Tropospheric contact networks: design, estimation and use for surveillance of airborne pathogens

Maria Choufany

Thèse de doctorat de l'université Aix-Marseille
Discipline : Mathématiques Appliquées

Présentée par :

Maria Choufany

Intitulée :

Réseaux de contact troposphérique : conception, estimation et utilisation pour la surveillance des pathogènes aéroportés.

Soutenue le 9 décembre 2020

Devant le jury composé de :

Elisabeta Vergu	Directrice de recherche, INRAE - Jouy-en-Josas	Rapporteure
Eric Matzner-Lober	Professeur des universités - Université Rennes 2	Rapporteur
Catherine Abadie-Fournier	Directrice de recherche, CIRAD - Montpellier	Examinatrice
Alain Franc	Directeur de recherche, INRAE - Bordeaux	Examinateur
Samuel Soubeyrand	Directeur de recherche, INRAE - Avignon	Directeur de thèse
Cindy E. Morris	Directrice de recherche, INRAE - Avignon	Co-directrice de thèse
Davide Martinetti	Chargé de recherche, INRAE - Avignon	Co-directeur de thèse

---

Université Aix-Marseille
ED 184 - Mathématiques et Informatique de Marseille
UR INRAE Biostatistique et Processus Spatiaux

Résumé

Anticiper les épidémies de maladies infectieuses (endémiques et émergentes) et atténuer leurs impacts sont des défis majeurs en pathologie humaine, animale et végétale. La surveillance pour la détection des maladies des plantes et les réactions ultérieures pour limiter leurs effets mobilisent des ressources humaines et économiques considérables. De nombreux agents pathogènes des plantes sont disséminés par l'air sur des échelles spatiales à la fois petites et grandes. Il est donc essentiel de développer des stratégies de surveillance adéquates (c'est-à-dire prenant en compte les caractéristiques des mouvements des agents pathogènes via l'air) permettant d'améliorer la détection précoce d'émergences de nouveaux pathogènes aéroportés et le suivi d'épidémies de pathogènes endémiques.

Dans ce contexte, la recherche menée dans ma thèse consiste essentiellement à concevoir des méthodes visant à modéliser, estimer et caractériser les réseaux de contact troposphérique (la troposphère est la couche inférieure de l'atmosphère) et à exploiter leurs propriétés pour mieux surveiller les pathogènes aéroportés des plantes. Cette recherche est essentiellement fondée sur des outils mathématiques et statistiques, sur un logiciel externe de reconstruction des trajectoires des masses d'air (en l'occurrence `HYSPLIT`; https://ready.arl.noaa.gov/HYSPLIT.php) exploitant des bases de données météorologiques massives, et sur la modélisation compartimentale des épidémies classiquement évoquée sous l'acronyme SIR (*susceptible, infected, recovered*). Le logiciel `HYSPLIT` est utilisé pour reconstituer de manière relativement réaliste les mouvements des masses d'air et leur non-stationnarité en temps et en espace, mouvements qui sont cruciaux pour prévoir la propagation à une large échelle des agents pathogènes microscopiques à dissémination aérienne.

Pour répondre à la problématique décrite ci-dessus, je développe dans une première partie de la thèse une approche généraliste destinée à modéliser et inférer un réseau de contact entre zones géographiques à partir d'un ensemble de trajectoires d'individus. Cette approche est présentée sous un formalisme mathématique dans le cadre de la théorie des graphes. Les liens entre les nœuds du graphe (en l'occurrence les zones géographiques sus-mentionnées) sont estimés en analysant le passage (au sens large) des trajectoires à travers les nœuds, ce qui permet d'inférer le réseau de contact (ou de connectivité) dans son ensemble. Différents estimateurs pour les liens entre noeuds, fondés sur différentes hypothèses bio-physiques, sont proposés. En appliquant l'approche de construction du réseau aux trajectoires des masses d'air sur des périodes temporelles successives, on obtient un réseau spatio-temporel de contact troposphérique donnant une estimation de la probabilité de connexion pour

chaque paire de nœuds *donneur–receveur* au fil du temps. De plus, je propose de mesurer l'erreur ou l'incertitude attachée à la quantification des connexions entre nœuds en considérant plusieurs contextes d'échantillonnage spatial et temporel. Une fois construit et estimé, le réseau peut être caractérisé topologiquement en utilisant des statistiques classiques de la théorie des graphes.

