrepancies. These discrepancies may be due in part to the continuous nature of the PDE, and also to the choice of the domain spatial partition only based on a measure of winter temperature, and to the need for supplementary introductions in the model. However, to correctly judge the adequacy of the model to the data, the reliability and the sharpness of the approach, we need to perform goodness-of-fit tests.

- ❖ Lower standard deviation values were obtained for model parameters compared to the values obtained in Chapter 3. Does this means that incorporating two introductions in the model has decreased the uncertainty about the parameters? Is it due to the supplementary data that were considered or to the failure of the inference algorithm to fully explore the parameter space? To answer these questions, a complementary study could be conducted to calibrate the credibility intervals under simulation studies.

- ❖ Eventhough the model with two introductions has lower standard deviations for the introduction times, these quantities remain strongly uncertain. To give more accurate results about these parameters, it is important to apportion the different sources of uncertainty, and as proposed in Chapter 3, to use additional information brought by genetic data for example.

- ❖ The preliminary inference achieved in this chapter, raises questions about whether the AMIS algorithm will still be tractable in a high dimensional problem such as a spatio-temporal PDMP with multiple introductions. An intuitive way for estimating the PDMP if the number of introductions N is low, consists of assuming a prior for N weighting only a few small values (e.g., between 1 and 10), estimating model parameters with AMIS for each possible value of N, and integrating over N the product between the conditional posterior of the parameters given N and the prior of N, as we did for $\tilde{T}$ in the BMA approach. The DAG in Figure 5.8 should be used to correctly design this approach. If N is potentially large, another algorithm not conditional on N (i.e., a sort of reversible jump algorithm) should certainly be developed.

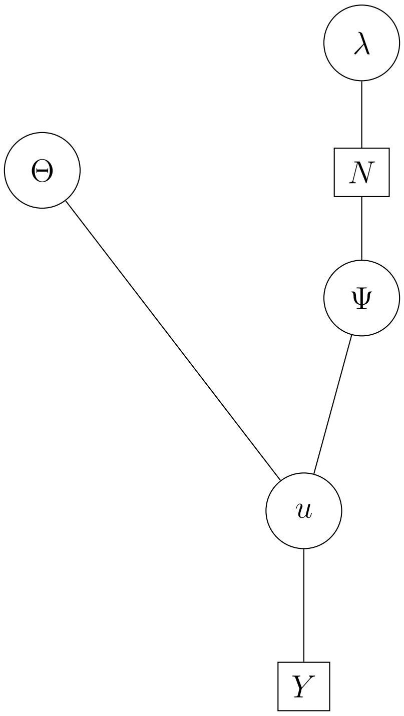


Figure 5.8.: DAG for inferring the Bayesian posterior distribution of the PDMP parameters and the latent variables (jump times and introduction locations).

6. Conclusion

6.1. Summary & Discussions

The main objective of the thesis is to propose an efficient methodology scalable to various invasive quarantine pests for which no dedicated model is available and whose initial introductions (date and time) are unknown. The detection of Xf in France for the first time during 2015 is the real case study that has motivated our research and enhanced the usefulness of the proposed methods.

How have we addressed the research questions raised in Section 1.2 towards meeting this aim, and what information have we brought forward to the decision-makers for supporting the decision process about the quarantine pest Xf?

Task #1: Tackle the problem of recovering the initial conditions of the pathogen based on post-introduction data

To achieve the first goal of my thesis, I followed the approach proposed by [Soubeyrand and Roques, 2014] with a simple reaction-diffusion model and a MCMC algorithm to link the model to simulated data. Thus, I adopted a mechanistic-statistical approach that handles the characterizations of the surveillance-based observation process. This approach is grounded on a coupled reaction-diffusion-absorption model that describes the dynamics of the invasive species through time and heterogeneous space with respect to growth. Initial conditions and model parameters were jointly estimated in a Bayesian framework. With this approach, one can first recover a distribution for any biological parameters of the model and the distribution of the initial location of invasion and the starting time of the invasion. Moreover, the challenge achieved was to control the degree of convergence of the proposed approach. This mathematical–statistical analysis is sufficiently general to be used for various types of invasions and can efficiently recover biological as well as demographic parameters of a biological invasion.

Using deterministic models offers a concise and parsimonious description of the pathogen dynamics

When one aims to link data to model, one should find a compromise between model realism and estimation difficulty. Deterministic models, in particular parabolic PDE, offer a rich and flexible modelling framework that has been applied to a large number of invasions. Eventhough a PDE does not describe all the processes involved in the pathogen dynamics, it can, however, help understanding its essential properties and inferring its major components.

In this first task, we focused on the inference of the date and location of pathogen introduction after assuming that only one introduction governs the emergence of the pathogen and that eventual subsequent introductions have negligible effects on the dynamics. However, estimating this introduction point requires the estimation of the propagation characteristics, and *vice versa* because these characteristics link the introduction to the observations. Thus, I conducted a joint estimation of the PDE parameters and initial conditions in a Bayesian framework. The limitation of this approach lies in the fact that the PDE model could not represent the small-scale variations as shown in the conducted analyses of Local Brier Scores.

Bayesian inference better represents parameter uncertainties

The benefit of the Bayesian approach is primarily to allow the incorporation of prior expertise into the statistical analysis and the rigorous assessment of dependencies and uncertainties in estimation via the joint posterior distribution of parameters. Unfortunately, informative priors are not always easily provided due to the lack of information on model parameters or because experts prefer to use objective (i.e., non-informative) priors, while a prior is never entirely objective [Kass and Wasserman, 1996]. Moreover, even if the Bayesian approach is generally more costly than the frequentist approach, the Bayesian approach can directly assess uncertainty in parameter estimation conversely to the frequentist approach, which requires an additional task to fulfill this assessment (e.g., via the computation of the Fisher information matrix or via parametric bootstrap). Besides, most people better understand the direct probabilistic judgments about the unknowns provided by the Bayesian paradigm when reporting uncertainty [O'Hagan, 2008].

AMIS has advantageous features

The main motivation for using AMIS is the gain in computation time comparing to Markov chain Monte Carlo (MCMC). In particular, AMIS is interesting in cases where the likelihood is computationally expensive [Cornuet et al., 2012] because all particles simulated during the process are recycled, which minimizes the number of calls of the likelihood function. This applies to our case, in the sense that non-negligible computation time is needed to accurately solve the PDE and obtain an accurate inference about the model parameters. Another milestone towards a less expensive inference approach is that AMIS can be easily parallelized. On the other hand, tuning coefficients of the proposal distribution are automatically adapted across the algorithm iterations, contrary to the basic MCMC and the ML approach frequently used in the mechanistic-statistical framework. This can be rather challenging, even in adaptive MCMC algorithms, where the convergence properties of nonhomogeneous Markov Chains are roughly achievable [Haario et al., 2001, Liu, 2008]. It has, however, to be noted that AMIS has to be appropriately initialized, which can be relatively easily done in practice by evaluating the marginal posterior distributions over 1D grids. Limitation in the AMIS algorithm shows up in Chapter 5, where one introduction was added to the inference approach. Adding three components to the parameters vector caused a slower algorithm stabilization and a potential difficulty in exploring the parameter space.

Selection criteria have determined the threshold of temperature under which Xf is hampered

To infer the posterior distribution of the parameter vector we proceed in two steps: (i) inference of the parameters of the dynamics given the temperature threshold $\tilde{T}$ used for partitioning the study domain with AMIS algorithm, and (ii) selection of $\tilde{T}$ using four criteria: the BIC, two versions of the DIC and IC. The limitation of this approach is that the uncertainty about the temperature threshold $\tilde{T}$ is not quantified. Ideally, to be able to give a better representation of this parameter (i.e., a random representation in the Bayesian paradigm) and to assess its uncertainty, two solutions can be proposed. The first solution is to jointly infer $\tilde{T}$ with the other parameters of the dynamics (but this solution induce technical adaptations of the algorithm that are expected to negatively impact computation times), and the second is to use the BMA approach as shown in Chapter 4 where the posterior model probabilities coincide with the posterior distribution of $\tilde{T}$. The temperature threshold obtained for the case study of Xf in our applications is presently estimated between 5 and 5.5 Celsius degrees. Note that, before submitting our work on BMA to a journal, we will consider supplementary temperature thresholds (e.g., every 0.25 °C) to provide a smoother posterior distribution of $\tilde{T}$.

Xf probably did not recently emerge in France

The conducted analyses tend to show that the introduction of Xf in South Corsica occurred probably near Ajaccio around 1959 (95%-posterior interval: $[1933, 1986]$), long time before its first detection in 2015. Our results on the estimation of the introduction time are relatively consistent with the results obtained by Denancé et al. [2017a] who assessed the introduction of the two main strains found in Corsica around 1965 and 1980, respectively, using a phylogenetic approach. Likewise, our estimation is compatible with the result of Soubeyrand et al. [2018], who dated the introduction around 1985 (95%-posterior interval: $[1978, 1993]$) with a statistical analysis of temporal data (indeed, the posterior intervals obtained from both studies overlap). Moreover, we helped in gaining knowledge about the spread velocity of vector and disease combined, which is obviously linked to the introduction time. The values obtained are on average of 2×10^5 m2/month (median $= 1.8 \times 10^5$, sd $= 0.7 \times 10^5$), that corresponds roughly to a front velocity of 155 m/month.

Task #2: Investigate the application of BMA in the context of pathogen dynamics using PDE models

To achieve this task, I have considered models grounded on a family of reaction-diffusion equations with eventual spatially heterogeneous diffusion and reproduction terms. I have computed, from post-introduction data, the BMA posterior distribution of the introduction time and location of the pathogen and its future spatial extent. Following the approach presented in Chapter 3, I have applied to each model the AMIS algorithm for providing an empirical approximation obtained via a weighted sample of the posterior distribution, given the specified model. Then, for drawing BMA posterior samples, I have computed posterior probabilities of models using different approximations of the integrated likelihood that have been proposed in the literature. I have first applied the approach on simulated data and then applied it to make predictions concerning the real case study of Xf in Southern Corsica, France.

BMA is expected to better reflect uncertainty about model predictions and provide more realistic inferences. However, in our case study, this technique does not seem to outperform on the best model

Based on the ample literature on *model-averaging*, we were expecting this technique to provide ameliorated predictions and a more realistic estimate of the uncertainty associated with model predictions than any single model [Hoeting et al., 1999, Draper, 1995, Wintle et al., 2003]. However, our application shows that the BMA does not seem to outperform the best model. In this case study, as discussed by Wintle et al. [2003], this may be due in part to the dominance of one or two predictors among all the models or the lack of complete independence between the training and testing data. Likewise, in our application, the credibility intervals provided with BMA were larger than the ones provided by the best model. This could reflect the fact that BMA has succeeded in better assessing the uncertainty about model predictions, avoiding overconfidence about predictions and misleading decision making obtained when using a single best model. However, to firmly confirm this result, complementary studies should be conducted to calibrate the credibility intervals.

BMA offers a direct way to model comparison

In addition to combining model inferences and predictions, the BMA can be used for a direct model comparison by computing model posterior probabilities. To compute these probabilities, we have proceeded trying various evidence-based and predictive methods proposed in the literature. Most of the used estimators agreed in that the best model is the one, for which the diffusion rate is spatially homogeneous and the threshold of temperature is equal to 5.5 °C, and associated the lowest weights to the same models, but differed in the detailed ranking. One non-biased estimator that we should implement, and whose application is pursuant to the nature of AMIS, can be obtained by averaging over the unnormalized importance weights computed at the final iteration of the algorithm [Bugallo et al., 2015]. This approach, viewed as a benchmark, might allow us to identify which of the implemented methods is the most consistent.

Model selection ignores model uncertainty

In Chapter 4, we first infer the posterior distribution of model parameters given the threshold of temperature, and then this threshold was chosen using selection criteria. This approach does not allow the quantification of the uncertainty about the selected value. The BMA approach presented in Chapter 4 allows ones to open up to smoothed threshold of temperature by means of the empirical approximation of the BMA marginal posterior distribution.

Task #3: Unravelling multiple introductions of pathogen dynamics

The third task is considered to be an extension of Task #1.

The PDE models used in Task #1 are generally not adapted to describe the dynamics of populations that expand their range not only by neighborhood dispersal but also because of new introductions in disease-free areas and by long-distance dispersal, which can correspond to rare but significant events. The term *stratified diffusion* was used to describe this twofold dispersal process [Hengeveld, 1989]. To further the work done in Task #1, *stratified diffusion* models and PDMP can be designed to incorporate into the model not only long-distance dispersal but also multiple introductions. Distinguishing these two types of events from surveillance data is not easy in general, except if one has at disposal genetic data or contact tracing data, but can anyway be modeled separately with a mixture of two kernels [Sapoukhina et al., 2010]. However, identifiability issues of the mixture components may arise if information in data is limited to infer in details the dispersal processes.

The coalescing colony model is a PDMP

PDMP have often been built to model temporal processes. Spatio-temporal PDMP are also occasionally encountered in the theoretical and quantitative population dynamics literature, even if they are not called PDMP. For instance, spatio-temporal PDMP have been built at the population level [Shigesada et al., 1995], the metapopulation level [Soubeyrand et al., 2009a] (which is a set of populations) and the individual level [Caillerie, 2017]. I have precisely shown in Chapter 5 that the coalescing colony model of [Shigesada et al., 1995], which was developed to represent *stratified diffusion* in biological invasions, can be formulated as a spatio-temporal PDMP.

Inferring parameters of the spatio-temporal PDMP

To achieve the third task, I proposed a spatio-temporal PDMP embedding multiple introductions. Then, I used a Bayesian inference with AMIS algorithm to jointly estimate the parameters of the model and its initial conditions for a fixed number of introductions. In the case of Xf, this allowed the extension of the study domain to the entire Corsican island. In fact, Xf multiplex is the only subspecies that has been observed in Corsica, but two sequence types (see Chapter 2) have been found with Xf-positive cases, which could be due to multiple introductions. Thus, the spatio-temporal PDMP offers a more realistic framework for the dynamics of Xf in Corsica. The preliminary inference achieved in this chapter, raises concerns about whether the AMIS algorithm will still be tractable in a high dimensional problem such as a spatio-temporal PDMP with multiple introductions (i.e, the number of introductions is no longer a fixed parameter, but a jump process). How can we adapt it to reduce computational time and allow for faster stabilization?

6.2. Outlook

To achieve the goal of the thesis, three main tasks have been performed in light of various parsimonious choices and assumptions. To conclude the thesis, the possibility of having a different standpoint on these choices and relaxing the considered assumptions is investigated. Several improvements and engaging perspectives, which are subject to future research, are also distinguished. The prospects I propose here, are divided into two parts: (i) the generic methodological prospects, which are related to the modelling, the inference and the prediction tools, and (ii) the application prospects, which concerns amelioration to best adapt the approach to the specific case study of Xf.

Generic Methodological Prospects

***"[...] even when the world is indeed a well-defined closed system, [...] different modellers can generate different nonequivalent descriptions of it, [...] more than one model may be compatible with the same set of data or evidence."*[Saltelli et al., 2008]**

The deterministic model that we proposed in Chapter 3 to represent the disease has survived a series of tests. However, this model only takes into account a few biological drivers of the disease dynamics. These drivers could be implicitly handled by investigating alternative representations of disease dynamics:

- ❖ A possible extension of this modelling approach is to replace the deterministic model by a stochastic version, for instance, an SDE model that would allow relaxing hypotheses made on the dynamics. However, this is not the ideal trade-off between model realism and easiness of inference, compared to the PDMP approach. Nevertheless, this should not prevent us from comparing stochastic and deterministic models in the framework of the BMA approach presented in Chapter 4. By averaging over miscellaneous stochastic and deterministic competing models of the dynamics, consistently with theoretical findings, the BMA would be expected to clearly outperform the best model.

- ❖ An additional perspective can be investigated by aggregating models in the BMA framework, to compare: (i) models including only short distance dispersal, with homogeneous/heterogeneous diffusion as the set of models already proposed in Chapter 4, to (ii) models allowing a finer quantification of local and long distance dispersal such as integro-differential equations [Bonnefon et al., 2014], spatial contact models [Mollison, 1977], mixed dispersal kernel models [Clark et al., 1998], stratified dispersal models [Shigesada et al., 1995] or PDMP [Abboud et al., 2019a]. Models incorporating both short and long distance dispersal are generally expected to yield better predictions [Higgins and Richardson, 1999, Nathan et al., 2008, Fayard et al., 2009, Gilioli et al., 2013, White et al., 2017].

- ❖ Estimating the parameters of the PDE in a Bayesian framework turned out to have a relatively high computational cost. To reduce this cost, approximating the input/output relation in the mechanistic model using meta-models necessitating less computer intensive calculations could be a valuable option, that could be incorporated in AMIS

[Osio and Amon, 1996, Giunta and Watson, 1998]. In particular, kriging meta-models show up to be an adequate solution for approximating deterministic models since they interpolate the observed or known data points [Simpson et al., 2001].

❖ All the aggregated PDE models in Chapters 3 and 4 assume a logistic form of the reproduction term. In Chapter 5, we proposed a PDMP model including an Allee effect. However, the exact form of the reproduction term is generally not known; for instance, there are many cases where one does not know whether there is an Allee effect or not. Thus, the following questions arise:

(1) Is it important to include an Allee effect in a model when this effect is present in the population?

(2) Does including an Allee effect when this effect is not present in the population will result in misleading predictions?

(3) Which method can be adapted to predict the expansion speed when one does not know if an Allee effect is present?

To address the above questions one can consider a continuous family of reaction-diffusion models $(\mathcal{M}_\rho(\theta))$, indexed by a parameter ρ and with coefficient θ:

$$\partial_t u = D\,\Delta u + f_\rho(u),\ t \geq 0,\ x \in \mathbb{R}^n, \tag{6.1}$$

where $u(t,x)$ represents the population density at time t and position $x \in \mathbb{R}^n$. The growth functions f_ρ are defined as follows:

$$f_\rho(u) = r\,u\,(1 - u/K)\,(u + \rho), \tag{6.2}$$

with $r > 0$. We note that $f_\rho(0) = f_\rho(K) = 0$. For this type of growth functions, provided that the initial condition $u_0(x) := u(0,x)$ satisfies $0 \leq u_0 \leq 1$, an immediate consequence of the parabolic maximum principle [Protter, MH and Weinberger, HF, 1967, Roques, 2013a] is

$$0 \leq u(t,x) \leq K \text{ for all } t \geq 0,\ x \in \mathbb{R}^n.$$

Depending on the value of ρ the model $\mathcal{M}_\rho$ either describes a strong Allee effect, a weak Allee effect, or the absence of Allee effect. In that respect:

- when $\rho \geq K$, there is no Allee effect, and the model fulfills the standard KPP assumption;
- when $0 \leq \rho < K$, $f_\rho(u)$ always remains positive for $u \in (0,K)$, but the maximum *per capita* growth rate is not reached as $u \to 0$: there is a weak Allee effect;
- when $-K/2 \leq \rho < 0$, $f_\rho(u) < 0$ for $u \in (0,\rho)$: there is a strong Allee effect;
- when $\rho < -K/2$, u always converges uniformly to 0 for large t, as $u(t,x)$ is smaller than the solution of $U'(t) = f_\rho(U)$ with $U(0) = K$ (it is a consequence of a standard comparison principle), which itself converges to 0.

The advantage of this type of models, is that analytic formulas for the spreading speeds v_ρ associated with the models $(\mathcal{M}_\rho)$ can be computed [Hadeler and Rothe, 1975] as follows:

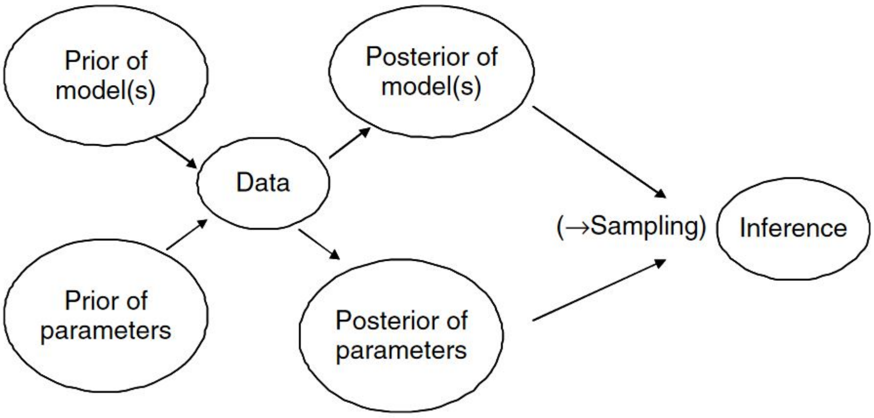


Figure 6.1.: Bayesian *model-averaging* as illustrated in [Saltelli et al., 2008] (Chapter 1, page 9).

- when $\rho \geq K/2, v_\rho = 2\sqrt{r\,\rho\,D}$;
- when $0 \leq \rho < K/2$, $v_\rho = \sqrt{rD}\left(\rho\sqrt{2}/\sqrt{K} + \sqrt{K}/\sqrt{2}\right)$;
- when $-K/2 \leq \rho < 0$, $v_\rho = \sqrt{2\,r\,D}(\sqrt{K}/2 - \rho/\sqrt{K})$.

The spreading speed is generally of paramount importance in population dynamics studies [Malchow et al., 2007, Mistro et al., 2012], and being able to derive explicit formula for the spreading speed allows ones to get an accurate description for it that can then be explicitly used for estimation purposes.

"***Uncertainty is everywhere and you cannot escape from it***" [Lindley, 2006].

This citation means that the uncertainty is always present when modelling and inferring a certain phenomenon, but it is important to quantify it. In my thesis, I have provided a coherent judgment quantifying a subjective degree of uncertainty in model output. In other words, I have provided a posterior distribution of the unknown parameters conditionally to data and assessed uncertainties about model output. To reduce these uncertainties and make our models more suitable for decision-making, it is important to precise how these quantified uncertainties can be apportioned to different sources of uncertainty in the model input [Saltelli et al., 2004]. This medium is the so-called sensitivity analysis. Ideally, if we had more time, we should have run sensitivity analyses in tandem with the uncertainty analyses, as proposed by [Saltelli et al., 2008]. Here, I propose a sensitivity analysis for two model inputs:

❖ **Input: The Prior distribution**

In the BMA analysis, one has to specify the priors for the candidate models and their respective parameters. This choice is crucial because the prior governs how posterior mass is spread among models, and it provides shrinkage correction for the estimated parameters. However, there may be few information available about the model and its parameters. Even if available, some experts still prefer an objective prior, to see what kind of information can be deducted from the data. For instance, in the case of Xf, a phylogenetic analysis of the data collected in Corsica and in regions where Xf has been present for a long period and which could be the source of the infection in Corsica, can help in eliciting an informative prior on the introduction date [Denancé et al., 2017a]. In addition, an informative prior of the dispersal rate can be elicited by using the knowledge about the movements of insect vectors. However, movements of infected seedlings, for which there is no precise knowledge, has also an influence on the diffusion rate (see Section 2.1.2.3). Thus, building an informative prior of the dispersal rate, which is only based on vectors movements, may lead to a "biased" prior distribution. Consequently, a non-informative prior should be embedded into the analysis. Typically, we have used a uniform prior over the model space in which we incorporated some biological knowledge and equal prior weights for all the models. An interesting perspective of our work on the BMA would be to run a sensitivity analysis regarding the prior's influence.