Dans une deuxième partie de la thèse, je construis un modèle épidémiologique conditionné par un réseau de contact dans le but d'évaluer diverses stratégies de surveillance. Le modèle épidémiologique est un modèle dynamique spatial et compartimental de type SIRS (*susceptible, infected, recovered, susceptible*) qui représente la propagation d'un pathogène à travers le réseau, que ce réseau soit estimé à partir des trajectoires des masses d'air ou obtenu d'une autre manière. Ensuite, des stratégies de surveillance spatio-temporelles, dont certaines dépendent du réseau de contact, sont proposées et évaluées en fonction de leurs capacités à détecter précocement une épidémie d'un pathogène émergent. L'objectif est de déterminer *où* et *quand* surveiller la population hôte du pathogène.

Ces travaux, qui reposent sur des données météorologiques massives et des codes numériques conséquents, contribuent à dépasser les paradigmes classiques de la surveillance des pathogènes aéroportés des plantes. Ils revêtent par ailleurs une dimension générique qui laisse envisager leur application à d'autres cadres que l'épidémiologie végétale.

Mots clés : statistiques, mathématiques appliqués, théorie des graphes, réseaux, épidémiologie, surveillance, science atmosphérique.

Abstract

Anticipating outbreaks of infectious diseases (endemic and emerging) and mitigating their impacts are major challenges in human, animal and plant pathology. Surveillance for the detection of plant diseases and subsequent responses to limit their effects mobilize considerable human and economic resources. Many plant pathogens are airborne on both small and large spatial scales. It is therefore essential to develop adequate surveillance strategies (i.e. taking into account the characteristics of pathogen movement via air) to improve early detection of new airborne pathogen emergence and monitoring of endemic pathogen outbreaks.

In this context, the research carried out in my thesis consists essentially in designing methods to model, estimate and characterize tropospheric contact networks (the troposphere is the lower layer of the atmosphere) and to exploit their properties to better monitor airborne plant pathogens. This research is essentially based on mathematical and statistical tools, on external software for reconstructing the trajectories of air masses (in this case `HYSPLIT`; https://ready.arl.noaa.gov/HYSPLIT.php) using massive meteorological databases, and on the compartmental modeling of epidemics, classically referred to by the acronym SIR (*susceptible, infected, recovered*). The software `HYSPLIT` is used to reconstruct in a relatively realistic way the movements of air masses and their non-stationarity in time and space, movements which are crucial for predicting the large-scale spread of microscopic airborne pathogens.

In order to answer the above described problem, I develop in a first part of the thesis a generalist approach intended to model and infer a network of contact between geographical areas from a set of individuals' trajectories. This approach is presented under a mathematical formalism within the framework of graph theory. The links between the nodes of the graph (in this case the above-mentioned geographical areas) are estimated by analyzing the passage (in the broad sense) of the trajectories through the nodes, which allows the inference of the contact network (or connectivity) as a whole. Different estimators for the links between nodes, based on different biophysical hypotheses, are proposed. By applying the network construction approach to air mass trajectories over successive time periods, we obtain a spatio-temporal tropospheric contact network giving an estimate of the probability of connection for each pair of nodes *donor-receiver* over time. In addition, I propose to measure the error or uncertainty attached to the quantification of connections between nodes by considering several spatial and temporal sampling contexts. Once constructed and estimated, the network can be characterized topologically using classical graph theory statistics.

In a second part of the thesis, I construct an epidemiological model conditioned by a contact network in order to evaluate various surveillance strategies. The epidemiological model is a dynamic spatial and compartmentalized model of the SIRS type (*susceptible, infected, recovered, susceptible*) which represents the propagation of a pathogen through the network, whether this network is estimated from the trajectories of air masses or obtained in another way. Then, spatio-temporal monitoring strategies, some of which depend on the contact network, are proposed and evaluated in terms of their ability to detect an outbreak of an emerging pathogen at an early stage. The objective is to determine *where* and *when* to monitor the pathogen's host population.

This work, which is based on massive meteorological data and consistent numerical codes, contributes to going beyond the classical paradigms of airborne plant pathogen monitoring. Moreover, it has a generic dimension that allows its application to frameworks other than plant epidemiology.

Keywords: statistics, applied mathematics, graph theory, networks, epidemiology, surveillance, atmospheric science.