- **Input: Models and parameters**

  We have quantified the uncertainties about model and parameters using the BMA approach, illustrated by [Saltelli et al., 2008] as in Figure 6.1. A sensitivity analysis could be executed after sampling the parameters and determining the posterior weights of the models. This analysis addresses the questions on how much of the uncertainty is due to the model selection, and on how much is due to the estimation of the parameters.

PDMP Embedding Multiple Introductions: A Promising Research Project

Due to lack of time, the analyses on the PDMP have not been accomplished. However, I tried to explore as much as possible, the proposal made in [Abboud et al., 2018] on inferring a PDMP with multiple introductions. This seems to be a promising avenue of research. Ideally, the work done on inferring PDMP can be forwarded in many interesting ways:

- In Chapter 5, I have fixed the number of introductions to two. An approach that one can tailor is to first infer the posterior distribution of model parameters given the number of introductions, for instance from 1 to 10 introductions, and then, this number can be inferred in the BMA framework, as we have done for the threshold of temperature in Chapter 4.

- An interesting avenue is to explore spatio-temporal PDMP with a parametric model of the conditional intensity of the jump process to describe the introduction process, such as the inhomogeneous Poisson process incorporating spatial covariables. Typically, these covariables could be environmental variables affecting the growth and mortality of invasive species.

- Another avenue of research for making more precise inferences to unravel the characteristics of the introduction process, is to incorporate genomic data (sub-families, sequence type, or genome of the pathogen) that may provide information about the jumps of the spatio-temporal PDMP.

- The AMIS algorithm that we adapted to a PDMP with a fixed number of introductions should be improved to better explore the parameter space, in order to apply it in a more complex framework where the introductions are governed by a spatio-temporal point process. An interesting approach that should be investigated in this respect is the *Markov adaptive multiple importance sampling* algorithm [MAMIS; Martino et al., 2015]. This algorithm applies the iterative IS approach using an adaptive proposal distribution. The location coefficients of the proposals are adapted according to an MCMC technique, such as Metropolis-Hastings or Gibbs sampler (see Section 2.2.2.2.3). The main difference with respect to the AMIS lies in the more streamlined adaptation procedure of MAMIS. Moreover, this MCMC-Driven Adaptive Multiple Importance Sampling technique reduces the dependence on the choice of the cloud of proposals, since the proposal density in the MCMC method can be adapted in order to optimize the performance. This approach is supposed to make "the best of both worlds".

Application Prospects

In Chapter 4, we use surveillance data collected only in South Corsica to estimate Xf introduction and the temperature threshold over which this bacterium is hampered, and then, predict its future spatial extent. A possible complement of our study is to extrapolate the risk beyond South Corsica. Figure 6.2 shows a basic example of static risk maps. These maps are constructed over a grid with a spatial resolution of 1×1km, using the two temperature thresholds $\tilde{T}$ that have the highest posterior weights in the BMA approach. Our static risk map indicates where Xf is likely to spread if it is introduced by only accounting on winter temperatures. It is worth noting that Xf-positive cases have been found in Northern Corsica, in PACA region and in the Balearic Islands in Spain, which are located in areas at high risk where $\{x : T(x) > 5\}$ or in their vicinity.

Our temperature-based risk map seems to be consistent with the map provided by EFSA Panel on Plant Health (PLH) et al. [2019] and shown in Figure 6.3, where Apulia, Corsica, the North-western Mediterranean coast, and the Balearic Islands are the areas at the highest risk. The consistency of our map with those provided by Godefroid et al. [2018] and Martinetti and Soubeyrand [2019] is however less clear (see Figures 6.4 and 6.5). Note that the risk maps provided by EFSA Panel on Plant Health (PLH) et al. [2019], Martinetti and Soubeyrand [2019], Godefroid et al. [2018] are obtained from analyses incorporating numerous environmental explanatory variables. To ameliorate the static prediction obtained from our model, we should incorporate other important factors influencing the propagation of Xf to partition the study domain. In particular, we could focus on variables, which have been identified by Martinetti and Soubeyrand [2019] with a machine learning approach, as very influential variables (e.g., precipitation seasonality, solar radiation, precipitation during the dry season).

Beyond the estimation of risk maps, Martinetti and Soubeyrand [2019] proposed sampling algorithms based on risk to improve the surveillance of Xf in terms of early detection. Likewise, our static risk map can be used to conceive sampling strategies in a risk-based sampling approach. But we could design epidemio-surveillance surveys by considering dynamical risk maps instead of the static risk map shown in Figure 6.2. At least two settings can be considered from this perspective: (1) adding a time dimension to our risk map in the context of global warming; and (2) exploiting the inferred information about disease diffusion to adapt the risk map with time.

Concerning point (1), global change in climate may generally impact population dynamics [Malchow et al., 2007, Mistro et al., 2012]. In particular, climatic change under global warming is susceptible to cause a significant increase in winter temperatures [Maxwell et al., 1992]. Therefore, the spread of Xf may probably increase, because cold winter temperatures are considered to be the regulatory "curing" mechanism of Xf dynamics [Purcell, 1977, Anas et al., 2008, Feil and Purcell, 2001]. In our approach, the sub-domain where Xf is susceptible to propagate will extend with the increase of winter temperatures. This, suggests that epidemio-surveillance strategies should be adapted to the inter-annual fluctuations of winter temperatures.

Concerning point (2), the results obtained in Chapter 3 provided an estimate of the spread velocity of vector and disease combined, which has an average of 2×10^5 m^2/month, (median $= 1.8 \times 10^5$, sd $= 0.7 \times 10^5$). This velocity corresponds roughly to a front velocity of 155 m/month. Hence, surveillance schemes could be designed to take into account the potential spread around already detected foci. For instance, the subspecies *pauca* has been recently detected for the first time on an olive tree on the East coast of PACA https://agriculture.gouv.fr/la-contamination-par-xylella-fastidiosa-de-2-oliviers-confirmee-en-paca). The surveillance of *pauca* on olive trees will be obviously reinforced around this focus and more largely in the whole region. Concerning the surveillance around the focus, our result about the dissemination of Xf should encourage to survey eventual *pauca*-infected olive trees in a time-increasing domain, whose border could move forward at a speed of approximately $155m \times 12 \approx 2km$ per year.

With the recent detection (August 2019) in Antibes and Menton of two olive trees infected by Xf (at least one of the trees being infected by the subspecies *pauca*, as mentionned above), the concern about this pathogen has raised in PACA and at the national level. A methodological work such as mine presented in this thesis contributes, from an applied perspective, to unravell the global behaviour of a new pathogen introduced in a new environment. The inferences that are made from such a methodological work can be exploited as scientific knowledge used to inform decision-makers and other stakeholders (e.g., the inferences that I obtained have been incorporated in the scientific opinion written by the European Food Safety Authority (EFSA) Panel on Plant Health and discussed in the scientific committee about Xf advising the Prefect of Corsica). With knowledge acquired since 2015 (e.g., by INRA research units and within the XF-ACTORS project) about the behaviour of Xf in European environments, the French authorities in charge of the surveillance and control of Xf are naturally better equipped for the potential future spread of the subspecies *pauca* recently detected in PACA than they were in 2015 when Xf was first *in situ* detected in Corsica and PACA.

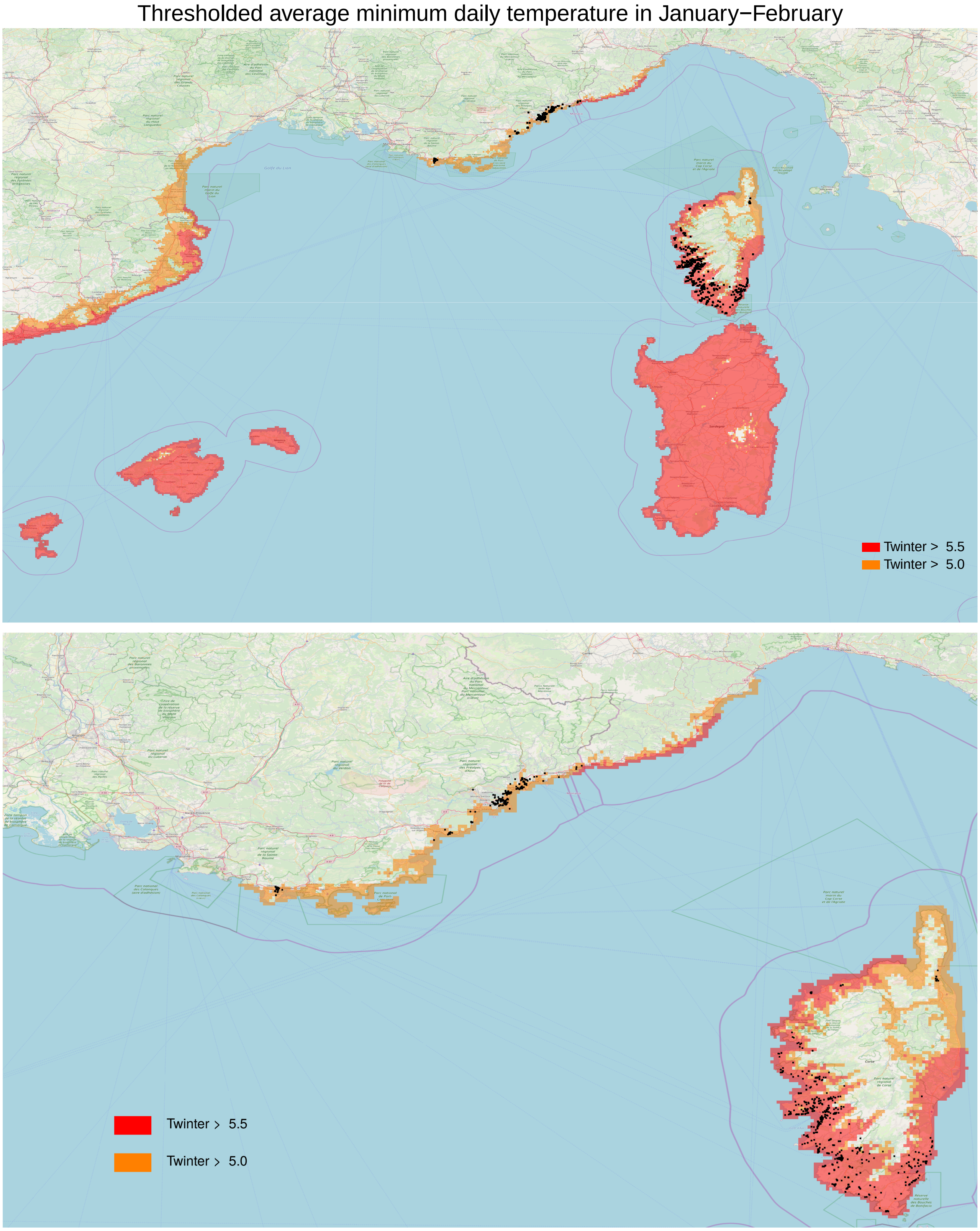

Thresholded average minimum daily temperature in January−February
Twinter > 5.5
Twinter > 5.0
Twinter > 5.5
Twinter > 5.0

Figure 6.2.: Risk maps constructed over a grid with spatial resolution of 1 × 1km, using the thresholds of temperatures $\tilde{T}$ (in Celsius degrees) associated to the models with the highest posterior probabilities in the BMA approach, namely 5°C and 5.5°C, and the location of Xf-positive cases (black dots) sampled between July 2015 and September 2019 in France.

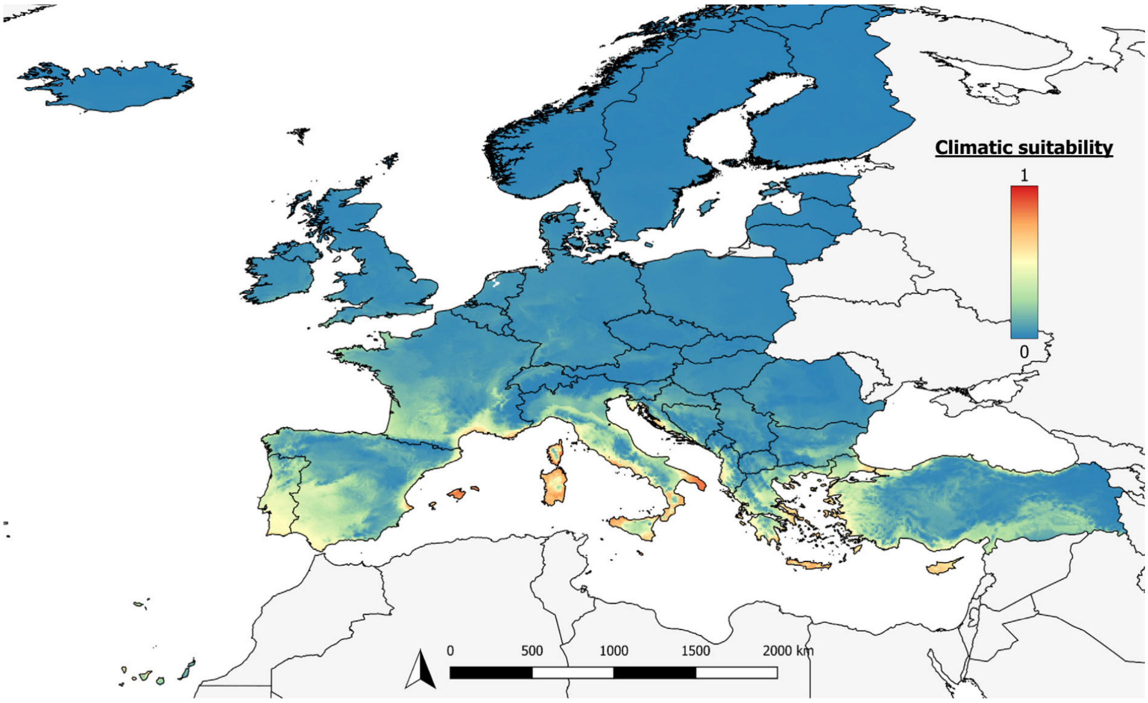


Figure 6.3.: Potential for the establishment of Xf according to climatic suitability, using an ensemble predictions model, published in EFSA Panel on Plant Health (PLH) et al. [2019].

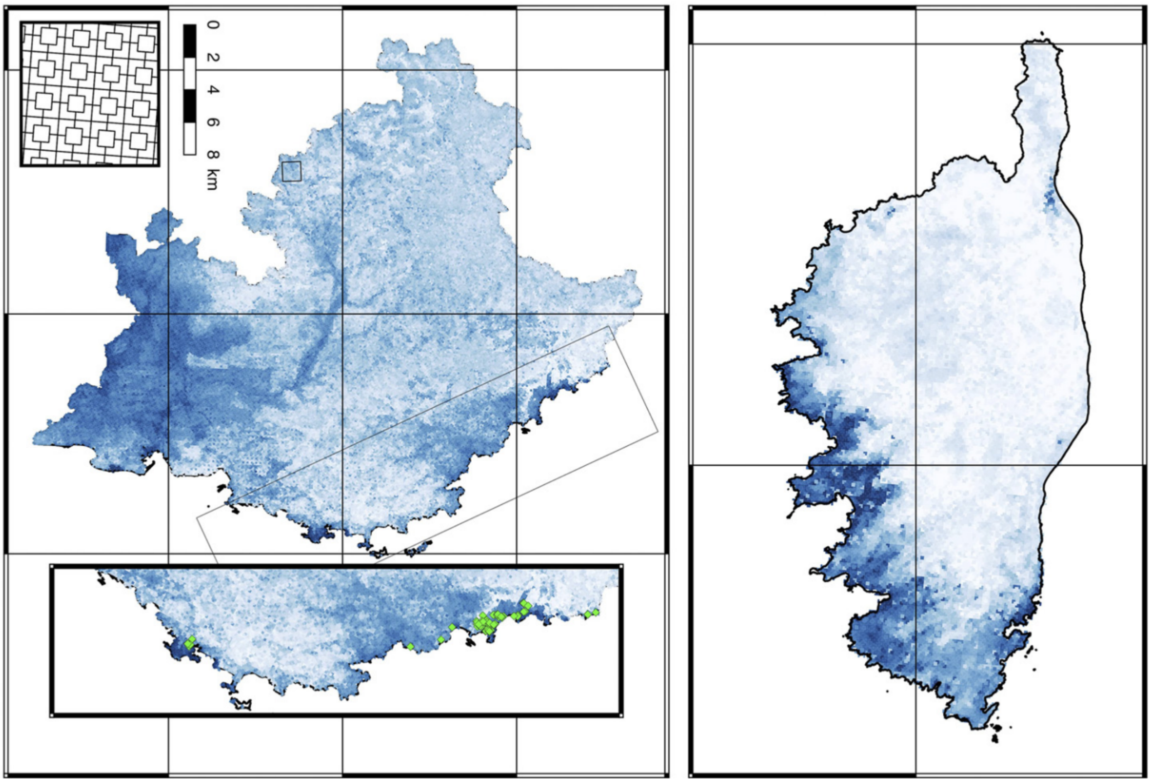


Figure 6.4.: Risk maps in Corsica and PACA inferred from the surveillance data of Xf collected in Corsica, published in Martinetti and Soubeyrand [2019].

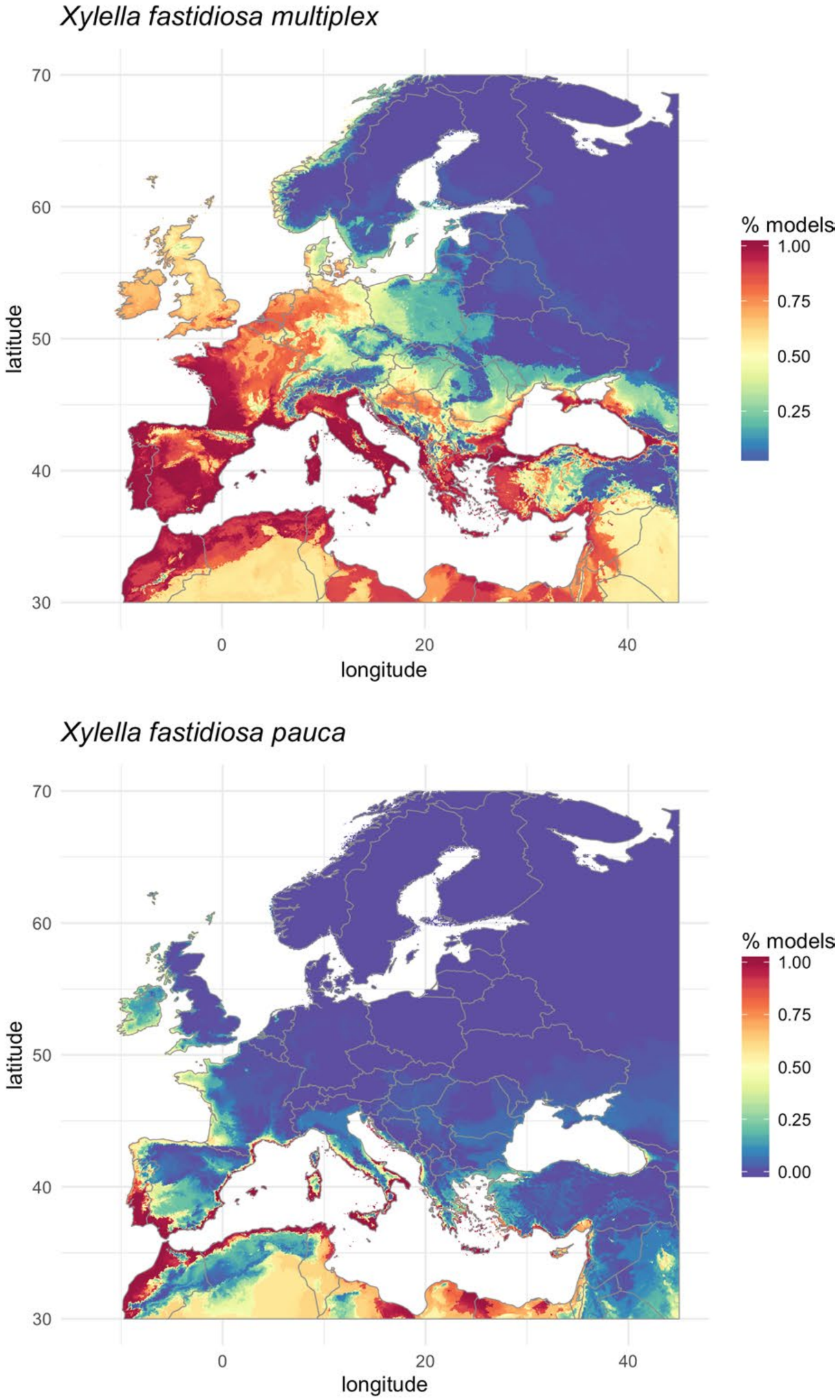


Figure 6.5.: Potential distribution of *multiplex* (on the left) and *pauca* (on the right) subspecies of Xf published in Godefroid et al. [2018].

Bibliography

C. Abboud, R. Senoussi, and S. Soubeyrand. Piecewise-deterministic Markov Processes for Spatio-temporal Population Dynamics. In Azaïs, Romain and Bouguet, Florian, editor, *Statistical Inference for Piecewise-deterministic Markov Processes*. ISTE Editions/Wiley, 2018.

C. Abboud, O. Bonnefon, E. Parent, and S. Soubeyrand. Dating and localizing an invasion from post-introduction data and a coupled reaction–diffusion–absorption model. *Journal of Mathematical Biology*, 2019a.

C. Abboud, O. Bonnefon, E. Parent, and S. Soubeyrand. Model&data-based prediction of pathogen dynamics. in progress, 2019b.

G. N. Agrios. *Plant pathology*. Elsevier Academic Press, San Diego, CA, 5 edition, 2005.

H. Akaike. Information theory as an extension of the maximum likelihood principle. In *Petrov, BN and Csaki, F*, pages 267–281, Akademiai Kiado, Budapest, 1973.

M. Alfaro, J. Coville, and G. Raoul. Travelling waves in a nonlocal reaction-diffusion equation as a model for a population structured by a space variable and a phenotypic trait. *Communications in Partial Differential Equations*, 38:2126–2154, 2013.

R. P. P. Almeida, C. Wistrom, B. L. Hill, J. Hashim, and A. H. Purcell. Vector transmission of xylella fastidiosa to dormant grape. *Plant Disease*, 89:419–424, 2005.

O. Anas, U. J. Harrison, P. M. Brannen, and T. B. Sutton. The effect of warming winter temperature on the severity of pierce's disease in the appalachian mountains and piedmont of the southeastern united states. *Plant Health Progress.*, 101094:450–459, 2008.

R. M. Anderson, C. A. Donnelly, N. M. Ferguson, M. E. J. Woolhouse, C. J. Watt, H. J. Udy, S. Mawhinney, S. P. Dunstan, T. R. E. Southwood, J. W. Wilesmith, J. B. M. Ryan, L. J. Hoinville, J. E. Hillerton, A. R. Austin, and G. A. H. Wells. Transmission dynamics and epidemiology of BSE in British cattle. *Nature*, 382:779–788, 1996.

H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

T. Ando. Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, 31:13–38, 2011.

D. Andow, P. M. Kareiva, S. A. Levin, and A. Okubo. Spread of invading organisms. *Landscape Ecology*, 4:177–188, 1990.

D. Andow, P. Kareiva, S. Levin, and A. Okubo. Spread of invading organisms: patterns of spread. *Evolution of insect pests: the pattern of variations*, pages 219–242, 1993.

R. Azaïs and F. Bouguet. *Statistical Inference for Piecewise-deterministic Markov Processes*. ISTE Editions/Wiley, 2018.

C. Baker, G. Bocharov, J. Ford, P. Lumb, S. Norton, C. Paul, T. Junt, P. Krebs, and B. Ludewig. Computational approaches to parameter estimation and model selection in immunology. *Journal of Computational and Applied Mathematics*, 184:50 – 76, 2005.