Acknowledgements

I would like to thank first of all my thesis directors: *Samuel Soubeyrand*, *Davide Martinetti* and *Cindy E. Morris*. I am so grateful for them, for all the time they devoted to the success of this work. Thank you for your patience, your availability and above all your judicious advice. I have learned a lot from you and I address all my gratitude for all of this. Besides, I would especially like to thank *Rachid Senoussi* for his involvement and contribution in my work.

I address all my thanks to *Elisabeta Vergu*, as well as to *Eric Matzner-Lober*, of the honor they did to me by accepting to be reporters of this thesis.

I associate with these, a special thanks to *Catherine Abadie* and *Alain Franc* for accepting to review my work.

"Family" BioSP, I would like to thank you one by one for all the unforgettable moments you have offered me. It was a great pleasure to work with such a nice and dynamic team. I am keen on thanking you for your support that helped me to succeed. CNAM' team, it was very pleasant to work with such a professional and friendly people. ACACIA' squad, thank you for the many good moments spent together, the thesis life was full of gladness with you.

A very special thanks to all my friends who have always been a comfort when I needed it the most. Words can never express how grateful I am that you're always there for me. Without your presence, this journey was going to be much less fun.

I add, and without forgetting a big thanks for my family and my in-laws to their non-stop support.

And last, but not least, my *husband* to be... I simply couldn't have done this without you. Thank you for your encouragement, care, love and support <3

Cachan' team, here I come ;)

Table of contents

List of Figures

List of Tables

1. Introduction

Since the early days, plant epidemics have taken center stage in populations by affecting them in umpteen ways: financially, biologically, culturally, etc. Their outbreaks go across centuries and continents. Hence, there's an essential need to preempt their emergence at the local scale before they reach the global one through dissemination via the air, water, human transportation, insects, etc. Amongst the common diffusion pathways, I will shed light on the possible long-range diffusion associated with the movement of air masses in the atmosphere. The dispersal of the pathogens via the air remains the most common pathway by which long-range diffusion of pathogens is carried. It is associated with the movement of air masses in the atmosphere, which are known for transporting promptly plant pathogens over hundreds or even thousands of kilometers (Aylor, 2003; Aylor et al., 1982; Brown and Hovmøller, 2002) and then causing major complications in the agricultural production, human health and ecosystem functioning (Mundt et al., 2009). Thus, it is crucial to improve effective control strategies to reduce the propagation of infections. Modeling such a dispersal can be approached by using networks ascertained through graph theory (Jeger et al., 2007; Kiss et al., 2017; Margosian et al., 2009; Moslonka-Lefebvre et al., 2011).

To apply graph theory to the modeling of dispersal, this thesis aims to study statistical and modeling tools for monitoring the spread of airborne plant pathogens and the emergence of new epidemics on a large geographical scale. This is done by estimating and characterizing tropospheric contact networks that delineate air mass connectivity between spatial areas through time. Thanks to the resulting estimated networks, we will be able to outline the pattern of the air masses trajectories over a spatio-temporal scale. This will be an essential beginning point for the ongoing improvement of monitoring mechanisms of airborne plant pathogens spread and the emergence of new epidemics on a large geographical scale. To secure this objective, we proceed by exploiting, in a complementary way, two research stages presented in the following manner:

First, based on a mathematical approach, we conceive and estimate the so-called contact network in line with the trajectories between individuals that is essentially based on two components: graph theory and trajectories. The concept of graph theory has been widely used to model pairwise connections between the individuals of a susceptible population (Kolaczyk and Csárdi, 2014; Newman, 2002; Pastor-Satorras et al., 2015). The trajectories are used to describe the spatio-temporal paths followed by an object in the space. They are renowned to characterize the movements of individuals

such as animals (Dalziel et al., 2008; Hooten et al., 2018), humans (Hoteit et al., 2014; Traunmueller et al., 2018), plant pathogens (Leyronas et al., 2018; Sadyś et al., 2014), etc. Combining these two concepts will provide a proficient tool to understand the dissemination of particles within spatial areas via a transfer mode composed of trajectories. Our concern is limited to trajectories involved in the troposphere and having a capacity to cover thousands of kilometers over a small period. Such trajectories can transport living organisms such as insects, fungi, bacteria, viruses, and pollen as well as inorganic material such as wildfire smoke, sand, dust or radioactive particles (Pérez et al., 2015). Hence, the need for modeling these trajectories to be able to fully follow the invasive species (Bogawski et al., 2019; Hallar et al., 2011; Monteil et al., 2014; Sadyś et al., 2014). Then, by considering such types of trajectories in the application of the contact network, we will be able to estimate the dynamic tropospheric contact network represented by spatial areas connection via the air mass trajectories over time.

Second, we explore the constructed networks by considering both network and node metrics that are essential in the characterization of network topology and in the identification of the nodes that can affect the epidemic spread within the population (De Arruda et al., 2014; Li et al., 2018; Meyers et al., 2003; Miller and Kiss, 2014; Moslonka-Lefebvre et al., 2011; Newman, 2002; Strona et al., 2017). Overall network metrics, such as the degree distribution, the average shortest path between nodes, the number of connected components, or the clustering coefficient, are used to summarize complex topological feature of the network and can provide useful insight on the likelihood of a major epidemic outbreak (Newman, 2002; Pastor-Satorras et al., 2015; Strogatz, 2001). On the other hand, node metrics assign a value to each node of the network according to some criterion, usually related to its centrality (even though there are several different definitions of centrality) or its capacity of spreading diseases or information to its contacts. Such metrics include node degree, betweenness, coreness and k-shell centrality, h-index, and more complex indices that have been proposed in the literature. A recent and rather **a** comprehensive list of such indices can be found in the review of Lü et al. (2016) and references therein. Another way of studying the epidemic spread in a population is to simulate the disease progression using compartmental models. These models usually assume that the host population can be divided into separate compartments according to their disease status. A typical example is the Susceptible-Infectious-Recovered (S-I-R or simply SIR) model, where each host, at each step of the simulation, can belong to only one of the three compartments. Furthermore, a rather simple set of rules and parameters allows describing how individuals progress within the different compartments (Diekmann and Heesterbeek, 1989; Kermack and McKendrick, 1932; Kermack et al., 1927). Different variations of the standard SIR model have been proposed in the literature to account for non-lethal non-immunizing diseases without the R compartment (SI or SIS), diseases with a non-infectious exposition period (SEIR, where E stands for Exposed) or diseases that do not confer lifelong immunity (SIRS). Traditionally, these

models are defined under the hypothesis of homogeneous mixing of individuals, i.e. they assume that all hosts have identical rates of disease-causing contacts. Whereas this assumption makes the analysis tractable using ordinary differential equations, once epidemiological parameters such as the infection and recovery rate have been defined, it may not adequately reflect the complex reality of heterogeneous contact patterns between individuals (Bansal et al., 2007). A more realistic, but less tractable approach is to assume a known underlying contact structure, that can be conveniently represented through a network of contacts between individuals (Newman, 2002). A prominent example of the use of SIR models on networks can be found in (Brockmann and Helbing, 2013), where the authors studied the worldwide 2009 H1N1 influenza pandemic and the 2003 SARS epidemic compared a simple wave-diffusion model based on geographical distance versus a network-based model issued from flight connections between countries to find out that the second model was capable of better predicting the arrival time of the disease into a new country.

Aside from the characterization of the epidemics, such models can be used to design efficient epidemic surveillance strategies based on the dynamic model and aim to anticipate the emergence of new epidemics and reduce the delay between an outbreak and its detection (Colman et al., 2019; Herrera et al., 2016; Martinetti and Soubeyrand, 2019; Polgreen et al., 2009).

To delve more deeply, we formulate and answer the following problems in this manuscript.

1. How to model, construct, and estimate a contact network via trajectories?
2. How can a contact network be developed through tropospheric connections?
3. How to evaluate the accuracy of the connectivity estimators according to spatio-temporal sampling schemes?
4. What are the topological characteristics of the tropospheric contact network?
5. How can infectious diseases spread through a tropospheric network?
6. What are the most effective spatio-temporal surveillance strategies for boosting the early detection of an epidemic?