H. G. Baker. The continuing evolution of weeds. *Economic Botany*, 45:445–449, 1991.

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

D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'académie royale des sciences*, pages 1–45, 1760.

O. Bonnefon, J. Coville, J. Garnier, and L. Roques. Inside dynamics of solutions of integro-differential equations. *Discrete & Continuous Dynamical Systems-B*, 19:3057–3085, 2014.

E. L. Boone, K. Ye, and E. P. Smith. Assessment of two approximation methods for computing posterior model probabilities. *Computational Statistics & Data Analysis*, 48:221 – 234, 2005.

E. L. Boone, S. J. Simmons, H. Bao, and A. E. Stapleton. Bayesian hierarchical regression models for detecting qtls in plant experiments. *Journal of Applied Statistics*, 35:799–808, 2008.

C. Bordier, H. Dechatre, S. Suchail, M. Peruzzi, S. Soubeyrand, M. Pioz, M. Pélissier, D. Crauser, Y. Le Conte, and C. Alaux. Colony adaptive response to simulated heat waves and consequences at the individual level in honeybees (apis mellifera). *Scientific reports*, 7:3760, 2017.

L. Bosso, D. Russo, M. D. Febbraro, G. Cristinzio, and A. Zoina. Potential distribution of *Xylella fastidiosa* in Italy: a maximum entropy model. *Phytopathologia Mediterranea*, 55: 62–72, 2016.

G. E. Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 1976.

R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18:125–135, 2008.

S. Brooks. Bayesian computation : A statistical revolution. *Trans. Roy. Statist. Soc., series A*, 15:2681–2697, 2003.

C. G. Broyden. A new double-rank minimisation algorithm. preliminary report. In *Notices of the American Mathematical Society*, volume 16, page 670, 1969.

M. F. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49, 2015.

K. P. Burnham, G. C. White, and D. R. Anderson. Model selection strategy in the analysis of capture-recapture data. *Biometrics*, 51:888–898, 1995.

N. Caillerie. *Stochastic and deterministic kinetic equations in the context of mathematics applied to biology*. Theses, Université de Lyon, July 2017.

G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46: 167–174, 1992.

V. Cavalieri, C. Dongiovanni, D. Tauro, G. Altamura, M. Di Carolo, G. Fumarola, M. Saponari, and D. Bosco. Transmission of the codiro strain of xylella fastidiosa by different insect species. In *Proceedings of the XI European Congress of Entomology, Naples, Italy*, pages 2–6, 2018.

N. Champagnat, R. Ferrière, and S. Méléard. Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models. *Theoretical Population Biology*, 69:297 – 321, 2006. ESS Theory Now.

D. S. Chapman, S. M. White, D. A. Hooftman, and J. M. Bullock. Inventory and review of quantitative models for spread of plant pests for use in pest risk assessment for the EU territory. *EFSA Supporting Publications*, 12, 2015.

J. S. Clark, C. Fastie, G. Hurtt, S. T. Jackson, C. Johnson, G. A. King, M. Lewis, J. Lynch, S. Pacala, C. Prentice, E. W. Schupp, T. Webb, III, and P. Wyckoff. Reid's paradox of rapid plant migration: Dispersal theory and interpretation of paleoecological records. *BioScience*, 48:13–24, 1998.

M. Clyde and E. I. George. Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62:681–698, 2000.

D. Cornara, V. Cavalieri, C. Dongiovanni, G. Altamura, F. Palmisano, D. Bosco, F. Porcelli, R. P. P. Almeida, and M. Saponari. Transmission of xylella fastidiosa by naturally infected *Philaenus spumarius* (hemiptera, aphrophoridae) to different host plants. *Journal of Applied Entomology*, 141:80–87, 2017.

J. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39:798–812, 2012.

M. Costello, S. Steinmaus, and C. Boisseranc. Environmental variables influencing the incidence of Pierce's disease. *Australian Journal of Grape and Wine Research*, 23:287–295, 2017.

D. Dacunha-Castelle and M. Duflo. *Probabilités et Statistiques. Problèmes à Temps Mobile*, volume 2. Masson, Paris, 1982.

I. Dattner, E. Miller, M. Petrenko, D. E. Kadouri, E. Jurkevitch, and A. Huppert. Modelling and parameter inference of predator&#x2013;prey dynamics in heterogeneous environments using the direct integral approach. *Journal of The Royal Society Interface*, 14:20160525, 2017.

M. Daugherty, D. Bosco, and R. Almeida. Temperature mediates vector transmission efficiency: inoculum supply and plant infection dynamics. *Annals of Applied Biology*, 155: 361–369, 2009.

M. P. Daugherty, A. R. Zeilinger, and R. P. P. Almeida. Conflicting effects of climate and vector behavior on the spread of a plant pathogen. *Phytobiomes Journal*, 1:46–53, 2017.

M. Davis, A. Purcell, and S. Thomson. Isolation media for the pierce's disease bacterium. *Phytopathology*, 70:425–429, 1980.

M. H. A. Davis. Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46: 353–376, 1984.

N. Denancé, S. Cesbron, M. Briand, A. Rieux, and M.-A. Jacques. Is *Xylella fastidiosa* really emerging in France? . In J. Costa and R. Koebnik, editors, *1st Annual Conference of the EuroXanth - COST Action Integrating Science on Xanthomonadaceae for integrated plant disease management in Europe*, volume 7, Coimbra, Portugal, 2017a. EuroXanth.

N. Denancé, B. Legendre, M. Briand, V. Olivier, C. Boisseson, F. Poliakoff, and M.-A. Jacques. Several subspecies and sequence types are associated with the emergence of *Xylella fastidiosa* in natural settings in France. *Plant Pathology*, 66:1054–1064, 2017b.

O. Diekmann and J. A. P. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. John Wiley & Sons, LTD, England, 2000.

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57:45–70, 1995.

EFSA Panel on Plant Health (PLH), C. Bragard, K. Dehnen-Schmutz, F. D. Serio, P. Gonthier, M. Jacques, J. A. J. Miret, A. F. Justesen, A. MacLeod, C. S. Magnusson, P. Milonas, J. A. Navas-Cortés, R. Potting, P. L. Reignault, H. Thulke, W. v. d. Werf, A. V. Civera, J. Yuen, L. Zappalà, D. Boscia, D. Chapman, G. Gilioli, R. Krugner, A. Mastin, A. Simonetto, J. R. S. Lopes, S. White, J. C. Abrahantes, A. Delbianco, A. Maiorano, O. Mosbach-Schulz, G. Stancanelli, M. Guzzo, and S. Parnell. Update of the scientific opinion on the risks to plant health posed by xylella fastidiosa in the eu territory. *EFSA Journal*, 17, 5 2019.

T. S. Eicher, C. Papageorgiou, and A. E. Raftery. Default priors and predictive performance in bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, 26:30–55, 2011.

L. C. Evans. *Partial differential equations*, volume 19 of *Graduate studies in mathematics*. American Mathematical Society, Providence, Rhode Island, 1998.

N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, D. Posada, M. Peeters, O. G. Pybus, and P. Lemey. The early spread and epidemic ignition of hiv-1 in human populations. *Science*, 346:56–61, 2014.

J. Fayard, E. K. Klein, and F. Lefèvre. Long distance dispersal and the fate of a gene from the colonization front. *Journal of evolutionary biology*, 22:2171–2182, 2009.

H. Feil and A. H. Purcell. Temperature-dependent growth and survival of *Xylella fastidiosa* in vitro and in potted grapevines. *Plant Disease*, 85:1230–1234, 2001.

H. Feil, W. S. Feil, and A. H. Purcell. Effects of date of inoculation on the within-plant movement of *Xylella fastidiosa* and persistence of Pierce's disease within field grapevines. *Phytopathology*, 93:244–251, 2003.

C. Fernández, E. Ley, and M. F. J. Steel. Bayesian modelling of catch in a north-west atlantic fishery. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51:257–280, 2002.

P. C. Fife. *Mathematical Aspects of Reacting and Diffusing Systems*. Springer-Verlag Berlin Heidelberg, 1979.

R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7:355–369, 1937.

R. Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13, 1970.

J. Frantzen. *Epidemiology and Plant Ecology: Principles and Applications*. World Scientific Publishing Company, USA, 2007.

N. Frazier. Xylem viruses and their insect vectors. In *Proceedings of the international conference on virus and vectors on perennial hosts, with special reference to Vitis*, pages 91–99, 1965.

R. Friis. *Epidemiology 101*. Jones & Bartlett Learning, 2017.

R. Friis and T. Sellers. *Epidemiology for public health practice.* Jones and Bartlett Publishers, Sudbury, USA, 3 edition, 2004.

R. A. Gatenby and E. T. Gawlinski. A reaction-diffusion model of cancer invasion. *Cancer research*, 56:5745–5753, 1996.

A. E. Gelfand and A. F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Gelman, G. O. Roberts, W. R. Gilks, et al. Efficient metropolis jumping rules. *Bayesian statistics*, 5:599–608, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Texts in statistical science series. Chapman & Hall/CRC, New York, 2 edition, 2003.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

G. Gilioli, S. Pasquali, and F. Ruggeri. Nonlinear functional response parameter estimation in a stochastic predator-prey model. *Mathematical Biosciences & Engineering*, 9:75, 2012.

G. Gilioli, S. Pasquali, S. Tramontini, and F. Riolo. Modelling local and long-distance dispersal of invasive chestnut gall wasp in europe. *Ecological Modelling*, 263:281 – 290, 2013.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, London, 1996.

M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard. Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. *Molecular Biology and Evolution*, 30:713–724, 11 2012.

C. S. Gillespie and A. Golightly. Bayesian inference for generalized stochastic population growth models with application to aphids. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:341–357, 2010.

A. Giunta and L. Watson. A comparison of approximation modeling techniques-Polynomial versus interpolating models. In *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, page 4758, St. Louis, MO, U.S.A., 1998. Multidisciplinary Analysis Optimization Conferences.

P. Gloaguen, M.-P. Etienne, and S. Le Corff. Stochastic differential equation based on a multimodal potential to model movement data in ecology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67:599–619, 2018.

M. Godefroid, A. Cruaud, J.-C. Streito, J.-Y. Rasplus, and J.-P. Rossi. Climate change and the potential distribution of Xylella fastidiosa in Europe. *bioRxiv*, 2018.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24:23–26, 1970.

D. Gonze, J. Halloy, and A. Goldbeter. Deterministic versus stochastic models for circadian rhythms. *Journal of biological physics*, 28:637–653, 2002.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7: 223–242, 2001.

K. Hadeler and F. Rothe. Travelling fronts in nonlinear diffusion equations. *Journal of Mathematical Biology*, 2:251–263, 1975.

O. Halkka, M. Raatikainen, L. Halkka, , and J. Lokki. Factors determining the size and composition of island population of *philaenus spumarius (l.) (hom.)*. *Acta entomologica Fennica*, 28:83–100, 1971.

W. H. Hamer. *Epidemic disease in England: the evidence of variability and of persistency of type*. Bedford Press, 1906.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 04 1970.

R. Hengeveld. *Dynamics of biological invasions*. Springer Science & Business Media, New York, 1989.

T. S. Henneberger. *Effects of low temperature on populations of Xylella fastidiosa in sycamore*. PhD thesis, University of Georgia, 2003.

J. Heydari, C. Lawless, D. A. Lydall, and D. J. Wilkinson. Fast bayesian parameter estimation for stochastic logistic growth models. *Biosystems*, 122:55 – 72, 2014.

S. I. Higgins and D. M. Richardson. Predicting plant migration rates in a changing world: the role of long-distance dispersal. *The American Naturalist*, 153:464–475, 1999.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors. *Statist. Sci.*, 14:382–417, 1999.

T. A. Huld, M. Šúri, E. D. Dunlop, and F. Micale. Estimating average daytime and daily temperature profiles within europe. *Environmental Modelling & Software*, 21:1650–1661, 2006.

P. E. Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46:10–18, 2009.

D. R. Jones and R. H. A. Baker. Introductions of non-native plant pathogens into great britain, 1970–2004. *Plant Pathology*, 56:891–910, 2004.

P. Kareiva and N. Shigesada. Analyzing insect movement as a correlated random walk. *Oecologia*, 56:234–238, 1983.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370, 1996.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *The Royal Society*, 115:700–721, 1927.

W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics–I. 1927. *Bull. Math. Biol.*, 53:33–55, 1991.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19: 27–43, 1982.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

I. P. A. Kolomogorov and N. Piscouno. Etude de l'equation de la diffusion avec croissance de la quantite de la matiere et son application a un problem biologique. *Moscow University Bull. Math*, 1:1–25, 1937.

I. Kyrkou, T. Pusa, L. Ellegaard-Jensen, M.-F. Sagot, and L. H. Hansen. Pierce's disease of grapevines: A review of control strategies and an outline of an epidemiological model. *Frontiers in microbiology*, 9, 2018.

C. Lago, E. Garzo, A. Moreno, and A. Fereres. Flight behaviour and patterns of directional movement on *P*hilaenus spumarius. In *Proceedings of the European Research on Emerging Plant Diseases – Contributions of the H2020 projects POnTE and XF-Actors – 2nd Joint Annual Meeting*, page 74, Valencia, Spain, 2018. IVIA.

E. C. Lamon and M. A. Clyde. Accounting for model uncertainty in prediction of chlorophyll a in lake okeechobee. *Journal of Agricultural, Biological, and Environmental Statistics*, 5: 297–322, 2000.

V. Lánska. Minimum contrast estimation in diffusion processes. *Journal of Applied Probability*, 16(1):65–75, 1979.

E. Lanzarone, S. Pasquali, G. Gilioli, and E. Marchesini. A Bayesian estimation approach for the mortality in a stage-structured demographic model. *Journal of mathematical biology*, 75:759–779, 2017.

E. Leamer. *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. John Wiley & Sons Incorporated, 1978.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, Paris, France, 1805.

Lewis, MA and Kareiva, P. Allee dynamics and the spread of invading organisms. *Theoretical Population Biology*, 43:141–158, 1993.

J. Li, J.-L. Wu, and G. Zhang. Estimation of intrinsic growth factors in a class of stochastic population model. *Stochastic Analysis and Applications*, 37:602–619, 2019.

D. Lindley. *Understanding Uncertainty*. John Wiley & Sons, INC, Hoboken, New Jersey, 2006.

J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

R. Luther. Räumliche Ausbreitung chemischer Reaktionen. *Zeitschrift für Elektrochemie*, 12: 596–600, 1906.

T. Luzyanina, D. Roose, and G. Bocharov. Distributed parameter identification for a label-structured cell population dynamics model using cfse histogram time-series data. *Journal of Mathematical Biology*, 59:581, Dec 2008.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

D. Madigan, J. Gavrin, and A. E. Raftery. Enhancing the predictive performance of bayesian graphical models. *Communications in Statistics-Theory and Methods*, 24:2271–2292, 1995.

H. Malchow, S. V. Petrovskii, and E. Venturino. *Spatiotemporal patterns in ecology and epidemiology: theory, models, and simulation*. Chapman and Hall/CRC, 2007.

D. C. R. K. Y. A. D. S. D. B. D. B. G. P. M. R. K. F. P. Maria Saponari, Giuliana Loconsole. Infectivity and transmission of *Xylella fastidiosa* by *Philaenus spumarius* (hemiptera: Aphrophoridae) in apulia, italy. *Journal of Economic Entomology*, 107:1316 – 1319, 2014.

J.-M. Marin, P. Pudlo, and M. Sedki. Consistency of the adaptive multiple importance sampling. *arXiv preprint arXiv:1211.2548*, 2012.

D. Martinetti and S. Soubeyrand. Identifying lookouts for epidemio-surveillance: Application to the emergence of xylella fastidiosa in france. *Phytopathology*, 109:265–276, 2019.

L. Martino, V. Elvira, D. Luengo, and J. Corander. Mcmc-driven adaptive multiple importance sampling. In A. Polpo, F. Louzada, L. L. R. Rifo, J. M. Stern, and M. Lauretto, editors, *Interdisciplinary Bayesian Statistics*, pages 97–109, Cham, 2015. Springer International Publishing.

B. Maxwell et al. Arctic climate: potential for change under global warming. In *Arctic ecosystems in a changing climate: an ecophysiological perspective*, pages 11–34, San Diego, 1992. Academic Press.

P. McCullagh. *Generalized linear models*. Routledge, 2019.

S. Méléard. Random modeling of adaptive dynamics and evolutionary branching. In F. A. C. C. Chalub and J. F. Rodrigues, editors, *The Mathematics of Darwin's Legacy*, pages 175–192. Springer Basel, Basel, 2011.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087–1091, 1953.

D. C. Mistro, L. A. D. Rodrigues, and S. Petrovskii. Spatiotemporal complexity of biological invasion in a space- and time-discrete predator–prey system with the strong allee effect. *Ecological Complexity*, 9:16 – 32, 2012.

D. Mollison. Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:283–326, 1977.

T. G. Müller, D. Faller, J. Timmer, I. Swameye, O. Sandra, and U. Klingmüller. Tests for cycling in a signalling pathway. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53:557–568, 2004.

J. Murray and P. Kulesa. On a dynamic reaction–diffusion mechanism: The spatial patterning of teeth primordia in the alligator. *Journal of the Chemical Society, Faraday Transactions*, 92:2927–2932, 1996.

R. Nathan, F. M. Schurr, O. Spiegel, O. Steinitz, A. Trakhtenbrot, and A. Tsoar. Mechanisms of long-distance seed dispersal. *Trends in ecology & evolution*, 23:638–647, 2008.

V. G. Oehler, K. Y. Yeung, Y. E. Choi, R. E. Bumgarner, A. E. Raftery, and J. P. Radich. The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data. *Blood*, 114:3292–3298, 2009.

A. Okubo. *Diffusion and ecological problems: mathematical models*, volume 10 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, 1980.

A. Okubo and S. Levin. *Diffusion and ecological problems - Modern Perspectives*. Springer-Verlag, New York, 2 edition, 2002.

I. G. Osio and C. H. Amon. An engineering design methodology with multistage bayesian surrogates and optimal sampling. *Research in Engineering Design*, 8:189–206, 1996.

A. O'Hagan. The Bayesian approach to statistics. *Handbook of probability: Theory and applications*, pages 85–100, 2008.

E. Parent and J. Bernier. *Le raisonnement bayésien: modélisation et inférence*. Springer Science & Business Media, 2007.

D. Parkinson and A. R. Liddle. Bayesian model averaging in astrophysics: a review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6:3–14, 2013.

R. O. Peterson, J. A. Vucetich, R. E. Page, A. Chouinard, et al. Temporal and spatial aspects of predator–prey dynamics. *Alces*, 39:215–232, 2003.

N. B. Pierce. *The California vine disease: a preliminary report of investigations*. US Government Printing Office, Washington, 1892.

D. Pimentel, R. Zuniga, and D. Morrison. Update on the environmental and economic costs associated with alien-invasive species in the united states. *Ecological Economics*, 52:273 – 288, 2005.

J. E. v. d. Plank. *Plant diseases: epidemics and control.* Academic Press, New York, 1963.

E. Plazio, N. Bodino, V. Cavalieri, E. Dongiovanni, G. Fumarola, A. Ciniero, L. Galetto, M. Saponari, and D. Bosco. Investigations on dispersal capability of philaenus spumarius by mark-release-recapture method. In *Book of Abstracts-European Conference on Xylella, Palma de Mallorca*, volume 56, 2017.

M. Porta. *A Dictionary of Epidemiology*. Oxford University Press, Inc., United Kingdom, 2008.

S. Portnoy and R. Koenker. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statist. Sci.*, 12:279–300, 11 1997. doi: 10.1214/ss/1030037960.

Protter, MH and Weinberger, HF. *Maximum Principles in Differential Equations*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

A. Purcell. Cold therapy of pierce's disease of grapevines. *Plant Disease Reporter*, 61:514–518, 1977.

A. Purcell et al. Environmental therapy for pierce's disease of grapevines. *Plant Disease*, 64: 388–390, 1980.

A. H. Purcell. Vector preference and inoculation efficiency as components of resistance to pierce's disease in european grape cultivars. *Phytopathology*, 71:429–435, 1981.

A. H. Purcell. *Homopteran Transmission of Xylem-Inhabiting Bacteria*, pages 243–266. Springer New York, New York, NY, 1990.

A. E. Raftery. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83:251–266, 1996.

A. E. Raftery and Y. Zheng. Discussion. *Journal of the American Statistical Association*, 98: 931–938, 2003.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174, 2005.

A. E. Raftery, M. Kárný, and P. Ettler. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52:52–66, 2010.

E. Renshaw. *Modelling biological populations in space and time*, volume 11. Cambridge University Press, Cambridge, 1993.

D. M. Richardson and W. J. Bond. Determinants of plant distribution: evidence from pine invasions. *The American Naturalist*, 137:639–668, 1991.

B. D. Ripley. Stochastic simulation, john willey & sons. *New York*, 1987.

C. Robert and G. Casella. Monte carlo statistical methods. *Springer-Verlag*, 1998.

G. Roberts and A. Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic Processes and their Applications*, 49:207 – 216, 1994.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

H. V. Roberts. Probabilistic prediction. *Journal of the American Statistical Association*, 60: 50–62, 1965.

L. Roques. *Modèles de réaction-diffusion pour l'écologie spatiale: Avec exercices dirigés*. Editions Quae, 2013a.

L. Roques. *Modèles de réaction-diffusion pour l'écologie spatiale: [avec exercices dirigés]*. Éd. Quae, Versailles, 2013b.

L. Roques, S. Soubeyrand, and J. Rousselet. A statistical-reaction–diffusion approach for analyzing expansion processes. *Journal of Theoretical Biology*, 274:43–51, 2011.

L. Roques, E. Walker, P. Franck, S. Soubeyrand, and E. Klein. Using genetic data to estimate diffusion rates in heterogeneous landscapes. *Journal of mathematical biology*, 73:397–422, 2016.

R. Ross. The Prevention of Malaria. *Public Health*, 24:287, 1911.

F. Rothe. Convergence to pushed fronts. *The Rocky Mountain Journal of Mathematics*, pages 617–633, 1981.

K. J. Rothman, S. Greenland, T. L. Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

D. B. Rubin. A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm. *Journal of the American Statistical Association*, 82(398): 543–546, 1987.

D. B. Rubin and N. Schenker. Efficiently simulating the coverage properties of interval estimates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 35:159–167, 1986.

A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. Sensitivity analysis in practice: a guide to assessing scientific models. *Chichester, England*, 2004.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

M. Saponari, D. Boscia, F. Nigro, and G. Martelli. Identification of dna sequences related to xylella fastidiosa in oleander, almond and olive trees exhibiting leaf scorch symptoms in apulia (southern italy). *Journal of Plant Pathology*, 95, 2013.

N. Sapoukhina, Y. Tyutyunov, I. Sache, and R. Arditi. Spatially mixed crops to control the stratified dispersal of airborne fungal diseases. *Ecological Modelling*, 221:2793 – 2800, 2010.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6:461–464, 1978.

D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24:647–656, 1970.

N. Shigesada and K. Kawasaki. *Biological invasions: theory and practice*. Oxford University Press, UK, 1997a.

N. Shigesada and K. Kawasaki. *Biological invasions: theory and practice*. Oxford series in ecology and evolution. Oxford University Press, Oxford, New York, 1 edition, 1997b.

N. Shigesada, K. Kawasaki, and Y. Takeda. Modeling Stratified Diffusion in Biological Invasions. *The American Naturalist*, 146:229–251, 1995.