To address these research questions, the manuscript will be structured in five main chapters representing each of these questions, starting from the conception of the networks to the epidemic characterization. Hereafter, a general overview of the manuscript's chapters is given :

Chapter 2 states the scientific context branched into two research axes: the epidemiological and the mathematical. We start by giving a brief review of the long-range dispersal of pathogens. Then, we introduce some key elements in the mathematical modeling of epidemics (such as dispersion). In chapter 3, we detail the proposed approach used to infer and construct networks via a set of trajectories connecting

individuals based on different mathematical definitions and properties. Different estimators of the connections are proposed and can be applied according to the context of the study. In our case, we apply the approach by considering the air masses trajectories in order to infer the aerial connectivity between spatial areas through a period of time. Then, we apply the approach in two spatial case studies: the coastline of the Mediterranean sea and the French region of Provence-Alpes-Côte d'Azur (PACA). Also, we assess the spatial and temporal patterns based on the probability of connections between the spatial areas in the temporal context. When referring to estimators in statistics, the computation of the attached uncertainty is paramount. As defined in chapter 3, several estimators, which are actually approximations of space-time integrals, were proposed. Consequently, in order to compute the accuracy of the estimator, it is of central importance to consider the impact of the spatio-temporal sampling scheme on the integral approximation. Hence, in chapter 4, we compute theoretically the accuracy of one of the connectivity estimators by considering three general sampling schemes: a classical Monte Carlo sampling scheme, a random sampling scheme with dependencies, and a deterministic sampling scheme. To endorse the theoretical results, we design a numerical comparison of the estimators' efficiency within the framework of the aerial contact-based connectivity within the watersheds of PACA. After completing the conceptual framework which results in a spatio-temporal tropospheric network, we tend to characterize it by considering statistical metrics. Therefore, in chapter 5, we develop an approach to bare the patterns of air masses connectivity based on the spatio-temporal networks with the aim of detecting strong sources or strong receptors of airborne pathogens according to a temporal period. In order to reach a better characterization of the affinity between the epidemic dispersal and the spatio-temporal tropospheric networks, we construct, in chapter 6, an epidemic model, in which dispersal events are drawn conditionally on the spatio-temporal networks defined above. The outputs of this model are essential to set up the appropriate surveillance strategies in order to detect the prevalence of the epidemic over the network by considering seasonality patterns. Finally, in chapter 7, we conclude and discuss further the results of this work with the possible perspectives in the future.

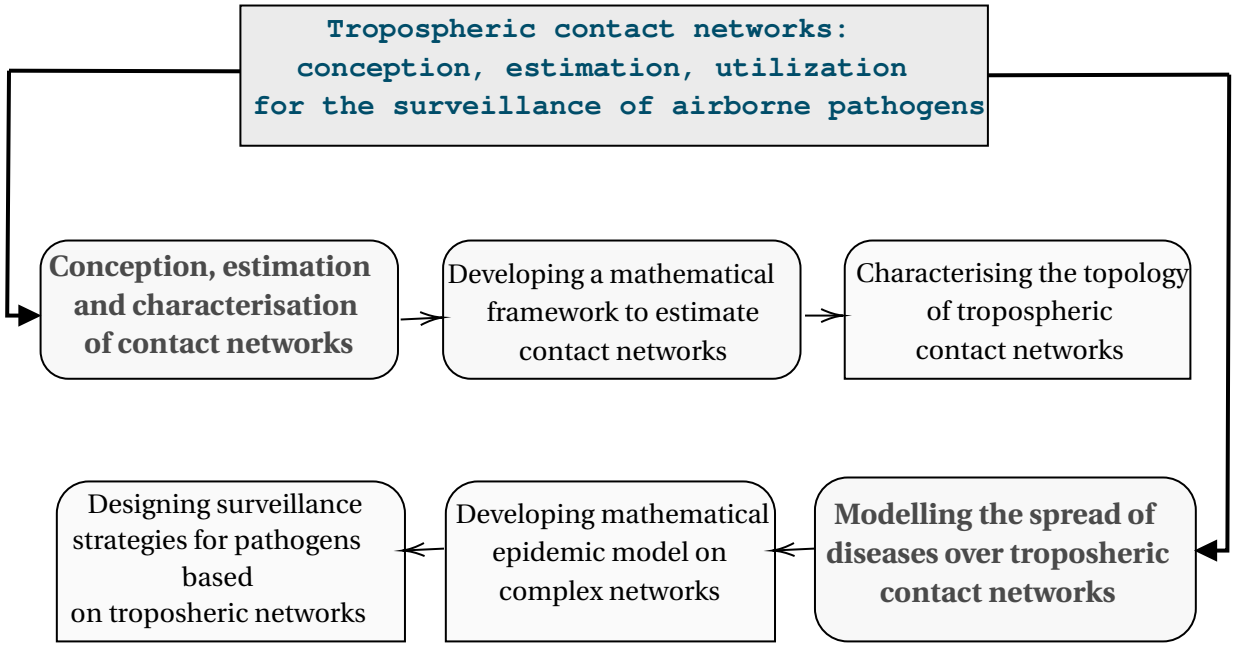


Figure 1.1. – Main objectives of the thesis.