A. Sicard, A. R. Zeilinger, M. Vanhove, T. E. Schartel, D. J. Beal, M. P. Daugherty, and R. P. Almeida. Xylella fastidiosa: Insights into an emerging plant pathogen. *Annual Review of Phytopathology*, 56:181–202, 2018.

A. H. Sidman, M. Mak, and M. J. Lebo. Forecasting non-incumbent presidential elections: Lessons learned from the 2000 election. *International Journal of Forecasting*, 24:237 – 258, 2008.

D. Simberloff. *Which insect introductions succeed and which fail?*, volume 37, pages 61–75. Wiley, Chichester, UK, 1989.

T. W. Simpson, J. Poplinski, P. N. Koch, and J. K. Allen. Metamodels for computer-based engineering design: survey and recommendations. *Engineering with computers*, 17:129–150, 2001.

J. G. Skellam. Random dispersal in theoretical populations. *Biometrika*, 38:196–218, 1951.

S. Soubeyrand and L. Roques. Parameter estimation for reaction-diffusion models of biological invasions. *Population Ecology*, 56:427–434, 2014.

S. Soubeyrand, A.-L. Laine, I. Hanski, and A. Penttinen. Spatio-temporal structure of host-pathogen interactions in a metapopulation. *The American Naturalist*, 174:308–320, 2009a.

S. Soubeyrand, S. Neuvonen, and A. Penttinen. Mechanical-statistical modeling in ecology: from outbreak detections to pest dynamics. *Bulletin of Mathematical Biology*, 71:318–338, 2009b.

S. Soubeyrand, L. Roques, J. Coville, and J. Fayard. Patchy patterns due to group dispersal. *Journal of Theoretical Biology*, 271:87 – 99, 2011.

S. Soubeyrand, P. de Jerphanion, O. Martin, M. Saussac, C. Manceau, P. Hendrikx, and C. Lannou. What dynamics underly temporal observations? Application to the emergence of *Xylella fastidiosa* in France: probably not a recent story. *New Phytologist*, 2018.

M. A. Spence, P. G. Blackwell, and J. L. Blanchard. Parameter uncertainty of a dynamic multispecies size spectrum model. *Canadian Journal of Fisheries and Aquatic Sciences*, 73: 589–597, 2016.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:583–639, 2002.

R. Stevens. How plants defend themselves. In J. Horsfall and A. Dimond, editors, *Plant Pathology, an Advanced Treatise*, volume 3, pages 357–429, New York, 1960. Academic Press.

G. Strona, C. J. Carstens, and P. S. Beck. Network analysis reveals why *Xylella fastidiosa* will persist in Europe. *Scientific Reports*, 7:71, 2017.

J. Timmer, T. G. Müller, I. Swameye, O. Sandra, and U. Klingmüller. Modeling the nonlinear dynamics of cellular signal transduction. *International Journal of Bifurcation and Chaos*, 14:2069–2079, 2004.

P. Turchin. *Quantitative Analysis of Movement: measuring and modeling population redistribution in plants and animals*. Sinauer, Sunderland, Massachusetts, 1998.

V. Viallefont, A. E. Raftery, and S. Richardson. Variable selection and bayesian model averaging in case-control studies. *Statistics in Medicine*, 20:3215–3230, 2001.

C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46:433–448, 1997.

S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11: 3571–3594, 2010.

S. Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.

C. R. Weaver and D. King. *Meadow spittlebug, Philaenus leucophthalmus (L.)*. Ohio Agricultural Experiment Station, 1954.

H. Weinberger. Asymptotic behavior of a model in population genetics. In J. Chadam, editor, *Nonlinear partial differential equations and applications*, pages 47–96. Springer, 1978.

S. Weisberg. *Applied Linear Regression*. Wiley, New York, NY, 2 edition, 1985.

J. M. Wells, B. C. Raju, H.-Y. Hung, W. G. Weisburg, L. Mandelco-Paul, and D. J. Brenner. Xylella fastidiosa gen. nov., sp. nov: gram-negative, xylem-limited, fastidious plant bacteria related to xanthomonas spp. *International Journal of Systematic and Evolutionary Microbiology*, 37:136–143, 1987.

S. M. White, J. M. Bullock, D. A. P. Hooftman, and D. S. Chapman. Modelling the spread and control of xylella fastidiosa in the early stages of invasion in apulia, italy. *Biological Invasions*, 19:1825–1837, 2017.

C. K. Wikle. Hierarchical models in environmental science. *International Statistical Review*, 71:181–199, 2003a.

C. K. Wikle. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84:1382–1394, 2003b.

B. A. Wintle, M. A. McCarthy, C. T. Volinsky, and R. P. Kavanagh. The use of bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, 17: 1579–1590, 2003.

K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21:2394–2402, 2005.

G. Yin and Y. Yuan. Bayesian model averaging continual reassessment method in phase i clinical trials. *Journal of the American Statistical Association*, 104:954–968, 2009.

Index

APPENDICES

A. Appendix for Chapter 3

Electronic Supplementary Material (EMS)

Dating and localizing an invasion from post-introduction data and a coupled reaction-diffusion-absorption model

C. Abboud[1], O. Bonnefon[1], E. Parent[2], and S. Soubeyrand[1]

[1]BioSP, INRA, 84914 Avignon, France

[2] AgroParisTech, UMR 518 Math. Info. Appli., Paris, France,
INRA, UMR 518 Math. Info. Appli., Paris, France

S1 Numerical Equation Solving

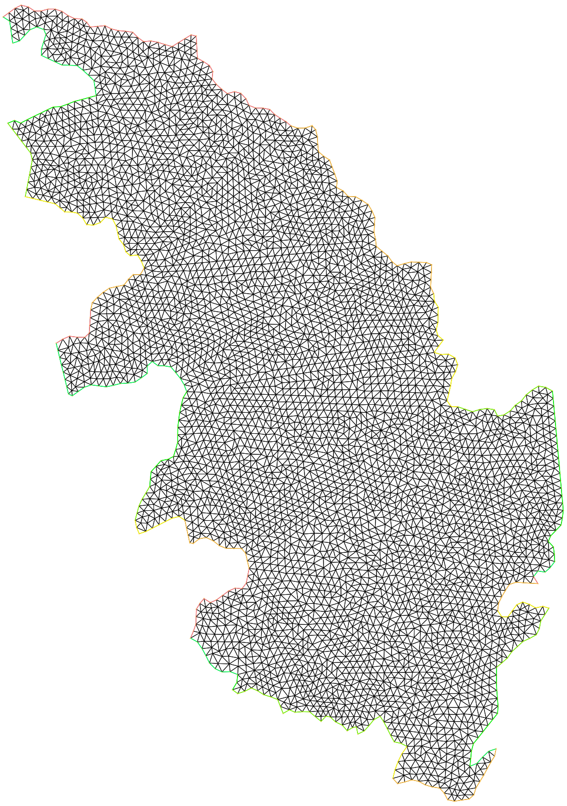


Figure S1: Mesh used for the Finite Element Method. This mesh contains 4791 nodes and 9141 triangles. The geometrical characteristics of this mesh were used to compute the accuracy of the simulator.

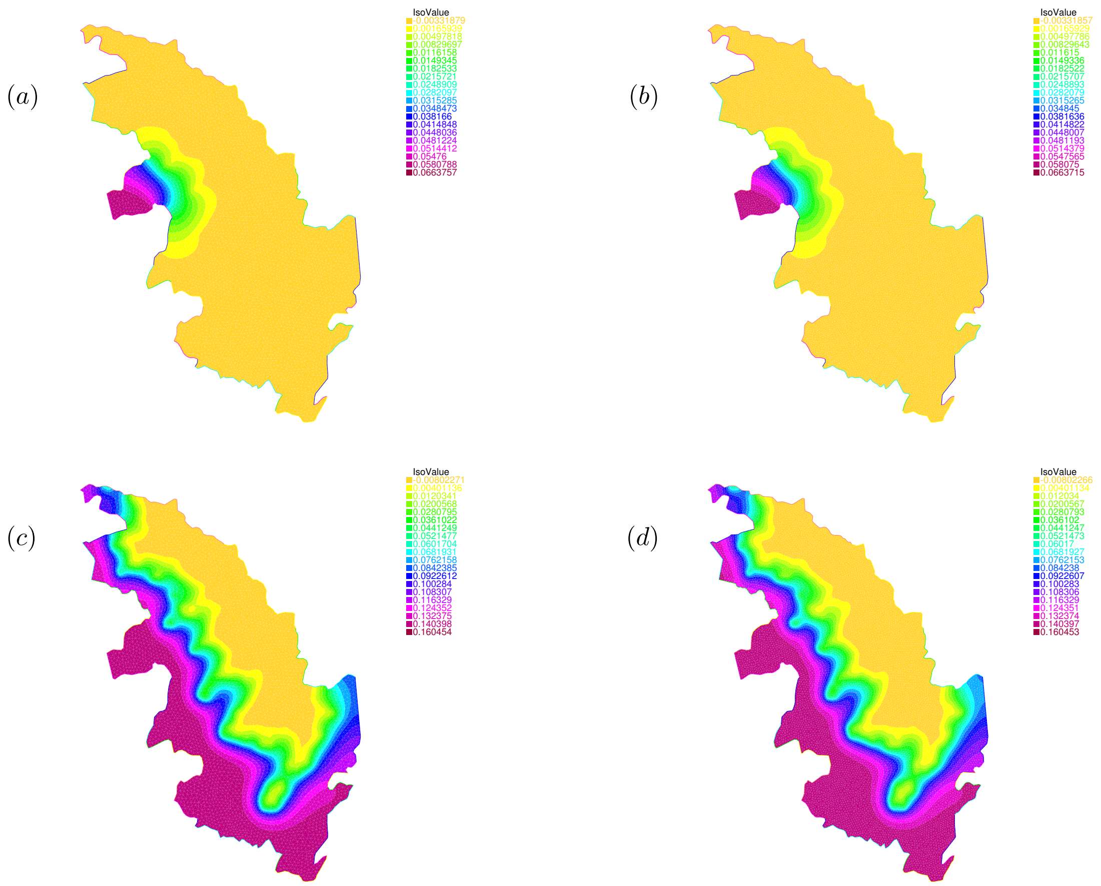


Figure S2: Probability of infection obtained at two different times and with two different meshes for the parameter vector corresponding to the posterior maximum. Top panels: 100 months after the introduction; Bottom panels: time of the last observation; Left panels: mesh composed of 4791 nodes; Right panel: finer mesh with 10703 nodes. Average difference between (a) and (b): $3e^{-5}$; Maximal difference: 0.002. Average difference between (c) and (d): $4e^{-5}$; Maximal difference: 0.02.

S2 Local Brier Score

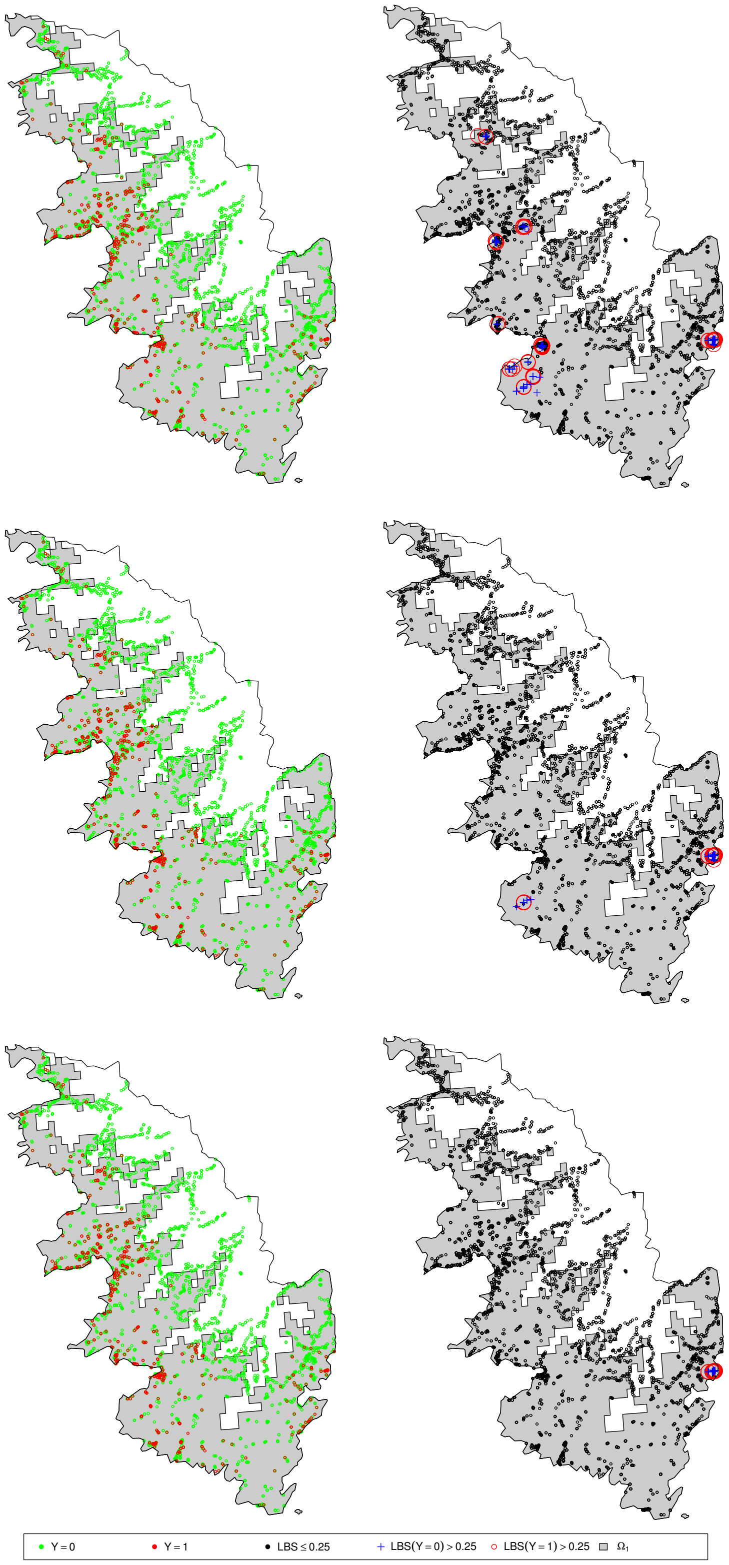


Figure S3: Locations where the LBS given in section 3.5.1 (main text) is larger than 0.25 with $k = 50$ (top), $k = 100$ (center), $k = 150$ (bottom). The gray surface gives the extent of Ω_1.

S3 Maximum likelihood estimation

Estimation of the parameter vector $\Theta = (D, b, K, \alpha, \tau_0, \tilde{\mathbf{x}}_{\mathbf{0}}, r_0, p_0)$ was also performed in the frequentist setting via maximum likelihood estimation. The maximization of the likelihood was made with the function `fmincon` of `Matlab R2015b`. This function searches for the optimum of a constrained nonlinear multivariable function using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. To take into account the risk of finding a local optimum, we carried out the maximization of the likelihood for 240 different initial values of the parameter vector drawn from the prior distribution. Figure S4 shows the evolution of the log-likelihood function from the initial values to the optimal values for the 240 calls of `fmincon`. We clearly see that most of the calls led to a relatively high likelihood (despite a few exceptions), but none of them led to a higher value than the highest value obtained with AMIS, which is not designed as an optimizer but is designed as a sampler in the posterior distribution (the highest log-likelihood value obtained with `fmincon` is -2467.8, whereas it is -2449.9 with AMIS). From a computational perspective, the maximum likelihood approach required 273 likelihood evaluations in average (i.e., 6.5×10^4 evaluations for the 240 optimizations), whereas we made 50×10^4 likelihood evaluations in AMIS. Thus, maximum likelihood estimation is less demanding, but is apparently stuck in local optimum with our model and data, and an additional task would be required to assess uncertainty in parameter estimation (e.g., via the computation of the Fisher information matrix), whereas AMIS directly provides estimation uncertainty.

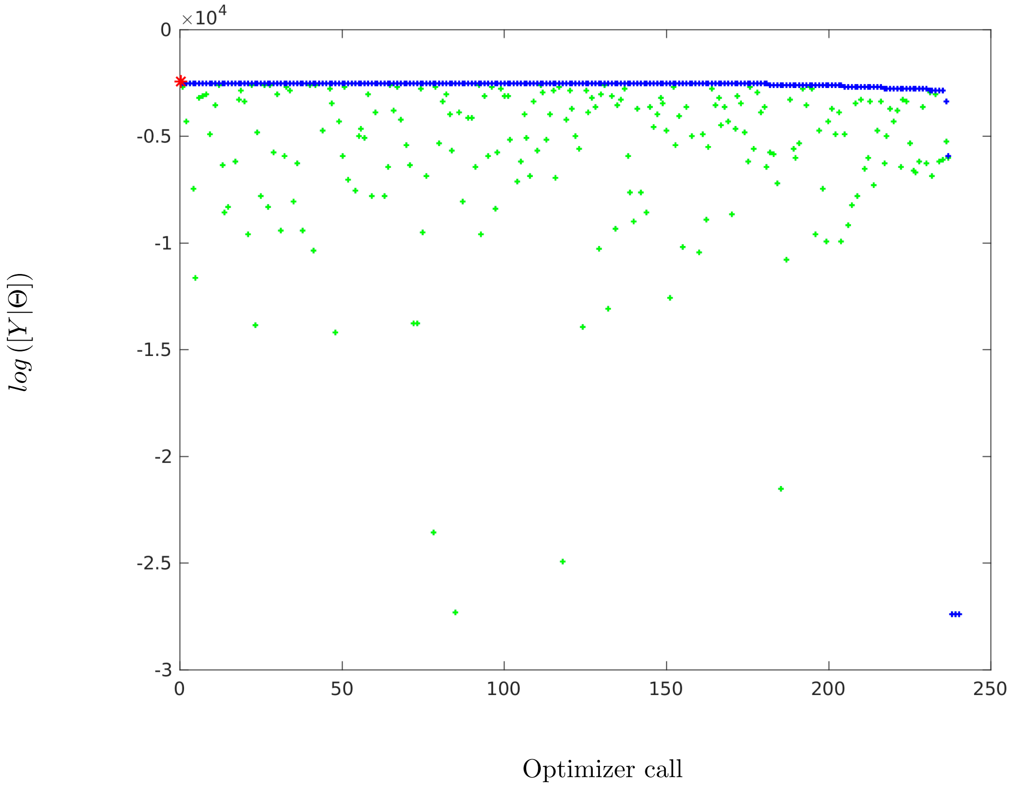


Figure S4: Values of the log-likelihood function evaluated at 240 different initial values of Θ used for the optimization (green crosses), and at the 240 resulting optimized values of Θ obtained with the fmincon function (blue crosses). The red asterisk indicates the maximum value of the log-likelihood obtained in the AMIS procedure.

B. Appendix for Chapter 4

Electronic Supplementary Material (EMS)

Model&data-based Prediction of Pathogen Dynamics

C. Abboud[1], O. Bonnefon[1], E. Parent[2], and S. Soubeyrand[1]

[1]BioSP, INRA, 84914 Avignon, France

[2] AgroParisTech, UMR 518 Math. Info. Appli., Paris, France,

INRA, UMR 518 Math. Info. Appli., Paris, France

S1 Priors

The prior distributions $[\Theta_{i\tilde{T}}|\mathcal{M}_i(\tilde{T})]$ combine vague uniform and Dirac distributions. Dirac distributions are considered for r_0 and p_0, which are related to the initial conditions. These parameters are kept fixed for identifiability issues.

$$\begin{aligned}[\Theta_{0\tilde{T}}|\mathcal{M}_0(\tilde{T})] = & \frac{1}{(10^8-50)\times 100\times 1\times 100\times 1000\times|\Omega_{\tilde{T}1}|} \\ & \times \mathbb{1}(D\in[50;10^8], b\in[0;100], K\in]0;1], \alpha\in[0;100], \tau_0\in[-1000;0], \tilde{\mathbf{x}}_0\in\Omega_{\tilde{T}1}) \\ & \times \mathrm{Dirac}_{5000}(r_0)\times\mathrm{Dirac}_{0.1}(p_0).\end{aligned}$$

$$\begin{aligned}[\Theta_{1\tilde{T}}|\mathcal{M}_1(\tilde{T})] = & \frac{1}{(10^8-100)\times(10^4-10)\times 100\times 1\times 100\times 1000\times|\Omega_{\tilde{T}1}|} \\ & \times \mathbb{1}(D_{1\tilde{T}1}\in[100;10^8], D_{1\tilde{T}2}\in[10;10^4], b\in[0;100], K\in]0;1], \alpha\in[0;100], \tau_0\in[-1000;0], \tilde{\mathbf{x}}_0\in\Omega_{\tilde{T}1}) \\ & \times \mathrm{Dirac}_{5000}(r_0)\times\mathrm{Dirac}_{0.1}(p_0).\end{aligned}$$

$$\begin{aligned}[\Theta_{2\tilde{T}}|\mathcal{M}_2(\tilde{T})] = & \frac{1}{(10^8-50)\times(10^5-50)\times 100\times 1\times 1000\times|\Omega_{\tilde{T}1}|} \\ & \times \mathbb{1}(D_{2\tilde{T}1}\in[50;10^8], D_{2\tilde{T}2}\in[50;10^5], b\in[0;100], K\in]0;1], \tau_0\in[-1000;0], \tilde{\mathbf{x}}_0\in\Omega_{\tilde{T}1}) \\ & \times \mathrm{Dirac}_{5000}(r_0)\times\mathrm{Dirac}_{0.1}(p_0).\end{aligned}$$

C. Appendix for Chapter 5

The following sections themselves are not incorporated in my Thesis because they were written by my co-autors, S. Soubeyrand and R. Senoussi, in the book chapter "Abboud et al. 2018".

7.3. Metapopulation Epidemic Model

7.3.1. *Spatially Realistic Levins Model*

In ecology, the class of Stochastic Patch Occupancy Models (SPOM) has been developed to characterize and infer the dynamics of metapopulations. A metapopulation is a set of spatially separated populations of the same species which interact via between-population migrations of individuals. Among this class of models, the spatially realistic Levins model (SRLM) is a major reference [OVA 04].

Consider a set of n circular habitat patches with areas $a_i > 0$ and centers $x_i \in \mathbb{R}^2$, $i \in \mathcal{I} = \{1, \dots, n\}$. Let $d_{i,j}$ denote the Euclidean distance between x_i and x_j. The binary variable $Y_i(t) \in \{0, 1\}$ gives the occupation status of patch i at time $t \in \mathbb{R}$: $Y_i(t) = 1$ if patch i is occupied by the species of interest at t, $Y_i(t) = 0$ otherwise. The random vector $\mathbf{Y}(t) = \{Y_1(t), \dots, Y_n(t)\}$ follows a binary-state continuous-time Markov process with inhomogeneous transition rates. Local extinctions independently occur with a constant rate e_i, which is typically proportional to the patch area a_i:

$$\mathbb{P}(Y_i(t + dt) = 0 \mid Y_i(t) = 1) = e_i dt.$$

Colonizations of unoccupied patches occur with a time-varying rate depending on the occupation status of the other patches and their distance with respect to the focal patch:

$$\mathbb{P}(Y_i(t + dt) = 1 \mid Y_i(t) = 0) = \sum_{\substack{j=1 \\ j \neq i}}^{n} c_{ij} Y_j(t) dt,$$

where c_{ij} is typically a function of the distance d_{ij} and other patch characteristics such as the areas a_i and a_j. In general, the larger d_{ij}, the lower c_{ij} (source patches send more migrants to close patches than to further patches), and the larger a_i and a_j, the larger c_{ij} (large patches send more migrants and have a higher propensity to receive migrants).

7.3.2. *A Colonization Piecewise-Deterministic Markov Process*

Here, we are interested in a pathogen metapopulation. Thus, in what follows, we adopt the vocabulary of epidemiology. In particular, thereafter, a patch is a set of hosts for the pathogen of interest, an occupied patch is a patch that is infected by a pathogen population, and an unoccupied patch is said to be healthy.

This section presents the metapopulation model proposed in [SOU 09a], which differs from the Levins model mainly because (i) extinctions and colonizations occur on distinct periods, (ii) the binary occupation status $Y_i(t)$ is augmented by a time-varying quantitative variable providing the size of the pathogen population within patch i, and (iii) observation variables are explicitly introduced in the model. To simplify the presentation of the model, we focus on the metapopulation dynamic 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.

In the following, *infection times* T_i $(i \in \mathcal{I})$ denote the times of initiation of local epidemics in the year under consideration; let $\mathbf{T} = \{T_i : i \in \mathcal{I}\}$. As a local epidemic can only occur during the growing season, $T_i \geq 0$. We assume that the pathogen survived in patch 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 a *colonization time*. By convention, we set $T_i \geq 1$ if patch i is still healthy at time $t = 1$.

Observation variables

Â The metapopulation dynamic is observed at the patch level at times $t = -1$ and $t = 1$, i.e. the end of successive years. Given that sampling is not complete (there are some patches whose health statuses are not observed) and that infections are not always detected, we introduce the observation variables $Y_{i,t}^{\text{obs}}$, $i \in \mathcal{I} = \{1, \ldots, n\}$ and $t \in \{0, 1\}$:

$$Y_{i,t}^{\text{obs}} = \begin{cases} 0 & \text{if the meadow is observed as healthy} \\ 1 & \text{if the meadow is observed infected} \\ \texttt{NA} & \text{if the meadow is not sampled.} \end{cases}$$

There are no false-positives (i.e. healthy patches observed as infected). In addition, vectors of explanatory variables are observed at the patch level, namely the patch coordinates x_i, the area a_i covered by the patch and $\{B_i, C_i, D_i\}$ that will arise in the model as regressors.

In the model, the response variables are the observations $\mathbf{Y}_1^{\text{obs}} = \{Y_{i,1}^{\text{obs}} : i \in \mathcal{I}\}$ at time $t = 1$, and we work conditionally on past observations $\mathbf{Y}_{-1}^{\text{obs}} = \{Y_{i,-1}^{\text{obs}} : i \in \mathcal{I}\}$ and covariates $\{x_i, a_i, B_i, C_i, D_i : i \in \mathcal{I}\}$. The observed final health statuses $Y_{i,1}^{\text{obs}}$ are assumed to be independently drawn from $\{0, 1, \texttt{NA}\}$ with unequal probabilities, given actual final health statuses:

$$Y_{i,1}^{\text{obs}} \mid Y_i(1) \sim \alpha_1 \text{Dirac}(0) + \alpha_2 \text{Dirac}(1) + (1 - \alpha_1 - \alpha_2) \text{Dirac}(\texttt{NA}),$$

where α_1 and α_2 account for misclassification and incompleteness in the observation process at $t = 1$ and satisfy:

$$\begin{aligned}
\alpha_1 &= r_1 \frac{p_1}{p_1 + q_1(1 - p_1)} \\
\alpha_2 &= r_1 \left(1 - \frac{p_1}{p_1 + q_1(1 - p_1)}\right) \\
p_1 &= \mathbb{P}(Y_{i1}^{\text{obs}} = 1 \mid Y_{i1}^{\text{obs}} \neq \texttt{NA}) \\
q_1 &= \mathbb{P}(Y_{i,1} = 1 \mid Y_{i1}^{\text{obs}} = 0) \\
r_1 &= \mathbb{P}(Y_{i1}^{\text{obs}} \neq \texttt{NA}).
\end{aligned}$$

Probabilities p_1, q_1 and r_1 are *observation parameters* whose values are assessed before fitting the model to data and plugged in the model.

Extinctions

Extinctions of the pathogen in infected patches can only occur during the dormancy period $[-1, 0)$. Times of extinction are not explicitly introduced into the model. We simply assume that extinctions between times -1 and 0 are, conditionally on observations $Y_{i,-1}^{\text{obs}}$, the result of independent Bernoulli draws for the infection statuses $Y_i(0)$ of patches:

$$\begin{aligned}
Y_i(0) \mid Y_{i,-1}^{\text{obs}} &\sim \text{Bernoulli}(b_i s(Y_{i,-1}^{\text{obs}})) \\
b_i &= \text{logit}^{-1}(B_i^T \beta) \\
s(Y_{i,-1}^{\text{obs}}) &= \begin{cases} 1 & \text{if } Y_{i,-1}^{\text{obs}} = 1 \\ q_{-1} & \text{if } Y_{i,-1}^{\text{obs}} = 0 \\ p_{-1} + q_{-1}(1 - p_{-1}) & \text{if } Y_{i,-1}^{\text{obs}} = \texttt{NA}, \end{cases}
\end{aligned} \qquad [7.7]$$

where b_i gives the conditional probability of pathogen survival given that patch i was infected in the beginning of dormancy, and s deals with misclassification and incompleteness of the observation process at time $t = -1$. b_i is a function of observed covariates B_i and a vector of parameters β (B_i^T is the transpose of B_i), $p_{-1} = \mathbb{P}(Y_{i,-1}^{\text{obs}} = 1 \mid Y_{i,-1}^{\text{obs}} \neq \texttt{NA})$ and $q_{-1} = \mathbb{P}(Y_{i,-1} = 1 \mid Y_{i,-1}^{\text{obs}} = 0)$. Probabilities p_{-1} and q_{-1} are *observation parameters* whose values are assessed before fitting the model to data and plugged in the model. By convention, $Y_i(0) = 1$ if and only if $T_i = 0$.

Colonizations

Healthy patches are immune during the dormancy and susceptible within the growing season. Infected patches are infectious only during the growing season. The degrees of susceptibility and infectiousness depend on explanatory variables and time as described below. In addition, already infected patches cannot be over-infected during the growing season.

The spread of the pathogen during the growing season is modeled as a spatio-temporal piecewise-Poisson point process [ILL 08] . 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 healthy patch with a standard degree of susceptibility. Thus, each point stands for a potential colonization event.

The point process is governed by an intensity function $\tilde{\lambda}$ quantifying the risk of infection at each space-time location, this risk being generated by the already infected patches. Therefore, $\tilde{\lambda}$ varies in time and space with the number, the spatial locations and the infectiousness of these patches. The expression of $\tilde{\lambda}$ at time t and location x is given by:

$$\tilde{\lambda}(t, x) = \sum_{j \in \mathcal{I}_t} c_j g_j(t - T_j) h(x, x_j), \qquad [7.8]$$

where $\mathcal{I}_t = \{j \in \mathcal{I} : T_j < t\}$ is the set of patches infected before time t; c_j encodes characteristics of patch 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 deterministic standardized disease progress function, which gives the shape of the pathogen growth within patch j; h is a dispersal function, which models pathogen dispersal as a function of the source location x_j and the location of the receiving patch x. The product $c_j g_j(t - T_j)$ specifies the degree of infectiousness of patch j at time t. In the beginning of the growing season, just after time zero, $\tilde{\lambda}$ is generated only by those patches in which the pathogen survived during the dormancy.

The standardized disease progress function is specified with a thresholded quadratic form:

$$g_j(t) = \min\{t^2, \omega a_j\} \mathbb{1}(t \geq 0), \qquad [7.9]$$

where ω is a positive parameter. The threshold ωa_j takes into account possible saturation effects, which are assumed to be proportional to the patch area a_j.

The dispersal function h is specified as an anisotropic exponential dispersal function parameterized by $\eta = (\eta_1, \ldots, \eta_5)$ [SOU 07]:

$$h(x, x') = \frac{h_1\{\phi(x - x')\}}{h_2\{\phi(x - x')\}^2} \exp\left(-\frac{||x - x'||}{h_2\{\phi(x - x')\}}\right),$$

where $\phi(x - x')$ is the angle made by the vector $x - x'$, $||x - x'||$ is the distance between x and x', $h_1(\phi)$ gives the probability that a spore is dispersed in direction ϕ, and $h_2(\phi)$

gives the expected distance travelled by a spore dispersed in direction ϕ. The angular function h_1 is assumed to be a von Mises density function [FIS 95] parameterized by a mean direction parameter $\eta_1 \in \mathbb{R}$ and a dispersion parameter $\eta_2 > 0$:

$$h_1(\phi) = \{2\pi I_0(\eta_2)\}^{-1} \exp\{\eta_2 \cos(\phi - \eta_1)\},$$

with $I_0(u) = (2\pi)^{-1} \int_0^{2\pi} \exp\{u \cos(\phi)\} d\phi$. The angular function h_2 is assumed to be proportional to a von Mises density function parameterized by a mean direction parameter $\eta_3 \in \mathbb{R}$, a dispersion parameter $\eta_4 > 0$

$$h_2(\phi) = \eta_5 \{2\pi I_0(\eta_4)\}^{-1} \exp\{\eta_4 \cos(\phi - \eta_3)\},$$

where $\eta_5 > 0$ is the constant of proportionality.

A healthy patch i is colonized during the growing season if a point of the piecewise 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 $a_i \tilde{\lambda}(t, x_i)$; a_i is considered as the effective capture area of patch i and $x \mapsto \tilde{\lambda}(t, x)$ is assumed to be approximately constant over patch i. 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 are jointly modeled as a function of explanatory variables: $c_j d_i = \exp(C_j^T \gamma + D_i^T \delta)$, where C_j and D_i are vectors of covariates, and γ and δ are vectors of parameters.

PDMP formulation of the colonization dynamic

Let $\mathbf{X}_t \in \mathcal{X}$, $t \in [0, 1]$, be the $[2 \times n]$ matrix satisfying:

$$\mathbf{X}_t = \begin{pmatrix} X_{11}(t) & \cdots & X_{1n}(t) \\ X_{21}(t) & \cdots & X_{2n}(t) \end{pmatrix} = \begin{pmatrix} c_1 g_1(t - T_1) & \cdots & c_n g_n(t - T_n) \\ Y_1(t) & \cdots & Y_n(t) \end{pmatrix},$$

where each column provides, for a given patch, the size of the pathogen population at time t and the health status of the patch at time t (remind that $Y_i(t) = \mathbb{1}(t \geq T_i)$).

We introduce the function $\Phi = (\Phi_1, \ldots, \Phi_n) : \mathcal{X} \times \mathbb{R}_+ \to \mathcal{X}$ whose j-th component satisfies:

$$\Phi_j(\mathbf{x}, t) = \begin{cases} \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } \mathbf{x}_{2j} = 0 \\ \begin{pmatrix} c_j \min\{(t + \sqrt{\mathbf{x}_{1j}/c_j})^2, \omega a_j\} \\ 1 \end{pmatrix} & \text{if } \mathbf{x}_{2j} = 1. \end{cases} \qquad [7.10]$$

Let T_i and $T_{i'}$ be two successive colonization times (i.e. $0 < T_i < T_{i'}$ and no colonization occurred in the time interval $(T_i, T_{i'})$), called *jump times* in the PDMP framework. The inter-jump duration $S_{i'} = T_{i'} - T_i$ has a survival function detailed in Equation [7.16] that takes an exponential form depending on the multivariate *jump rate* $\lambda : \mathcal{X} \mapsto \mathbb{R}_+^n$:

$$\lambda(\mathbf{X}_t) = \begin{pmatrix} d_1 a_1 \tilde{\lambda}(t, x_1)(1 - Y_1(t)) \\ \vdots \\ d_n a_n \tilde{\lambda}(t, x_n)(1 - Y_n(t)) \end{pmatrix},$$

where $\tilde{\lambda}$ was defined in Equation [7.8] and can be expressed as a function of $\mathbf{X}_t$, and the variables $Y_1(t), \dots, Y_n(t)$ are the components of the 2nd row of $\mathbf{X}_t$.

Using Equations [7.9] and [7.10], $\mathbf{X}_t$ is a PDMP with *flow function* Φ:

$$\mathbf{X}_t = \begin{cases} \Phi(\mathbf{X}_{T_i}, t) & \text{if } t \in [T_i, T_{i'}) \\ \mathbf{U}_{i'} & \text{if } t = T_{i'}, \end{cases}$$

where $\mathbf{U}_{i'}$ is drawn from the *jump kernel* $Q_{i'}(\Phi(\mathbf{X}_{T_i}, S_{i'}), \cdot)$. In the simplest case (the one which is considered thereafter), the jump kernel is a Dirac distribution, which changes only the health status $X_{2i'}(t) = Y_{i'}(t)$ of i' from healthy to infected:

$$\mathbf{U}_{i'} = \Phi(\mathbf{X}_{T_i}, S_{i'}) + \begin{pmatrix} \mathbf{0}_n \\ \mathbf{1}_n(i') \end{pmatrix},$$

where $\mathbf{0}_n$ is the raw vector with n zeros and $\mathbf{1}_n(i')$ is the raw vector whose i'-th element is equal to 1 and the $n-1$ other elements are equal to 0. This form could be generalized by drawing a random value for the size of the pathogen population $X_{1i'}(t)$ in patch i' when this patch is colonized:

$$\mathbf{U}_{i'} = \Phi(\mathbf{X}_{T_i}, S_{i'}) + \begin{pmatrix} \min\{U_{i'}, \omega a_{i'} c_{i'}\}\mathbf{1}_n(i') \\ \mathbf{1}_n(i') \end{pmatrix},$$

where the real variable $U_{i'}$ should be randomly drawn in $\mathbb{R}_+$. As mentioned above, we use thereafter the simplest case:

$$\mathbf{X}_t = \begin{cases} \Phi(\mathbf{X}_{T_i}, t) & \text{if } t \in [T_i, T_{i'}) \\ \Phi(\mathbf{X}_{T_i}, S_{i'}) + \begin{pmatrix} \mathbf{0}_n \\ \mathbf{1}_n(i') \end{pmatrix} & \text{if } t = T_{i'}. \end{cases}$$

7.3.3. *Bayesian Inference Approach*

We aim to infer infection times $\mathbf{T}$ and parameters $\mathbf{\Theta} = (\omega, \eta, \beta, \gamma, \delta)$ given observed health statuses $\mathbf{Y}_{i,-1}^{\text{obs}}$ and $\mathbf{Y}_{i,1}^{\text{obs}}$, covariates $\mathbf{Z} = \{x_i, a_i, B_i, C_i, D_i : i \in \mathcal{I}\}$ and observation parameters $\kappa_{-1} = (p_{-1}, q_{-1})$ and $\kappa_1 = (p_1, q_1)$ (we will see below, in *Remark 1*, that the observation parameter r_1 can be removed from the model in the inference stage). The inference is made by using the probability distribution $P(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z})$, which can be written as follows:

$$P(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z}) = \int_{\mathbf{T}} P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T}) dP_{\mathbf{\Theta},\kappa_{-1}}(\mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z}). \qquad [7.11]$$

Equation [7.11] highlights the hierarchical structure of the model. In the first stage, the term $P_{\mathbf{\Theta},\kappa_{-1}}(\mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z})$ gives the distribution of infection times given the observed initial statuses and covariates. This term incorporates the survival process during dormancy and the colonization PDMP parameterized by $\mathbf{\Theta}$, and the observation process at time $t = -1$ parameterized by κ_{-1}. In the second stage, the term $P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T})$ gives the distribution of the observed final statuses given infection times. This term corresponds to the observation process at time $t = 1$ parameterized by κ_1. Note that when $\mathbf{T}$ is known, $\mathbf{Y}_{-1}^{\text{obs}}$ and $\mathbf{Z}$ bring no information on $\mathbf{Y}_1^{\text{obs}}$, i.e. $P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T}, \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z}) = P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T})$.

Equation [7.11] can be used to infer the unknowns $\mathbf{T}$ and $\mathbf{\Theta}$. However, the integral at the right-hand-side cannot be calculated analytically. To overcome this difficulty, the infection times $\mathbf{T}$ can be considered as latent variables, whose distribution is specified by $P_{\mathbf{\Theta},\kappa_{-1}}(\mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z})$, and inference can be carried out with a Markov chain Monte Carlo (MCMC) method in the Bayesian context [ROB 99] or a Monte Carlo expectation maximization method in the frequentist context [WEI 90].

In this study, we chose the Bayesian approach and we applied MCMC using a Metropolis-Hastings algorithm to draw a sample from the posterior distribution of the parameters and the infection times. The posterior distribution, up to a normalizing constant, can be written as

$$P_{\kappa_{-1},\kappa_1}(\mathbf{\Theta}, \mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Y}_1^{\text{obs}}, \mathbf{Z}) \propto P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}}|\mathbf{T}) P_{\mathbf{\Theta},\kappa_{-1}}(\mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Z}) \pi(\mathbf{\Theta}), [7.12]$$

where π is the prior distribution of $\mathbf{\Theta}$ and the symbol '$\propto$' means 'proportional to'. The following paragraphs provide the expressions of the terms appearing in Equation [7.12].

Expression of $P(\mathbf{T} \mid \mathbf{Y}_{-1}^{obs}, \mathbf{Z})$

Here, we give the expression of the conditional probability of any space-time configuration $\mathbf{T}$, describing what patches are infected at what times, given the observed initial health statuses $\mathbf{Y}_{-1}^{\text{obs}}$ and covariates $\mathbf{Z}$. Thereafter, for the sake of convenience, we omit the conditioning covariates and the conditioning parameters in the notation.

We make the three following assumptions in addition to those made above. First, the infection potential is constant within each patch. Second, the degree of susceptibility of a healthy patch at time zero is independent of the initial health status at time $t = -1$. Third, points of the Poisson point process located in susceptible patches independently succeed in initiating local epidemics. The success of a point in initiating a local epidemic is patch dependent. It is measured by the success probability d_i which reflects the degree of susceptibility of i and encodes individual characteristics such as local climatic conditions.

Let $t_1, \dots, t_n$ be times in $[0,1]$ and $\mathcal{I}_A = \{i \in \mathcal{I} : t_i = 0\}$, $\mathcal{I}_B = \{i \in \mathcal{I} : 0 < t_i < 1\}$ and $\mathcal{I}_C = \{i \in \mathcal{I} : t_i = 1\}$. $\mathcal{I}_A$, $\mathcal{I}_B$ and $\mathcal{I}_C$ are associated, respectively, with the sets of patches where the pathogen survived during the dormancy, which were colonized during the season and which remained healthy. We show below that:

$$\begin{aligned} &P(\{T_i = 0 : i \in \mathcal{I}_A\}, \{T_i = t_i : i \in \mathcal{I}_B\}, \{T_i \geq 1 : i \in \mathcal{I}_C\} \mid \mathbf{Y}^{\text{obs}}_{-1}) \\ &\qquad = \prod_{i \in \mathcal{I}_A} b_i s(Y^{\text{obs}}_{i,-1}) \prod_{i \in \mathcal{I}_B} \{1 - b_i s(Y^{\text{obs}}_{i,-1})\} e^{-d_i a_i \tilde{\Lambda}(t_i, x_i)} d_i a_i \tilde{\lambda}(t_i, x_i) \\ &\qquad \times \prod_{i \in \mathcal{I}_C} \{1 - b_i s(Y^{\text{obs}}_{i,-1})\} e^{-d_i a_i \tilde{\Lambda}(1, x_i)}, \end{aligned} \qquad [7.13]$$

where $\tilde{\Lambda}(t,x) = \int_0^t \tilde{\lambda}(s,x)ds$ is the time-cumulated infection risk affecting location x. Quantities d_i and c_j are only contained in $d_i a_i \tilde{\lambda}(t_i, x_i)$ and $d_i a_i \tilde{\Lambda}(t_i, x_i)$ as the product form $d_i c_j$. This product was directly modeled (instead of separately modeling d_i and c_j) to avoid identifiability difficulties in parameter estimation.

In Equation [7.13], the term $b_i s(Y^{\text{obs}}_{i,-1})$ is the probability of pathogen survival in i during the dormancy. In the second product of [7.13], the term $1 - b_i s(Y^{\text{obs}}_{i,-1})$ is the probability of pathogen extinction in i during the dormancy. The term $e^{-d_i a_i \tilde{\Lambda}(t_i, x_i)} d_i a_i \tilde{\lambda}(t_i, x_i)$ is the probability that i remained susceptible during $[0, t_i)$ and was infected at t_i. The product $d_i a_i \tilde{\lambda}(t, x_i)$ of the degree of susceptibility d_i, the capture area a_i, and the infection risk $\tilde{\lambda}(t, x_i)$ measures the instantaneous risk of infection of patch i at time t. Finally, in the third product of [7.13], $1 - b_i s(Y^{\text{obs}}_{i,-1})$ is the probability of pathogen extinction in i during the dormancy and $e^{-d_i a_i \tilde{\Lambda}(1, x_i)}$ is the probability that i remained healthy during the epidemic period $[0,1]$.

Proof of Equation [7.13]

Let $\tau_0, \dots, \tau_{n+1}$ be $n+2$ ordered times in $[0,1]$ satisfying

$$0 = \tau_0 = \cdots = \tau_q < \cdots < \tau_r = \cdots = \tau_{n+1} = 1,$$

and $\mathcal{I}^* = \{i_1, \dots, i_n\}$ be a permutation of $\mathcal{I} = \{1, \dots, n\}$. We want to determine the conditional probability that, given the observed initial statuses $\mathbf{Y}^{\text{obs}}_{-1}$ and the covariates $\mathbf{Z}$,

– patch i_k ($k \leq q$) is infected at time $\tau_k = 0$ (survival of the pathogen during dormancy),

– patch i_k ($q < k < r$) is the k-th patch to be infected and its infection time is $\tau_k \in (0, 1)$ (colonization),

– patch i_k, $k \geq r$, is still susceptible at time $\tau_k = 1$.

In other words, we want to determine

$$p(\mathcal{I}^*, \boldsymbol{\tau}; \mathbf{Y}_{-1}^{\text{obs}}) = P(\{T_{i_k} = \tau_k : k < r\}, \{T_{i_k} > \tau_k : k \geq r\} \mid \mathbf{Y}_{-1}^{\text{obs}})$$

where $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_n\}$. Note that times $\tau_{q+1}, \cdots, \tau_{r-1}$ corresponding to colonization events are mutually different and different from one under the Poisson assumption.