2. State of Art

Table of Contents

2.1. Epidemiological Context

The ease of disease transmission has been increasing throughout the past decades. However, handling such situations is becoming much more complicated since it has tremendous impacts on both, the ecosystem functioning and the human welfare. The failure of managing such situations, mainly those who have been transmitted over large scales, shows the lack of sufficient understanding of disease epidemiology. This field of study has been highly developed during recent years in a multidisciplinary context by epidemiologists, ecologists, pathologists, etc. in a shared context: understanding the spread, evolution, and control of invasive organisms.

2.1.1. Epidemiological terms and definitions

A *disease* is deterioration over time of the health of living beings that affects negatively their structure or parts of their bodies. The diseases can be infectious and/or transmissible or not. An *infectious disease* can arise from bacteria, fungi, pathogenic microbial agents, etc. It is recognized by its asymptomatic infection during its *incubation period*, the time between the exposure to the disease and the appearance of the first symptoms. A *non-infectious disease* is all other diseases, that can not be transmitted between organisms. They can be caused by toxic components, nutrient deficiencies, or climatic conditions for example. In table 2.1, we mentioned some examples of infectious and non-infectious diseases that can affect humans, plants, and animals.

When the *occurrence, frequency, incidence and prevalence* of a disease, number of infected individuals during a defined time period, increases dramatically, it is said to have become *epidemic*. If this *epidemic* spreads permanently in a region or population, it is called *endemic*, and if it spreads over global scale by covering multiple countries and even continents, it is called *pandemic*. The *occurrence* is the appearance of the disease over time. The *frequency* is the number of occurrences. The *incidence*, is the rate of occurrence of the disease at a specific point in time t computed by:

$$\text{Incidence}(t) = \frac{\text{Number of new infected cases at time } t}{\text{Total size of the population}}.$$

The prevalence is the proportion of infected cases at a specific point in time t defined as:

$$\text{Prevalence}(t) = \frac{\text{Number of existing infected cases at time } t}{\text{Total size of the population}}.$$

The *prevalence* indicates the extent to which the disease is widespread whereas the *incidence* indicates the risk of catching the disease.

	Infectious diseases	Non-infectious diseases
Humans	Covid19, Cholera, Smallpox	Alzheimer, cancer, cardio-vascular diseases
Plants	Curly top, Downy mildew, Mosaic	Bronze leave, Sun scald, Chemical damage
Animals	Rabies, Psittacosis, Diphtheria	Glaucoma, Hip Dysplasia, Arthritis

Table 2.1. – Examples of infectious and non-infectious diseases for humans, animals, and plants.

In what follows, I will be focusing on infectious plant diseases. The research and development context focuses on general approaches for plants with perspectives for animals.

2.1.2. Plant Epidemic dispersal

In recent decades, the infectious agents are expanding geographically (Mayer, 2000; Wilson, 2010). This expansion lies at the root of health crises and even natural disasters. This involves a need to look up after the causes and the circumstances related to their appearance and the development of particular pathogens. Pathogens, i.e microorganisms such as fungi, bacteria, and viruses that cause diseases, can reduce the productivity, quality, and even cause the death of plants. They can also infect farm and wild animals, but here, we will focus on plant pathogens. They can be introduced and spread to the host plants in different ways. Wind, rain, soil, vector insects, etc. facilitate the transport of pathogens from an infected plant to the uninfected one (Brown and Hovmøller, 2002; Dwyer and Elkinton, 1995; Fitt et al., 1986; Wallace, 2012). In addition, environmental changes, globalization, increased international trades, etc. are causing the scaling up of the movement of infectious diseases their dispersal over a wide geographical scale (Elad and Pertot, 2014; Hulme, 2009; Seherm and Coakley, 2003; Shaw and Osborne, 2011).