Let $\mathcal{A} = \{T_{i_k} = \tau_k : k \leq q\}$, $\mathcal{B} = \{T_{i_k} = \tau_k : q < k < r\}$, $\mathcal{C} = \{T_{i_k} > \tau_k : k \geq r\}$ and $\mathcal{D} = \{T_{i_k} > 0 : k > q\}$. As $\{T_{i_k} = \tau_k : k < r\} = \mathcal{A} \cap \mathcal{B}$ and the event $\mathcal{D}$ is included in $\mathcal{B} \cap \mathcal{C}$,

$$\begin{aligned} p(\mathcal{I}^*, \boldsymbol{\tau}; \mathbf{Y}_{-1}^{\text{obs}}) =& P(\mathcal{A}, \mathcal{B}, \mathcal{C} \mid \mathbf{Y}_{-1}^{\text{obs}}) \\ =& P(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D} \mid \mathbf{Y}_{-1}^{\text{obs}}) \\ =& P(\mathcal{C} \mid \mathcal{A}, \mathcal{B}, \mathcal{D}, \mathbf{Y}_{-1}^{\text{obs}}) P(\mathcal{B} \mid \mathcal{A}, \mathcal{D}, \mathbf{Y}_{-1}^{\text{obs}}) P(\mathcal{A} \mid \mathcal{D}, \mathbf{Y}_{-1}^{\text{obs}}) P(\mathcal{D} \mid \mathbf{Y}_{-1}^{\text{obs}}). \end{aligned}$$

The two last terms at the right-hand-side of the previous equation correspond to survivals and extinctions during the dormancy and can be written as

$$P(\mathcal{A} \mid \mathcal{D}, \mathbf{Y}_{-1}^{\text{obs}}) = P(\mathcal{A} \mid \mathbf{Y}_{-1}^{\text{obs}}) = \prod_{k \leq q} P(T_{i_k} = 0 \mid Y_{i_k,-1}^{\text{obs}}) = \prod_{k \leq q} b_{i_k} s(Y_{i_k,-1}^{\text{obs}}) \tag{7.14}$$

$$P(\mathcal{D} \mid \mathbf{Y}_{-1}^{\text{obs}}) = \prod_{k > q} P(T_{i_k} > 0 \mid Y_{i_k,-1}^{\text{obs}}) = \prod_{k > q} \{1 - b_{i_k} s(Y_{i_k,-1}^{\text{obs}})\}. \tag{7.15}$$

where function s, satisfying $s(y) = (q_{-1})^{\mathbb{1}(y=0)} \{p_{-1} + q_{-1}(1 - p_{-1})\}^{\mathbb{1}(y=\texttt{NA})}$, $y \in \{0, 1, \texttt{NA}\}$, comes from [7.7].

The term $P(\mathcal{B} \mid \mathcal{A}, \mathcal{D}, \mathbf{Y}_{-1})$ is the conditional probability density function of the colonization times. So, it corresponds to the pathogen spread during the season modeled using a piecewise spatio-temporal Poisson point process with intensity $\tilde{\lambda}$ (see eq. [7.8]). Assuming that the degree of susceptibility of a patch not infected at time zero is not affected by the initial health status, $P(\mathcal{B} \mid \mathcal{A}, \mathcal{D}, \mathbf{Y}^{\text{obs}}_{-1})$ can be decomposed into

$$\begin{aligned}
P(\mathcal{B} \,|\mathcal{A}, \mathcal{D}, \mathbf{Y}^{\text{obs}}_{-1}) &= \prod_{q<k<r} P(T_{i_k} = \tau_k \mid \{T_{i_j} = \tau_j : j < k\}) \\
&= \prod_{q<k<r} P(T_{i_k} = \tau_k, \{T_{i_j} > \tau_k : j > k\} \mid \{T_{i_j} = \tau_j : j < k\}) \\
&= \prod_{q<k<r} \left(- \left. \frac{\partial P(T_{i_k} > t, \{T_{i_j} > \tau_k : j > k\} \mid \{T_{i_j} = \tau_j : j < k\})}{\partial t} \right|_{t=\tau_k} \right).
\end{aligned}$$

$P(T_{i_k} > t, \{T_{i_j} > \tau_k : j > k\} \mid \{T_{i_j} = \tau_j : j < k\})$ is the probability that the k-th patch to be infected is not infected during the time interval $[\tau_{k-1}, t]$, and that the other remaining susceptible patches are not infected during the time interval $[\tau_{k-1}, \tau_k]$. Hence,

$$\begin{aligned}
P(T_{i_k} > t, &\{T_{i_j} > \tau_k : j > k\} \mid \{T_{i_j} = \tau_j : j < k\}) \\
&= P(N_{i_k}(\tau_{k-1}, t) = 0, \{N_{i_j}(\tau_{k-1}, \tau_k) = 0 : j > k\} \mid \{T_{i_j} = \tau_j : j < k\}),
\end{aligned}$$

where $N_i(t_1, t_2)$ is the number of points —of the Poisson point process— which (i) are located in the spatial surface A_i covered by patch i, (ii) are located in the time interval $[t_1, t_2]$, and (iii) are effectively efficient for initiating a local epidemic. Condition (iii) depends on the degree of susceptibility of the patch in question. We assume that the filter due to (iii) is an independent thinning operator [DIG 83, STO 95] with the probability d_i of thinning which depends on local characteristics. From the Poisson and thinning assumptions, $N_i(t_1, t_2)$ is Poisson distributed with mean value $d_i \int_{A_i} \int_{t_1}^{t_2} \tilde{\lambda}(t, x) dt dx$. Assuming that the infection risk is constant on the spatial surface A_i (with area a_i and centroid x_i) yields

$$\begin{aligned}
N_i(t_1, t_2) \mid \{T_j : j \in \mathcal{I}_{t_2}\} &\sim \text{Poisson}(d_i a_i \tilde{\Lambda}(t_1, t_2, x_i)) \\
\tilde{\Lambda}(t_1, t_2, x_i) &= \int_{t_1}^{t_2} \tilde{\lambda}(s, x_i) ds.
\end{aligned}$$

The distribution of $N_i(t_1, t_2)$ is conditional on infection times in the past of t_2 because $\tilde{\lambda}$ is a function of these times on $[t_1, t_2]$ (see eq. [7.8]). Moreover, $N_{i_j}(\tau_{k-1}, t)$ $(j \geq k)$ are independent for $t \in [\tau_{k-1}, \tau_k]$. This yields

$$\begin{aligned}
P(T_{i_k} > t, &\{T_{i_j} > \tau_k : j > k\} \mid \{T_{i_j} = \tau_j : j < k\}) \\
&= \exp\{-d_{i_k} a_{i_k} \tilde{\Lambda}(\tau_{k-1}, t, x_{i_k})\} \prod_{j>k} \exp\{-d_{i_j} a_{i_j} \tilde{\Lambda}(\tau_{k-1}, \tau_k, x_{i_j})\}.
\end{aligned} \quad [7.16]$$

It follows

$$P(\mathcal{B} \mid \mathcal{A}, \mathcal{D}, \mathbf{Y}^{\text{obs}}_{-1}) = \prod_{q<k<r} \left(d_{i_k} a_{i_k} \tilde{\lambda}(\tau_k, x_{i_k}) \prod_{j \geq k} \exp\{-d_{i_j} a_{i_j} \tilde{\Lambda}(\tau_{k-1}, \tau_k, x_{i_j})\} \right). \qquad [7.17]$$

The term $P(\mathcal{C} \mid \mathcal{A}, \mathcal{B}, \mathcal{D}, \mathbf{Y}^{\text{obs}}_{-1})$ corresponds to the patches which remain susceptible at the end of the season. Its expression was also derived using the Poisson point process. Indeed, $P(\mathcal{C} \mid \mathcal{A}, \mathcal{B}, \mathcal{D}, \mathbf{Y}^{\text{obs}}_{-1})$ is the probability that patches i_k $(k \geq r)$ remain susceptible during the time interval $[\tau_{r-1}, 1]$, i.e. after the infection of the $(r-1)$-th patch to be infected. Thus,

$$P(\mathcal{C} \mid \mathcal{A}, \mathcal{B}, \mathcal{D}, \mathbf{Y}^{\text{obs}}_{-1}) = \prod_{k \geq r} P(N_{i_k}(\tau_{r-1}, 1) = 0 \mid \{T_{i_j} = \tau_j : j < r\}) \qquad [7.18]$$

$$= \prod_{k \geq r} \exp\{-d_{i_k} a_{i_k} \tilde{\Lambda}(\tau_{r-1}, 1, x_{i_k})\} \qquad [7.19]$$

From [7.14], [7.17] and [7.18], it follows

$$\begin{aligned} p(\mathcal{I}^*, \boldsymbol{\tau}; \mathbf{Y}_{-1}) = & \prod_{k \leq q} b_{i_k} s(Y_{i_k,-1}) \\ & \times \prod_{q<k<r} \{1 - b_{i_k} s(Y_{i_k,-1})\} d_{i_k} a_{i_k} \tilde{\lambda}(\tau_k, x_{i_k}) \exp\{-d_{i_k} a_{i_k} \tilde{\Lambda}(0, \tau_k, x_{i_k})\} \\ & \times \prod_{k \geq r} \{1 - b_{i_k} s(Y_{i_k,-1})\} \exp\{-d_{i_k} a_{i_k} \tilde{\Lambda}(0, 1, x_{i_k})\} \end{aligned}$$

Expression of $P(\mathbf{Y}_1^{obs} \mid \mathbf{T})$

It is assumed that infected patches remain infected until the end of the season, i.e. if $T_i < 1$, then $Y_i(1) = 1$. Moreover, we add one assumption to those made on the observation process when survivals during the dormancy were modeled: the success in detecting an infection does not depend on the infection time.

Using material provided in the paragraph entitled *Observation variables* in Section 7.3.2, the distribution $P(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T})$ satisfies:

$$\begin{aligned} P(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T}) = & \prod_{i \in \mathcal{I}} P(Y_{i1}^{\text{obs}} \mid T_i) \\ = & \prod_{i: Y_{i1}^{\text{obs}}=1} \frac{p_1 \mathbb{1}(T_i < 1)}{p_1 + q_1(1-p_1)} \prod_{i: Y_{i1}^{\text{obs}}=0} \left(1 - \frac{p_1 \mathbb{1}(T_i < 1)}{p_1 + q_1(1-p_1)}\right) \qquad [7.20] \\ & \times (r_1)^{\sum_i \mathbb{1}(Y_{i1}^{\text{obs}} \neq \texttt{NA})} (1-r_1)^{\sum_i \mathbb{1}(Y_{i1}^{\text{obs}} = \texttt{NA})}, \end{aligned}$$

Remark 1. Assessing r_1 prior to the estimation procedure is not required since the term $(r_1)^{\sum_i \mathbb{1}(Y_{i1}^{\text{obs}} \neq \texttt{NA})}(1-r_1)^{\sum_i \mathbb{1}(Y_{i1}^{\text{obs}} = \texttt{NA})}$ in [7.20] brings no information on the dynamics and can be removed from the posterior distribution in the MCMC.

Remark 2. In [7.20], the fraction $p_1 \mathbb{1}(T_i < 1)/\{p_1 + q_1(1-p_1)\}$ is the probability that $Y_{i1}^{\text{obs}} = 1$ given the infection time T_i and given that the patch is sampled at time $t = 1$. It equals zero if $T_i \geq 1$ since a healthy patch is never observed as infected. It is less than one if $T_i < 1$ since the pathogen presence in an infected patch can be undetected.

7.3.4. *Markov Chain Monte Carlo (MCMC) Algorithm*

This section shows how to sequentially update the parameters and the infection times in the MCMC algorithm, by exploiting the decomposition properties of the posterior distribution (block updating).

The posterior distribution can be decomposed as follows. We split $\mathbf{\Theta}$ into two subsets: $\mathbf{\Theta} = (\theta_1, \theta_2)$, where θ_1 is the parameter vector used to specify the survival probabilities b_i ($i \in \mathcal{I}$), and θ_2 is the parameter vector used in the infection risk $\tilde{\lambda}$. Actually, θ_2 parameterize c_i, d_i, g and h which appear in $\tilde{\lambda}$. The posterior distribution $P_{\kappa_{-1},\kappa_1}(\mathbf{\Theta}, \mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Y}_1^{\text{obs}})$ can be decomposed into, up to a multiplicative constant,

$$P_{\kappa_{-1},\kappa_1}(\mathbf{\Theta}, \mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \mathbf{Y}_1^{\text{obs}}) \propto P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} | \mathbf{T}) Q_{\kappa_{-1}}(\mathbf{T}, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1) Q(\mathbf{T}, \theta_2) \pi_1(\theta_1) \pi_2(\theta_2) \qquad [7.21]$$

where π_1 and π_2 are the prior distributions for θ_1 and θ_2, $P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} \mid \mathbf{T})$ is given by [7.20], and

$$Q_{\kappa_{-1}}(\mathbf{T} \mid \mathbf{Y}_{-1}^{\text{obs}}, \theta_1) = \prod_{i:T_i=0} b_i s(Y_{i,-1}^{\text{obs}}) \prod_{i:T_i>0} \{1 - b_i s(Y_{i,-1}^{\text{obs}})\} \qquad [7.22]$$

$$Q(\mathbf{T} \mid \theta_2) = \prod_{i:0<T_i<1} d_i a_i \tilde{\lambda}(T_i, x_i) e^{-d_i a_i \tilde{\Lambda}(T_i, x_i)} \prod_{i:T_i \geq 1} e^{-d_i a_i \tilde{\Lambda}(1, x_i)}, \qquad [7.23]$$

are obtained from [7.13].

Let $\mathbf{T}^c$, θ_1^c and θ_2^c denote current values for the infection times and the parameters in the algorithm. Let $\mathbf{T}^*$, θ_1^* and θ_2^* be candidate values respectively drawn from the proposal distributions $q(\cdot \mid \mathbf{T}^c)$, $q(\cdot \mid \theta_1^c)$ and $q(\cdot \mid \theta_2^c)$. First, $\mathbf{T}^*$ replaces $\mathbf{T}^c$ with probability

$$\min\left\{1, \frac{P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} | \mathbf{T}^*) Q_{\kappa_{-1}}(\mathbf{T}^*, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1^c) Q(\mathbf{T}^*, \theta_2^c) q(\mathbf{T}^c \mid \mathbf{T}^*)}{P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}} | \mathbf{T}^c) Q_{\kappa_{-1}}(\mathbf{T}^c, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1^c) Q(\mathbf{T}^c, \theta_2^c) q(\mathbf{T}^* \mid \mathbf{T}^c)}\right\}.$$

No significant simplification is possible in the calculation of this acceptance probability (only the priors disappear). Second, θ_1^* replaces θ_1^c with probability

$$\min\left\{1, \frac{Q_{\kappa_{-1}}(\mathbf{T}^c, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1^*)\pi_1(\theta_1^*)q(\theta_1^c \mid \theta_1^*)}{Q_{\kappa_{-1}}(\mathbf{T}^c, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1^c)\pi_1(\theta_1^c)q(\theta_1^* \mid \theta_1^c)}\right\}.$$

Here, only the new value of [7.22] and $\pi_1(\theta_1^*)$ must be computed. Third, θ_2^* replaces θ_2^c with probability

$$\min\left\{1, \frac{Q(\mathbf{T}^c, \theta_2^*)\pi_2(\theta_2^*)q(\theta_2^c \mid \theta_2^*)}{Q(\mathbf{T}^c, \theta_2^c)\pi_2(\theta_2^c)q(\theta_2^* \mid \theta_2^c)}\right\}.$$

Here, only the new value of [7.23] and $\pi_2(\theta_2^*)$ must be computed.

If the number of infection times is large, then the proposed infection times will certainly be always rejected. To overcome this issue, one can sequentially update subsets of infectious times. For any subset $\mathcal{J}$ of $\mathcal{I}$, we can draw candidate values $\mathbf{T}_{\mathcal{J}}^* = \{T_i^* : i \in \mathcal{J}\}$ from a proposal distribution $q(\cdot \mid \mathbf{T}_{\mathcal{J}}^c)$, where $\mathbf{T}_{\mathcal{J}}^c = \{T_i^c : i \in \mathcal{J}\}$, and accept it with probability

$$\min\left\{1, \frac{P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}}|\mathbf{T}^*)Q_{\kappa_{-1}}(\mathbf{T}^*, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1^c)Q(\mathbf{T}^*, \theta_2^c)q(\mathbf{T}_{\mathcal{J}}^c \mid \mathbf{T}_{\mathcal{J}}^*)}{P_{\kappa_1}(\mathbf{Y}_1^{\text{obs}}|\mathbf{T}^c)Q_{\kappa_{-1}}(\mathbf{T}^c, \mathbf{Y}_{-1}^{\text{obs}}, \theta_1^c)Q(\mathbf{T}^c, \theta_2^c)q(\mathbf{T}_{\mathcal{J}}^* \mid \mathbf{T}_{\mathcal{J}}^c)}\right\},$$

where component i of $\mathbf{T}^*$ is T_i^* if $i \in \mathcal{J}$, and T_i^c else. Note that a similar procedure can be applied for θ_1 and θ_2 if their dimensions are extensive.

7.3.5. *Example of Results*

The inference approach presented above was applied to infer the metapopulation dynamic of the powdery mildew *Podosphaera plantaginis*, which is a fungal pathogen of the host plant *Plantago lanceolata*, in the Åland Islands. Host plants are spread in more than 4000 meadows (i.e. patches) in this archipelago. Figure 7.6 shows patches observed as infected in 2003 and 2004. Details about data, prior distributions, MCMC tuning and results can be found in [SOU 09a]. Here, we only illustrate the type of output that can be obtained, namely the posterior distributions of the infection times in 2004 of six different patches; see Figure 7.7. Each of the six distributions shows a typical pattern, from the patch that was certainly infected in the beginning of the growing season (patch 1) to the patch that certainly remained healthy until the end of the season (patch 6).

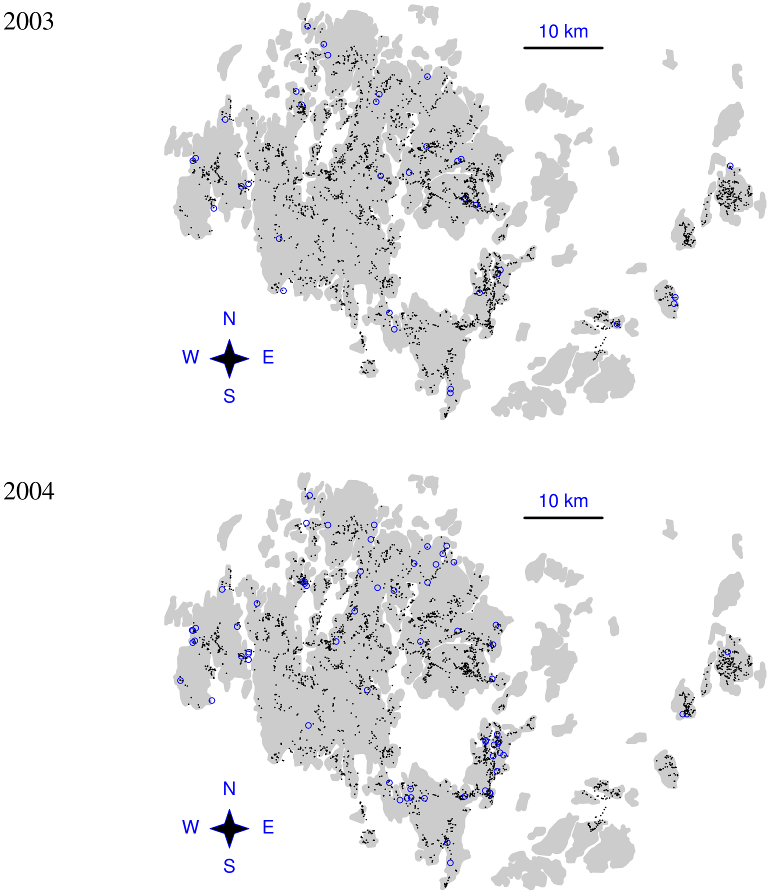


Figure 7.6: Map of the Åland Islands and patches of *Plantago lanceolata* that are healthy (dots) and infected (circles) in 2003 (top panel) and 2004 (bottom panel).

7.4. Stochastic Approaches for Modeling Spatial Trajectories

The study of animal movements informs on both individual behaviors of focal species and population-level dynamics. In particular, the characterization of territories used by individuals can be assessed via an estimation of the expected movements of animals, using discretely located data obtained at some given observation times. Many other application domains (e.g. physics of particles and transportation science) actually share the same questions regarding statistical inference of movements and trajectory reconstitution conditional on observations.

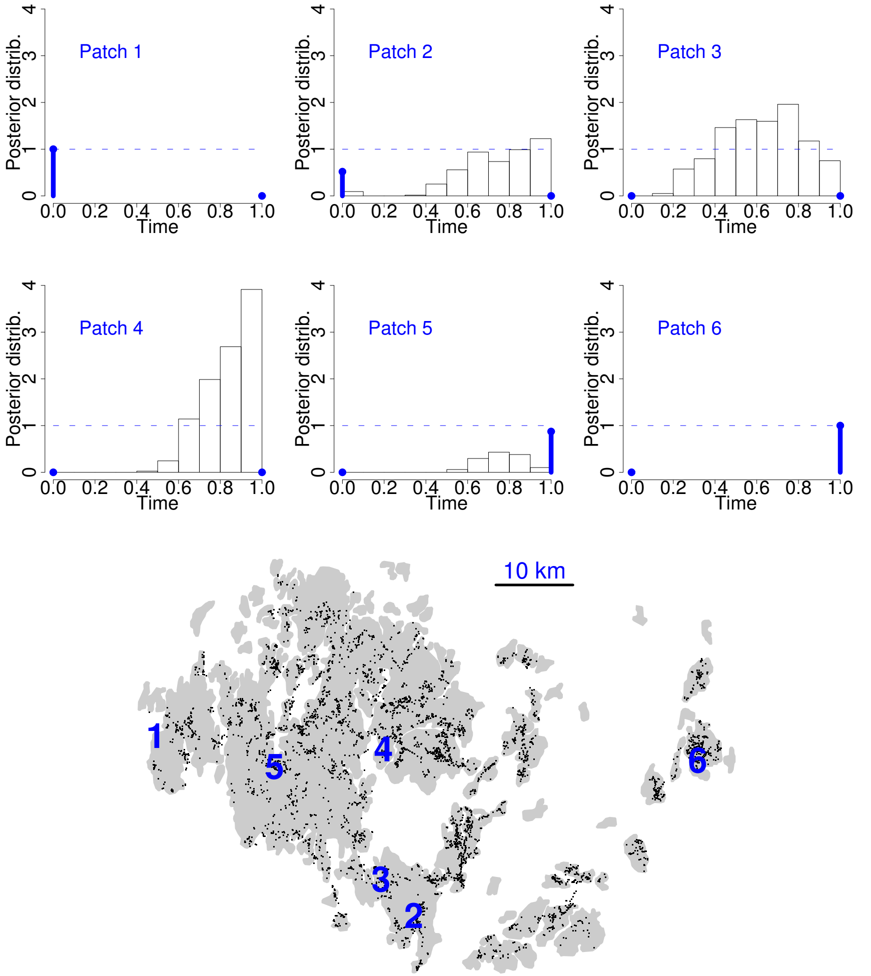


Figure 7.7: Zero-one-inflated posterior distributions of the infection times in year 2004 of six different patches (top panels). Locations of patches in the Åland Islands are indicated in the top panel. In each top panel, the dots at times zero and one give the posterior probabilities that the infection time is zero and one, respectively.

Various theoretical models for describing movements are available. Initially, continuous-time movements were often assumed to be simple Brownian motions [HOR 07], but then more and more general stochastic differential equations have been proposed [IAC 08]. Other approaches consider movements in a discrete-time context, mainly using multivariate Markov chains.In this section and in connection with the topic of the book, we will only consider time-continuous processes. From the numeric and inferential point of view, several `R` packages are available for performing statistical analyses of trajectories (e.g. `Move`, `BBMM` and `MovementAnalysis`).

In what follows, we present the simple case of interpolating punctual observations along a d-dimensional Brownian motion giving rise to the so called Brownian bridge. Then, we show how one can use the stochastic machinery, namely the martingale machinery of predictable compensation for jumps, for building models of trajectories with jumps that can be viewed as PDMPs. We illustrate our approach by exhibiting the diversity of behaviors that elementary examples may exhibit.

Notation

We will assume that the continuous index set for processes is time. Naturally, depending on the topic, one can replace the time index by any other real variable that have a pertinent meaning with respect to the underlying dynamics. Scalar elements (either constants, functions or processes) will be denoted by capital letters (e.g., X), vector elements by bold letters ($\mathbf{x}$), and matrix elements by capital bold letters ($\mathbf{X}$). Moreover, note that random functions include deterministic ones, and that the term *process* is used with a generic meaning, whereas the term *sequence* denotes only discrete-time random processes.

7.4.1. *Conditioning a Brownian Motion by Punctual Observations*

Due to the lack of relevant knowledge or because of their characteristics, movements of animals or particles in spatial domains are often modeled as realizations of Brownian processes, which are viewed as reference models for trajectories. We recall that a standard d-dimensional Brownian motion $\mathbf{w}(t)$ in $\mathbb{R}^d$ simply consists of d independent copies of one-dimensional standard Brownian motions $W_i(t)$ with $W_i(0) = 0, i = 1, \cdots, d$. $\mathbf{w}(t)$ being Gaussian, it is entirely characterized by its first order moments: $E(\mathbf{w}(t)) = 0$ and $E(\mathbf{w}(t)\mathbf{w}^T(s)) = (t \wedge s)\mathbf{I}_d$ where $\mathbf{I}_d$ stands for the d-unit matrix.

Observations of a random processes $\mathbf{x}(t)$ representing a trajectory, even when they are dense in time, always yield a sequential data set $\mathbf{y}_n = \mathbf{x}(T_n^{\mathbf{y}})$ for observation time $T_n^{\mathbf{y}}$, $n = 1, 2, \ldots$. Assuming that these observation times are independent of the process, one can infer some statistical characteristics of $\mathbf{x}(t)$ and then take into account observations to simulate (i.e. reconstruct or interpolate) the non-observed part of the trajectory. In the case of the Brownian motion, the conditioning with respect to observations gives the so-called Brownian bridge.

Brownian bridge on $\mathbb{R}^d$

The Brownian bridge $X(t),\ t \in [0,1]$, in $\mathbb{R}$ is defined (in distribution) as a Brownian motion $W(t), t \in [0,1]$, conditional on the knowledge that at $t = 1$, $W(t) = 0$. A path-wise definition exists: $X(t) = W(t) - tW(1),\ t \in [0,1]$.

This definition can be straightly extended to any interval $[T_1, T_2]$. Using the specific properties of conditional expectation for Gaussian distribution, one can easily prove that conditionally on $\{W(T_1), W(T_2)\}$, the Brownian bridge $X(t), t \in [T_1, T_2]$, is a Gaussian process, with $E(X(t)) = \frac{W(T_1)(T_2-t)+W(T_2)(t-T_1)}{T_2-T_1}$ and $E(X(t)X(s)) = \frac{(T_2-t)(s-T_1)}{T_2-T_1} = C(t,s)$, independent of $W(T_1),\ W(T_2)$, for $T_1 \leq s \leq t \leq T_2$.

In particular, $X(t)$ follows a Gaussian distribution with mean $\mu(t) = a_1 + \frac{t-T_1}{T_2-T_1}(a_2 - a_1)$ and variance $\sigma^2(t) = C(t,t)$ where $a_1 = W(T_1)$ and $a_2 = W(T_2)$.

A d-dimensional Brownian bridge $\mathbf{x}(t) = (X_1, \ldots, X_d)(t),\ t \in [T_1, T_2]$, with $\mathbf{x}(T_j) = \mathbf{a}_j = (a_{1,j}, \ldots, a_{d,j}) \in \mathbb{R}^d,\ j = 1, 2$, is defined as a vector of d independent Brownian bridges $X_i(t)$ with $X_i(T_j) = a_{i,j},\ j = 1, 2$. More explicitly, $\mathbf{x}(t)$ has a Gaussian density $\varphi(x|\mu(t), \Sigma(t))$ with mean $\mu(t) = \mathbf{a}_1 + \frac{t-T_1}{T_2-T_1}(\mathbf{a}_2 - \mathbf{a}_1)$ and covariance matrix $\Sigma(t) = \frac{(T_2-t)(s-T_1)}{T_2-T_1}\mathbf{I}_d$.

Brownian bridge with noisy extremal points

Due to measurement errors, the points $\mathbf{a}_j,\ j = 1, 2$, are generally random. If we assume these points to be independent with densities $f_{\mathbf{a}_j}, j = 1, 2$, the distribution of $\mathbf{x}(t)$ can be written:

$$P\left(\mathbf{x}(t) \in \mathbb{D}\right) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \left(\int_{\mathbb{D}} \varphi(x|\mu(t), \Sigma(t)) dx \right) f_{\mathbf{a}_1}(u) f_{\mathbf{a}_2}(v) du dv,\ \mathbb{D} \subset \mathbb{R}^d.$$

In the case of Gaussian densities $f_{\mathbf{a}_j}(u) = \prod_{i=1}^d \varphi(u_i|a_{i,j}, \sigma_j^2),\ j = 1, 2$, the process $\mathbf{x}(t)$ remains Gaussian with mean $\mu(t, \mathbf{a}_1, \mathbf{a}_2) = \mathbf{a}_1 + \frac{t-T_1}{T_2-T_1}(\mathbf{a}_2 - \mathbf{a}_1)$ whereas its covariance matrix satisfies $\Sigma^*(t) = \sigma^{*2}(t)\mathbf{I}_d$ with

$$\sigma^*(t) = \frac{(T_2-t)(t-T_1)}{T_2-T_1} + \sigma_1^2 \left(\frac{T_2-t}{T_2-T_1} \right)^2 + \sigma_2^2 \left(\frac{t-T_1}{T_2-T_1} \right)^2.$$

Mean occupation time

An important index in ecological studies is the mean occupation time of space domain $\mathbb{D}$ during a time interval $[t_1, t_2]$, which is defined as the non-negative random

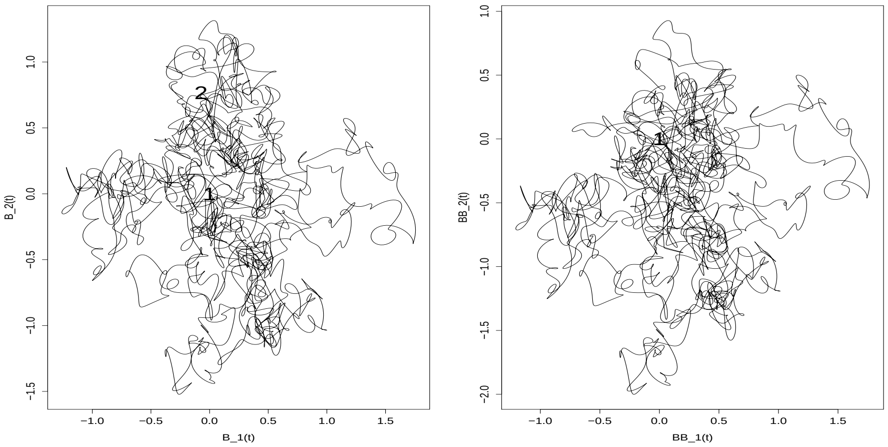


Figure 7.8: Paths in the plane of a standard Brownian motion starting at point (0,0) marked by "1" and arriving at an unconditioning point marked by "2" (left) and a standard Brownian bridge starting and arriving at point (0,0) marked by "1" (right).

variable $\tau_{\mathbb{D}} = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} \mathbf{1}_{\{\mathbf{x}(t)\in\mathbb{D}\}}dt$. Its expectation $\nu(\mathbb{D}) = E(\tau_{\mathbb{D}})$ induces an absolutely continuous measure with density:

$$h(x) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \varphi(x \mid \mu(t, \mathbf{a}_1, \mathbf{a}_2), \Sigma^*(t))dt.$$

Related statistical issues

For ecological and territory planing purposes, one can be interested in the estimation of the density $h(x)$ after collecting a set of observations $(T_j, \mathbf{x}(T_j) = \mathbf{a}_j),\ j = 1, \ldots, n+1$. Assume that these data are drawn from a d-dimensional Brownian motion with diffusion coefficient σ^2 and variances of measurement errors σ_j^2 depending on locations $\mathbf{a}_j$, and for $t \in [T_j, T_{j+1}]$, and let

$$\mu_j(t) = \mu(t, \mathbf{a}_j, \mathbf{a}_{j+1})$$

$$\sigma_j^*(t) = \sigma^2\frac{(T_{j+1}-t)(t-T_j)}{T_{j+1}-T_j} + \sigma_j^2\left(\frac{T_{j+1}-t}{T_{j+1}-T_j}\right)^2 + \sigma_{j+1}^2\left(\frac{t-T_j}{T_{j+1}-T_j}\right)^2.$$

Let the process $\mathbf{x}(t)$, $t \in [T_1, T_{n+1}]$, be formed by the set of independent Brownian bridges connecting $\mathbf{a}_j$ to $\mathbf{a}_{i+1}$ within time interval $[T_j, T_{j+1}]$, $j = 1, \ldots, n$. Then, the total mean occupation time of space has density

$$h(x) = \frac{1}{T_{n+1} - T_1} \sum_{j=1}^{n} \int_{T_j}^{T_{j+1}} \varphi(x \mid \mu_j(t), \Sigma_j^*(t))dt. \qquad [7.24]$$

The variances of measurement errors σ_j^2 are generally specified and one only has to estimate the diffusion coefficient σ^2 to compute the occupation time density.

The following trick was used to build a simple conditional likelihood for data. Assume that n is even, then one can prove that observations $\mathbf{x}(T_{2k})$, $k = 1, \ldots, n/2$, conditional on the values of observations $\mathbf{x}(T_{2k-1})$, $k = 1, \ldots, n/2$, are independent Gaussian random vectors with mean vectors $\mu_{2k-1}(T_{2k})$ and covariance matrices $\Sigma^*_{2k-1}(T_{2k})$. Hence, we can get an estimate $\hat{\sigma}^2$ by maximizing the following likelihood:

$$\prod_{k=1}^{n/2} \varphi(\mathbf{a}_{2k} \mid \mu_{2k-1}(T_{2k}), \Sigma^*_{2k-1}(T_{2k})).$$

The estimation of density h can be performed with standard numerical methods approximating the integral form in Equation [7.24]. This approach was compared to kernel methods considering observed locations as i.i.d. random vectors drawn from h, and was proven to be much more efficient since it accounts for measurement errors and temporal dependencies between observed locations. Moreover, the first approach yields more realistic domains for level sets of h.

Extension to further movement dynamics

Beyond the Brownian bridge, there exists today a wide range of literature about more general (and more realistic) diffusion bridges in $\mathbb{R}^1$ and $\mathbb{R}^d$ related to some specific stochastic differential equations of the form:

$$d\mathbf{x}(t) = f(t, \mathbf{x})dt + \sigma(t, \mathbf{x})d\mathbf{w},$$

driven by a d-dimensionnal Brownian motion $\mathbf{w}(t)$ and a vectorial drift function f.

There are many results about the characterization (in distribution as well as in a path-wise sense) of these diffusions when they are considered conditionally on their values $\mathbf{x}(T_j) = \mathbf{a}_j$ at times T_j, $j = 1, 2$. These results are however more complicated to obtain since they are grounded on sophisticated tools such as the Girsanov theorem.

7.4.2. *Movements with Jumps, Including Mathematical Preliminaries*

Thereafter, we assume that there exists a complete probability space $(\Omega, \mathbb{F}, \mathbb{P})$ with a filtration (or history) $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ such that processes are $\mathbb{F}$-adapted, stopping times refer to $\mathbb{F}$ and martingales to $(\mathbb{F}, \mathbb{P})$. We shall neither develop the classical theory of the predictable σ-algebra nor insist on other definitions such as predictable processes. One has to know that a process with everywhere càglàd paths (i.e. left continuous with right limits) are predictable. We also adopt standard notations for a process $X(t)$ such as $X(t-) = \lim_{s \uparrow t} X(s)$ and $\Delta X(t) = X(t) - X(t-)$. For càdlàg processes X, the continuous part is defined as $X^c(t) = X(t) - \sum_{s \leq t} \Delta X(s)$.

Point processes and predictable projections

In what follows, an 1D temporal point process $N(t)$ is given by a strictly increasing sequence of stopping times $(T_i)_{i \geq 0}$ with the convention that $T_0 = 0$. The associated counting process is defined as $N(t) = \Sigma_{i>0} \mathbf{1}_{\{T_i \leq t\}}$. Under this definition N is adapted. Moreover, it is assumed to be a *simple* point process in the sense that all jumps are only 1-valued.

As an adapted increasing process, N is a submartingale (i.e. $E(N(t)|\mathcal{F}_s) \geq N(s)$; for all $t \geq s$) and by the Doob-Meyer Theorem [PRO 05], there exists an increasing predictable process $\tilde{N}(t)$ such that $M(t) = N(t) - \tilde{N}(t)$ is a martingale. $\tilde{N}(t)$ is called the predictable compensator of $N(t)$. Theoretically, it is defined as a conditional expectation with respect to the predictable σ-field. In most interesting cases, $\tilde{N}(t)$ is almost surely absolutely continuous with respect to the Lebesgue measure with a random density function $\lambda(t)$, called the intensity function, that is $\tilde{N}(t) = \int_0^t \lambda(s) ds$.

When the filtration reduces to natural history of the process $N(t)$, the intensity can be deduced as follows (see [DAL 88] for details): If regular versions $G_{i+1}(dt|\mathcal{F}_{T_i})$ of the conditional distribution functions of interval lengths $D_{i+1} = T_{i+1} - T_i$ exist, then $\tilde{N}(t) = \Sigma_{i>0} \Lambda_i(t)$, with

$$\Lambda_{i+1}(t) = \begin{cases} 0 & \text{if } t \leq T_i \\ \int_0^{(t-T_i) \wedge D_{i+1}} \frac{G_{i+1}(ds|\mathcal{F}_{T_i})}{1 - G_{i+1}(s-|\mathcal{F}_{T_i})} & \text{if } t > T_i. \end{cases} \qquad [7.25]$$

If $G_i(dt|\mathcal{F}_{T_i}) \ll dt, i \geq 1$, then one can paste the different pieces into a single formula $\tilde{N}(t) = \int_0^t \lambda(s) ds$. As we shall see below, this expression is largely used in survival analysis, where $\Lambda_i(t)$ stands for the cumulative hazard function of the random variable D_i and its derivative is the hazard function.

For filtrations richer than the natural history, the calculation of the compensator is generally out of reach, but it conserves the same interpretation, namely the best cumulative predictor of the jumps of N. Authors generally assume some specific forms for the intensity process grounded on relevant hypotheses for the application domain of interest, because in many interesting cases, the form of the compensator uniquely determines the probability distribution underlying the point process $N(t)$. For example a deterministic continuous compensator refers to a Poisson processes (see an example of sample paths for N and $\tilde{N}$, and the associated compensating process M in Figure 7.9).

A useful and *universal* property, under the natural history, is that a simple point process N with continuous and a.s. unbounded compensator $\tilde{N}$ undergoing the random time change $\tilde{N}^{-1}(t)$, yields a standard homogeneous Poisson process $N^*(t) = N(\tilde{N}^{-1}(t))$. A partial converse is that a standard Poisson process $N^*(t) = \sum_{i>0} \mathbf{1}_{\{T_i^* \leq t\}}$ and a positive function $\lambda(t)$ jointly give rise to a Poisson process $N(t) = \sum_{i>0} \mathbf{1}_{\{T_i \leq t\}}$ of intensity $\lambda(t)$ with $T_i = \int_0^{T_i^*} \lambda(s)ds = \Lambda(T_i^*)$.

Generalization to multivariate and marked point processes

A d-dimensional point process $\mathbf{N}(t) = (N_1, \ldots, N_d)(t)$ is defined similarly as above by a probability space with d sequences of stopping times (T_i^j), $j = 1, \ldots, d$, $i \geq 0$, with corresponding vectorial compensator $\tilde{\mathbf{N}}(t) = (\tilde{N}_1, \ldots, \tilde{N}_d)(t)$ and martingales $M^j(t) = (N^j - \tilde{N}^j)(t)$.

However, in the context of movements with random jump sizes, we need a wider generalization, namely the marked point processes and their dual predictable projections [JAC 75]. We avoid details of the theory by simply restricting our presentation of marked point processes within the following framework. A sequence of random vectors $(T_i, \varepsilon_i)_{i\geq 0}$ taking values in $\mathbb{R}_+ \times \mathbb{R}^d$ with $T_i < T_{i+1}$ defines a random measure $N(dt, dx) = \sum_i \delta_{(T_i, \varepsilon_i)}$. A stochastic machinery similar than above can be developed to enable us to assert that there exists a predictable random measure (on an extended probability space) $\tilde{N}(dt, dx)$ such that for every predictable process $Y(s, x)$, the process $M_Y(t) = \int_0^t \int_{\mathbb{R}^d} Y(s, x)(N - \tilde{N})(dt, dx)$ is a martingale.

In the case of a filtration corresponding to the natural history, a formulation similar to [7.25] gives the predictable projection, by replacing the previous conditional probability function $G_{i+1}(dt|\mathcal{F}_{T_i})$ by the distributions $G_{i+1}(dt \times dx|\mathcal{F}_{T_i})$ of the variable $(T_{i+1} - T_i, \varepsilon_{i+1})$ conditionally on $\mathcal{F}_{T_i}$. More precisely, we have $\tilde{N}(dt, dx) = \Sigma_{i>0}\Lambda_i(dt, dx)$, with

$$\Lambda_{i+1}(dt, dx) = \frac{G_{i+1}(dt - T_i, dx|\mathcal{F}_{T_i})}{G_{i+1}([t - T_i, \infty] \times \mathbb{R}^d|\mathcal{F}_{T_i})} \mathbf{1}_{\{T_i < t \leq T_{i+1}\}}.$$

Example 1

In the case of an 1D point process, let us assume that $G_{i+1}(ds|\mathcal{F}_{T_i})$ is the Weibull distribution $W(\alpha, \beta)$, with hazard function $h(t) = \beta\alpha^\beta t^{\beta-1}$ and cumulative hazard function $H(t) = (\alpha t)^\beta$. Parameters α and β are the scale and shape characteristics. According to Equation [7.25], the compensator is written $\tilde{N}(t) = \alpha^\beta(\sum_{i=1}^n (T_i - T_{i-1})^\beta + (t - T_n)^\beta)$ for $T_n < t \le T_{n+1}$.

This compensator is stochastic since its expression depends on the stopping times T_i. Actually, N is a renewal process and it is not a Poisson process, unless $\beta = 1$ since for that case $\tilde{N}(t) = \alpha^\beta t$ is deterministic.

Example 2

We now consider the analogous Poisson process with intensity $\lambda(t) = \beta\alpha^\beta t^{\beta-1}$. Thus, for any interval $I = [\tau_1, \tau_2]$, the number of point events $N(I)$ in I is Poisson distributed with parameter $\Lambda(I) = \int_{\tau_1}^{\tau_2} \lambda(s)ds = \alpha^\beta(\tau_1^\beta - \tau_2^\beta)$. In addition, given the number of point events $N(I) = k$, its realization $\{Y_1, \dots Y_k\}$ within I, are i.i.d random variables with probability density $g(t) = \beta\frac{t^{\beta-1}}{\tau_2^\beta - \tau_1^\beta}\mathbf{1}_{\{\tau_1 \le t \le \tau_2\}}$.

For simulation purpose, note that Y_j has the same distribution as $(U(\tau_2^\beta - \tau_1^\beta) + \tau_1^\beta)$, where U is uniformly distributed over $[0, 1]$. Note also that the time transformation $\Lambda^{-1}(t)$ makes $N^*(t) = N(\Lambda^{-1}(t)$ to be a standard Poisson process. Observing that $D_{i+1}^* = T_{i+1}^* - T_i^*$ is exponentially distributed with rate 1, one can prove that the inter-event length time $D_{i+1} = T_{i+1} - T_i$, conditionally on T_i (or T_i^*), has the following survival function:

$$S_{i+1}(t) = \exp\{-\alpha^\beta((T_i + t)^\beta - T_i^\beta)\}\ ,\ t \ge 0.$$

This formula states that T_{i+1} conditionally on the event $T_{i+1} > T_i$ behaves as a Weibull distributed random variable $Y \sim W(\alpha, \beta)$, conditioned by the event $Y > T_i$; this is the memory loss property of a Poisson process. Figure 7.9 illustrates sample paths for N, $\tilde{N}$ and M for parameter values $\alpha = 1$ and $\beta = 1.2$.

Stochastic integrals for purely discontinuous martingales

In the context of point processes, stochastic integration reduces to path-wise integrals (in the sense of Stieltjes-Lebesgue integrals for bounded variation integrands), but nevertheless requires care. For sake of completeness, let us first recall that a semimartingale $X(t)$ is defined by the identity $X(t) = M(t) + A(t)$, where $M(t)$ is a local martingale and $A(t)$ is a locally bounded variation process. For any semimartingale X, one can define its quadratic variation process $[X, X](t) = X^2(t) - 2\int_0^t X(s-)dX(s)$, which is also a locally bounded variation process and satisfies $\Delta[X, X](t) = (\Delta X(t))^2$. The continuous part of $[X, X]$ is defined by $[X, X]^c(t) = [X, X](t) - \sum_{0 \le s \le t}(\Delta X(s))^2$. The quadratic co-variation process of two semimartingales is defined by duality as $[X, Y](t) = ([X + Y, X + Y] - [X, X] - [Y, Y])/2$, and similarly satisfies $\Delta[X, Y](t) = \Delta X(t)\Delta Y(t)$.

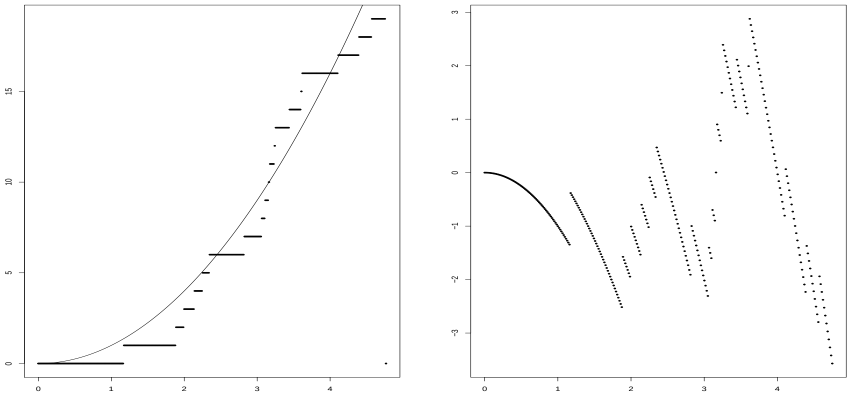


Figure 7.9: Left: Counting Poisson process $N(t)$ of Example 2 (broken line) and its compensator $\tilde{N}(t) = t^{\beta}$ (continuous line) with $\alpha = 1$ and $\beta = 1.2$. Right: The corresponding compensating martingale $M(t) = (N - \tilde{N})(t)$.

For a *simple* counting process N, we have $[N, N](t) = N(t)$. More generally $[X, X](t) = \sum_{s<0} \Delta X^2(s)$ holds for any adapted process X with locally bounded variation, so that $[X, X](t) \equiv 0$ if in addition X is continuous. In fact, the machinery of stochastic calculus intervenes only when the martingale component M has a non-purely discontinuous part (i.e. $[M^c, M^c] \neq 0$).