In general, when we talk about pathogen spread we highlight three basic factors as reproduced in the plant disease triangle in Figure: the susceptible host, the pathogen, and favorable environment:

- **Susceptible host:** the victim plant;
- **Pathogen:** certain microorganisms including fungi, nematodes, viruses and bacteria;
- **Favorable environment:** every pathogen need an adapted weather condition to thrive, such as specific humidity and temperature conditions;

If any of these factors is missing, i.e. the triangle is not complete, a pathogen cannot colonize a new spatial area. In addition to these factors, disease occurrence and severity depend on the time scale during which it is propagating. Time is a vector

along which the whole triangle moves. The amount of disease is defined by the surface of the triangle at any point along vector.

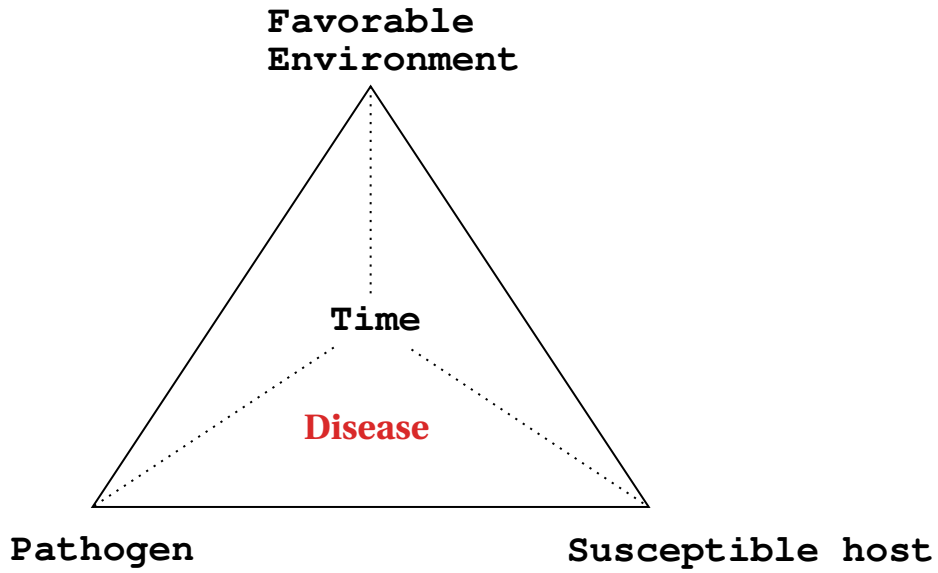


Figure 2.1. – The plant disease triangle: the intersection of the three factors (susceptible host, pathogen, favorable environment) approve the presence of the disease.

2.1.2.1. Dispersal routes

The transmission of infectious diseases from an infected host to a susceptible one occurs through many different routes. This latter can differ according to the populations, its cultural and economic situation, etc. Two main routes of transmission can be outlined: Direct and indirect.

- Direct transmission: infected seed, direct contact with infected planting material (bulbs, sets, cuttings, etc.);
- Indirect transmission: through intermediary that carry the pathogens such as water, animals, insects, wind, etc.

The understanding of the contribution of routes in disease transmission favors the identification of efficient strategies of prevention for pathogens that can be transmitted through these routes (Jeger et al., 2009). In the same vein, it will help to understand unanticipated infectious diseases and will allow management in the best possible way (Brett et al., 2017; Jones, 2004).

2.1.2.2. Airborne dissemination

Amongst the different known pathways of microbial pathogen diffusion, airborne diffusion is essentially associated with the dispersal of pathogens through the air via

small particles that can travel up to hundreds or thousands of kilometers. This dispersal can be found locally in a field where the disease is produced or it can extend to very long distances by crossing other countries, oceans, continents (Brown and Hovmøller, 2002; Mundt et al., 2009; Pady and Kapica, 1955; West, 2014). We can distinguish two types of airborne diseases spread: the direct transport of spores over long distances that can travel through a single step, or through step-wise transport gradually. The number of such diseases are increasing and have different forms such as viruses, bacteria, fungi. Airborne transmission can be affected by different factors such as the environment and socio-economic conditions. One of the strongest environmental causes is climate change. Due to climate change plant pathogens are showing a tendency in their geographic expansion since their comfort area is expanding (Das et al., 2017). This factor leaves an important impact on the spatiotemporal dynamics of the diseases since it implies a better climate for the emergence of certain types at certain moments. Furthermore, the socio-economic situation of the country might mark the propagation of the pathogens, since it is highly dependent on the level of investments in plant health.