Point processes, compensators and martingales

We recall that if $M(t) = N(t) - \tilde{N}(t)$ denotes the martingale compensating the jumps of a simple point process N, then any adapted, integrable predictable (in particular left continuous) $f(t)$, gives rise to a martingale $M_f(t) = \int_0^t f(s)dM(s)$. These processes are also purely discontinuous martingales and their quadratic co-variation processes satisfy the following formula $[M_f, M_g](t) = \int_0^t f(s)g(s)dN(s)$. As a by-product, we see that $[M_f, M_g](t)$ is compensated by $\int_0^t f(s)g(s)d\tilde{N}(s)$, so that for $t \geq s$, we have:

$$E(M_f(t)M_g(t) \mid \mathcal{F}_s) = E\left(\int_0^t f(u)g(u)d\tilde{N}(u) \mid \mathcal{F}_s\right).$$

This formula is particularly appealing for deterministic functions f and g and/or handy expressions of the compensator $\tilde{N}(dt)$ to explicitly calculate the covariance functions. Figure 7.10 shows two examples of 2D-trajectories whose coordinates are correlated martingales defined by stochastic integrals as above.

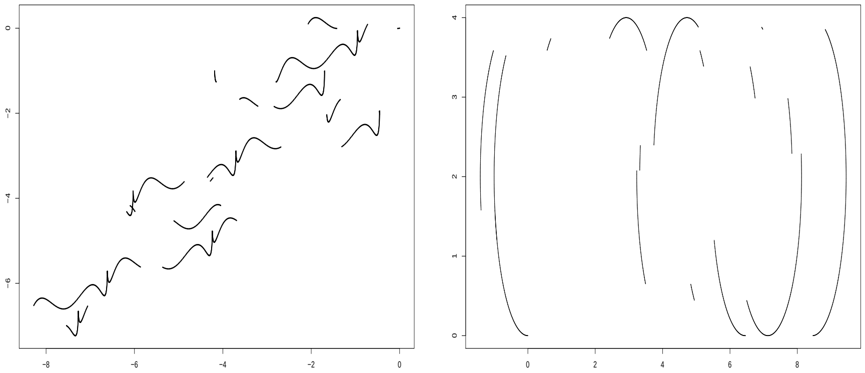


Figure 7.10: Realization of 2D martingales $X_i(t) = \int_0^t f_i(s)d(N - \tilde{N})(s)$, $i \in \{1, 2\}$, with $f_1(t) = 1 + \cos(t)$ and $f_2(t) = 0.5 - 2\sin(3t)$ (left panel), and $f_1(t) = \cos(0.3t)$ and $f_2(t) = 2\sin(0.3t)$ (right panel).

Example 3: Stochastic differential equations with impulsions

We now illustrate an other type of dynamical systems based on stochastic differential equations driven by compound point processes via a particular but nevertheless generic example for many dynamics.

We consider an autonomous system undergoing random shocks at random times. For $\mathbf{x} = (x_1, x_2)$, we consider the quadratic function $C(\mathbf{x}) = x_1^2 + \beta x_2^2$ on the plane $\mathbb{R}^2$. The level curves of C are either ellipses (when $\beta > 0$) or hyperbolas (when $\beta < 0$). This is obvious for $\beta > 0$. For $\beta < 0$, let $\beta = -\rho^2$, then the equation $C(\mathbf{x}) = c$ can be written $(x_1 - \rho x_2)(x_1 + \rho x_2) = c$, which reduces to $u_1 u_2 = c$ after a linear transformation.

Besides, using classical tricks for ordinary differential equations (ODE), one can prove that functions $\mathbf{x}(t)$ satisfying $C(\mathbf{x}(t)) = c$ are governed by the following homogeneous linear ODE:

$$\mathbf{x}'(t) = A\mathbf{x}(t), \text{ with } \mathbf{x}(0) = \mathbf{x}^* \qquad [7.26]$$

whose solution is $\mathbf{x}(t) = e^{At}\mathbf{x}^*$. More explicitly,

– if $\beta = \rho^2 > 0$, we have $A = \begin{pmatrix} 0 & -\rho^2 \\ 1 & 0 \end{pmatrix}$ and the solution of ODE [7.26] is:

$$\begin{aligned} x_1(t) &= \ x_1^* \cos(\rho t) \ \ - x_2^* \rho \sin(\rho t) \\ x_2(t) &= x_1^* \sin(\rho t)/\rho + \ x_2^* \cos(\rho t) \end{aligned}$$

– whereas for $\beta = -\rho^2 < 0$, we get $A = \begin{pmatrix} 0 & -\rho \\ -1/\rho & 0 \end{pmatrix}$, yielding the following solution of ODE [7.26]:

$$\begin{aligned} x_1(t) &= \quad x_1^*(e^t + e^{-t})/2 \quad + x_2^*\rho(-e^t + e^{-t})/2 \\ x_2(t) &= x_1^*(-e^t + e^{-t})/(2\rho) + \quad x_2^*\rho(e^t + e^{-t})/2 \end{aligned}$$

Next, let us consider the marked point process $\sum_{i>0} \delta_{(T_i, \varepsilon_i)}$ in $\mathbb{R}_+ \times \mathbb{R}^2$, and the bi-dimensional stochastic differential equation (SDE):

$$d\mathbf{z}(t) = A(\mathbf{z}(t-))dt + dM(t) \qquad [7.27]$$

where $M(t) = \sum_{i>0} \varepsilon_i \mathbf{1}_{\{T_i \leq t\}}$. The sequence $(\varepsilon_i)_{i\geq 1}$ is formed by i.i.d. elements and is independent of $(T_i)_{i\geq 1}$.

The solution of this SDE consists in a particle trajectory formed by a sequence of disjoint curve arcs, each being a solution of Eq. [7.26]: at random times T_i, the particle jumps by a size ε_i from its present orbit at a new location, initiates a new orbit, and so on.

Figure 7.11 illustrates sample paths for both ellipsoidal and hyperbolic orbits, depending on the sign of β, with standard Gaussian variables ε_i.

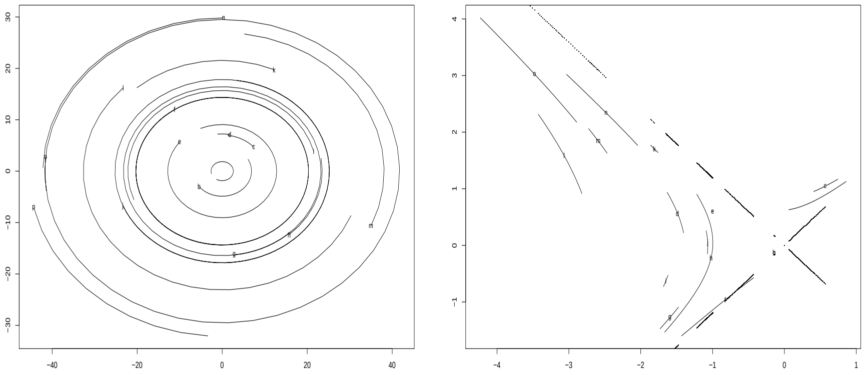


Figure 7.11: Realizations of path obtained with an SDE with jumps (Example 3). Left: Ellipsoidal orbits ($\beta = 2$). Right: Hyperbolic orbits ($\beta = -0.7$).

When are these stochastic differential equations PDMP?

As seen earlier in this chapter and the introductory chapter, piecewise deterministic Markov processes (PDMP) introduced by M.H.A Davis [DAV 84] enrich the usual classes of Markov Processes (diffusions, jump processes,...) by allowing a part of determinism in paths while inheriting the appealing Markovian properties [COS 08]. PMDP are time homogeneous $\mathbb{R}^d$-valued processes $\mathbf{x}(t)$ with $c\grave{a}dl\grave{a}g$ sample paths.

PDMPs can be sequentially *constructed* via an increasing sequence of stopping times $(T_n)_{n\geq 0}$ with $T_0 = 0$ and $\mathbf{x}(0) = \mathbf{x}_0$. For $\mathbf{x}(T_n) = \mathbf{x}_n$ and $t \in [T_n, T_{n+1}[$, the process $\mathbf{x}(t)$ obeys a deterministic rule, e.g. an ODE $d\mathbf{x}(t) = V(\mathbf{x}(t))dt$, governed by a regular vector field V. Then, conditionally to the past $\mathcal{F}_{T_n}$, the lifetime $D_{n+1} = T_{n+1} - T_n$ has hazard function $\lambda(s) = h(\mathbf{x}(T_n + s))$, where h is a non-negative bounded measurable function on $\mathbb{R}^d$. At time T_{n+1}, the process $\mathbf{x}(t)$ jumps to a state $\mathbf{x}(T_{n+1}) = \mathbf{x}(T_{n+1}-) + \varepsilon_{n+1}$, in accordance with a probability transition $Q(d\varepsilon|\mathbf{x}(T_{n+1}-))$. The triplet (V, h, Q) characterizes entirely the probability distribution of the PMDP. Note that when h satisfies $< \nabla h(\mathbf{x}), F(\mathbf{x}) > \equiv 0$, for all $\mathbf{x}$, i.e. h is a first integral for dynamical system, then $\lambda(s)$ is constant on the deterministic parts of paths and therefore the D_n are exponentially distributed.

Piecewise deterministic processes presented in this paper are based on a little more general marked point processes $N(dt, dx) = \sum_{i\geq 0} \delta_{(T_i,\varepsilon_i)}$ and so are neither Markovian nor time homogeneous in general and, therefore, are not PDMP in general. For the class of processes developed here to be PDMP, it is sufficient that the conditional cumulative intensities are separable measures in dt and dx and have the following form:

$$\Lambda_i(dt - T_i, dx) = h(\mathbf{x}(T_i + t))dt \times Q(dx|\mathbf{x}(T_{n+1}-)).$$

7.4.3. *The Doléans Dade Exponential Semimartingales*

The following theorem is borrowed from Protter [PRO 05] and is a consequence of the change of variables theorem as regards to Ito calculus for semimartingales.

THEOREM 7.1.– If X is a semimartingale with $X(0) = 0$, then there exists a unique semimartingale Z satisfying the equation $dZ(t) = Z(t-)dX(t)$, with $Z(0) = 1$ which is given by:

$$Z(t) = \exp^{(X(t) - \frac{1}{2}[X,X]^c(t))} \prod_{s\leq t} \left((1 + \Delta X(s)) \exp^{-\Delta X(s)} \right). \qquad [7.28]$$

The solution $Z(t)$, usually denoted $\mathcal{E}_X(t)$, is called the stochastic (or Doléans Dade) exponential of X. This theorem encompasses many useful results and applications. The formula reduces a lot for locally bounded variation processes X, since in this case $[X,X]^c(t) \equiv 0$ implies that

$$Z(t) = \exp^{X^c(t)} \prod_{s \leq t} (1 + \Delta X(s)) \, .$$

Under a mild integrability condition, if $X(t)$ is a martingale, then $Z(t)$ is also a martingale. A multivariate version of the theorem exists [JAC 82] and corresponds to the analog of deterministic linear differential equations $dZ(t) = Z(t-)dX(t)$ with a matrix process A and a vector semimartingale X.

In what follows, we present several applications of exponential martingales.

Example 4: Deterministic semimartingales

Theorem 7.1 includes extensions of the case of deterministic homogeneous linear differential equations. For instance, for any $d \times d$ matrix A, there exists a unique solution $Z(t) = \exp^{At} z_0$ to equation $dZ(t) = AZ(t)dt$, with $Z(0) = z_o \in \mathbb{R}^d$, taking here the deterministic matrix semimartingale $X(t) = At$.

Example 5: Cumulative hazard function

The probability distribution function $F(t) = P(T \leq t)$ and the survival function $S(t) = 1 - F(t)$ of a non-negative random variable T, with $dS(t) = -dF(t)$, are monotonic functions and have bounded variations. The cumulative hazard function $\Lambda(t) = \int_0^t \frac{dF(s)}{1-F(s-)}$ satisfies the equation $d\Lambda(t)(1 - F(s-)) = d(F(s))$. Conversely, given a positive increasing function Λ with $\Lambda(0) = 0$, there exists a unique function S with $S(0) = 0$, which satisfies $dS(t) = -S(t-)d\Lambda(t)$. Equation [7.28] implies that S satisfies:

$$S(t) = 1 - F(t) = \exp^{-\Lambda^c(t)} \prod_{s \leq t} (1 - \Delta\Lambda(s)) \, .$$

Note that the absolutely-continuous case $d\Lambda(t) = \lambda(s)ds$ yields $S(t) = e^{-\int_0^t \lambda(s)ds}$.

Example 6 : Survival analysis

Survival analysis in statistics is based on the simple case of a simple point process with at most one event at time T. Let $S(t)$ and $\Lambda(t)$ be respectively the survival and cumulative hazard functions of T; then according to formula [7.25], the associated compensating martingale is written $M(t) = \mathbf{1}_{\{T \leq t\}} - \int_0^t \mathbf{1}_{\{s<T\}} \lambda(s)ds = \mathbf{1}_{\{T \leq t\}} - \Lambda(t \wedge T)$.

Since M is a pure jump martingale, with $[M,M]^c(t) \equiv 0$ and $M^c(t) = -\Lambda(t \wedge T)$, its exponential is also a pure jump martingale and satisfies: $Z(t) = \exp^{-\Lambda(t \wedge T)} \left(1 + \mathbf{1}_{[T,\infty[}(t)\right) = S(t \wedge T) \left(1 + \mathbf{1}_{[T,\infty[}(t)\right)$.

For statistical purposes, we have however to deal with a little more sophisticated exponential martingale. Assume for example that T has hazard functions $\lambda_0(t)$ under probability $\mathbb{P}_0$ and $\lambda_\theta(t)$ under probability $\mathbb{P}_\theta$, such that $\lambda_\theta(t) = \mu_\theta(t)\lambda_0(s)$. Now, if we consider the $\mathbb{P}_0$ martingale $X_\theta(t) = \int_0^t(\mu_\theta(s) - 1)dM_0(s)$, we find that its stochastic exponential $Z_\theta(t) = \exp^{\int_0^t \log(\mu_\theta(s))dN(s) - \int_0^t(\mu_\theta(s)-1)\lambda_\theta(s)ds}$ is also a $\mathbb{P}_0$ martingale that exactly corresponds to the likelihood ratio $L_\theta(t) = E(\frac{d\mathbb{P}_\theta}{d\mathbb{P}_0}|\mathcal{F}_t) = f_\theta(t)^{\mathbf{1}_{\{T\leq t\}}}(1 - F_\theta(t))^{\mathbf{1}_{\{T>t\}}}$.

This construction is in fact a major key for dealing with more general statistical contexts (see Section 7.4.4).

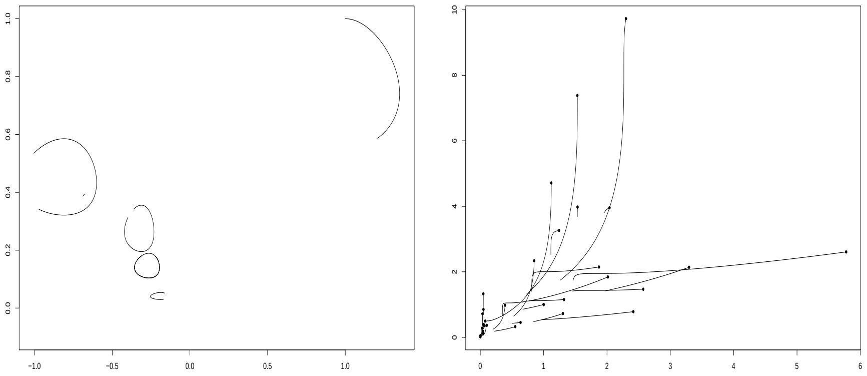


Figure 7.12: Sample paths of 2D stochastic exponentials $Z(t) = \mathcal{E}(M(t))$ (see Equation [7.28]) driven by compensating martingales $M_j(t); j = 1, 2$, based on a standard Poisson process $N(t)$. Left: $M_1(t) = \int_0^t \cos(s)(dN(s) - ds)$ and $M_2(t) = -\int_0^t \sin(s)(dN(s) - ds)$. Right: $M_1(t) = \int_0^t(1 + \cos(s))(dN(s) - ds)$ and $M_2(t) = \int_0^t(1 - \sin(s))(dN(s) - ds)$.

7.4.4. *Statistical Issues*

General case

We present here an important application of the exponential semimartingale theorem allowing a statistical approaches for marked point processes and related models such as PDMPs. It is a sort of Girsanov theorem characterizing the ratio of probability measures. Given two equivalent probability measures $\mathbb{P}$ and $\mathbb{Q}$ on some complete filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\leq 0}$, we already knows that the Radon-Nycodym derivative $Z(\omega) = \frac{d\mathbb{P}}{d\mathbb{Q}}$ is a positive $\mathbb{Q}$-integrable random variable implying therefore that the process $Z(t) = E(Z|\mathcal{F}_t)$ is a positive uniformly integrable martingale that equals $\frac{d\mathbb{P}_{|\mathcal{F}_t}}{d\mathbb{Q}_{|\mathcal{F}_t}}$, such that $Z(t)$ corresponds to a ratio of likelihoods.

In a statistical framework, considering a parametric set of probabilities $(\mathbb{P}_\theta, \theta \in \Theta)$ equivalent to $\mathbb{Q}$, such that $d\tilde{N}_\theta(t) = \mu_\theta(t)d\tilde{N}(t)$, where $\tilde{N}_\theta$ and $\tilde{N}$ are the respective compensators of N, one may expect to find , under some mild conditions, a particular $\mathbb{Q}$-martingale $W_\theta(t) = \int_0^t \rho_\theta(s)d(N - \tilde{N})(t)$ such that the likelihood ratio $Z_\theta(t)$ corresponds to the positive stochastic exponential of $W_\theta(t)$. Indeed, one can promptly and heuristically prove that it is true (and only true) for the process $\rho_\theta(t) = \lambda_\theta(t) - 1$.

In the context of multivariate/marked point processes, Jacod [JAC 75] gives a plain formula for the Radon-Nycodym derivative $\frac{d\mathbb{P}_{|\mathcal{F}_t}}{d\mathbb{Q}_{|\mathcal{F}_t}}$ under the natural filtration. This formula corresponds to the solution of the Doléan Dade equation for the martingale $W_\theta(t)$.

An Example

The statistical approach proposed above is applied here to the process presented in Example 3 of Section 7.4.2, which is piecewise driven by an ODE and randomly jumps at times T_i with jump amplitudes ε_i; in other words, this process statisfies the stochastic differential equation: $dX(t) = V(X(t-))dt + \sum_i \varepsilon_i \delta_{T_i}$.

Among the many potential measurements of the movement (eg length, kinetic energy,...), let us take the travel length L as a movement characteristic of a particle on orbits. For a particle starting from $\mathbf{x}_0$ at time $t = 0$, this is defined by:

$$L(t, X_0) = \int_0^t |V((X(s))|ds.$$

On one hand, let us assume that the random measure $N(dt, dx) = \sum_i \delta_{(T_i, \varepsilon_i)}$ has under probability $\mathbb{Q}$ the conditional intensities:

$$\Lambda_i(dt, dx \mid \mathcal{F}_{T_i}) = dt\mathbf{1}_{\{T_i \leq t < T_{i+1}\}} \times \varphi(x|\mathbf{0}, \mathbf{I}_2)dx,$$

where $\varphi(x|\mathbf{m}, \Sigma)$ stands for the Gaussian density with mean $\mathbf{m}$ and covariance Σ in $\mathbb{R}^2$. In that case we obtain $\tilde{N}(dt, dx) = dt\varphi(x|\mathbf{0}, \mathbf{I}_2)dx$, meaning that $N(dt, dx)$ is a Poisson measure under $\mathbb{Q}$.

On the other hand, let us assume that under $\mathbb{P}_\theta$, the conditional intensities depends on paths as follows:

$$\Lambda_{i,\theta}(dt, dx \mid \mathcal{F}_{T_i}) = \alpha h_{\beta_1}(t - T_i, X(T_i))\mathbf{1}_{\{T_i \leq t < T_{i+1}\}}dt \times \varphi(x|\mathbf{m}_\theta(t), \mathbf{I}_2)dx,$$

where for $T_i \leq t < T_{i+1}$, $\theta = (\alpha, \gamma, \beta_1, \beta_2)$ and $\gamma = (\gamma_1, \gamma_2)$, we define:

$$h_{\beta_1}(s, X_0)) = \frac{d}{ds} L^{\beta_1}(s, X_0)$$

$$\mathbf{m}_\theta(t) = \gamma\, L^{\beta_2}(t - T_i, X(T_i)).$$

The function $h_\beta(s, X_0)$, should be interpreted as the hazard function of the Weibull distribution $W(1, \beta)$ related to the positive travel length variable L on the orbit starting from X_0.

The previous equations ultimately tells that $\tilde{N}_\theta(dt, dx) = \lambda_\theta(t, x)\tilde{N}_\theta(dt, dx)$, with

$$\lambda_\theta(t, x) = \alpha \sum_{i \geq 0} h_{\beta_1}(t - T_i, X(T_i))$$
$$\exp^{-\frac{1}{2}[<\gamma,\gamma> L^{2\beta_2}(t-T_i, X(T_i)) - 2<x,\gamma> L^{\beta_2}(t-T_i, X(T_i))]}\, \mathbf{1}_{\{T_i \leq t < T_{i+1}\}}.$$

Next, for the sake of simplicity, let us suppose that the process is observed in the random time interval $[0, T_n]$, such that the likelihood ratio corresponds to the stochastic exponential of the $\mathbb{Q}$ martingale $W_\theta(t) = \int_0^t \int_{\mathbb{R}^2} (\lambda_\theta(s, x) - 1)(N - \tilde{N})(ds, dx)$. According to formula [7.28], the log-likelihood is equal to:

$$\begin{aligned}
\log(Z_\theta(T_n)) = &-\int_0^{T_n} \int_{\mathbb{R}^2} (\lambda_\theta(s, x) - 1)\tilde{N}(ds, dx) \\
&+ \int_0^{T_n} \int_{\mathbb{R}^2} \log(\lambda_\theta(s, x)) N(ds, dx). \\
= &-\alpha \left(\sum_{i=0}^{n-1} L^{\beta_1}(T_{i+1} - T_i, X(T_i)) \right) - T_n + n \log(\alpha) \\
&+ \sum_{i=0}^{n-1} \log(h_{\beta_1}(T_{i+1} - T_i, X(T_{i+1}))) \\
&- \frac{1}{2} < \gamma, \gamma > \sum_{i=0}^{n-1} L^{2\beta_2}(T_{i+1} - T_i, X(T_{i+1})) \\
&+ \sum_{i=0}^{n-1} < \gamma, \Delta X(T_{i+1}) > L^{\beta_2}(T_{i+1} - T_i, X(T_{i+1})).
\end{aligned}$$

One can therefore easily derive the set of equations for the maximum likelihood estimate (MLE) $\hat{\theta}$ and apply a classical optimization procedure. As an illustration, we deal here with the simple case where the parameters β_1 and β_2 are known, which allows us to get explicit formulas for the MLE of α and $\gamma = (\gamma_1, \gamma_2)$:

$$\hat{\alpha} = \frac{n}{\sum_{i=0}^{n-1} L^{\beta_1}(T_{i+1} - T_i, X(T_i))}$$

$$\hat{\gamma} = \frac{\sum_{i=0}^{n-1} \Delta X(T_{i+1}) L^{\beta_2}(T_{i+1} - T_i, X(T_{i+1}))}{\sum_{i=0}^{n-1} L^{2\beta_2}(T_{i+1} - T_i, X(T_{i+1}))}.$$

As a perspective, one can expect to use asymptotic techniques for discrete time indexed martingales in order to derive the asymptotic behaviors (in almost sure and in distribution senses) of these estimators and, therefore, perform sensible null hypothesis testing such as $\gamma = 0$ and $\alpha = \alpha_0$.