Over time, different aerial pathogens have grown over the continental scale and have touched different types of plants (Brown and Hovmøller, 2002). The figure 2.2 shows the trajectories of some examples of diseases that have gathered importance throughout the centuries (quoted in table 2.2 with their history and characteristics).

Pathogens	Host plant	Symptom	Origin	Starting year	References
Peronospora hyoscyami	tobacco	blue mold	Florida, Georgia	1921	Johnson et al. (1989)
Puccinia striiformis	wheat	rust	Italy	1767	Ali et al. (2014)
Mycosphaerella musicola	banana	leaf spot	Indonesia	1902	Meredith et al. (1970)
Hemileia vastatrix	coffee	leaf rust	Sri Lanka	1868	Ward et al. (1882)
Powdery Mildew	wheat, soybeans, etc.	leaf and stem powdery spot	Wisconsin	1834	Pozzebon et al. (2009)
Cryphonectria parasitica	chestnut	Cankers	East Asia	1900	Rigling and Prospero (2018)

Table 2.2. – Examples of airborne plant pathogens and some of their characteristics: host plants, symptoms, spatial first detection, and year of detection.

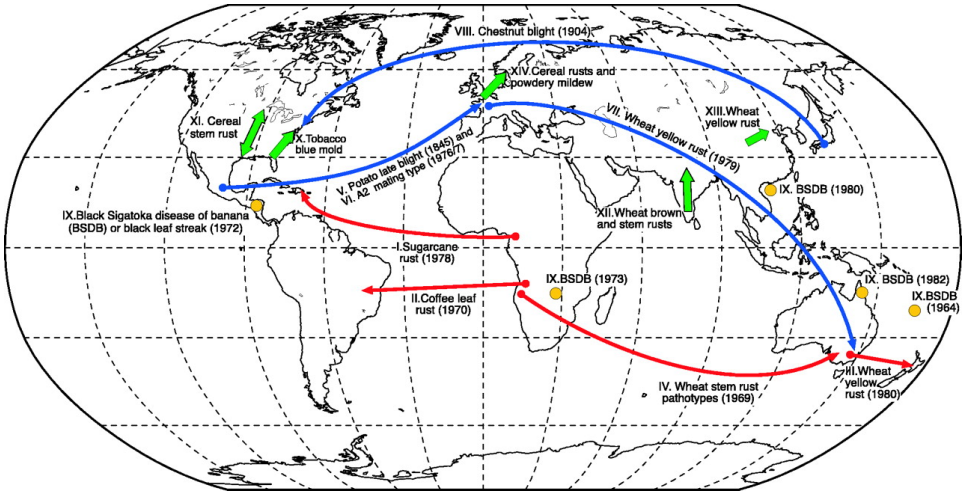


Figure 2.2. – Dispersal events of selected fungal pathogen as presented in Brown and Hovmøller (2002): Red arrows indicate the invasion of new territories probably occurred by the direct movement of airborne spores. Blue arrows indicate the invasion of new territories through humans or infected materials. Orange circles represent the worldwide spread of black Sigatoka disease of banana. Green arrows indicate periodic migrations of airborne spores in extinction-recolonization cycles.

2.1.2.3. Consequences of Long-distance dispersal

Controlling long-distance dispersed pathogens remains one of the most tough tasks since the LDD (long-distance dispersal) is over different and distinct populations (Jordano, 2017; Severns et al., 2019). LDD boosts the genetic connectivity which may spawn a change in the structure of genetic pools and can make a differentiation in the population by mutating (Garant et al., 2007). This differentiation can have a negative impact by aggravating the infection where the eradicated pathogen becomes more virulent (Flor, 1971). Besides the inter-pathogens effects, the plant diseases impact harshly human well being through economic and agricultural loss (Anderson et al., 2004) and ecosystem functions (Crowl et al., 2008). In a socio economic context, LDD has often had a wide-ranging impact (Smith et al., 2019). It affects the agricultural sector and the food animal production systems since those two sectors are dependant and in addition, it will affect humans. Over time, LDD can cause massive food shortages around different countries in the world which was followed by famines over the years such as the outbreak of the Brown spot fungus and the potato late blight disease. (Flood, 2010). Moreover, it can induce an economic crisis, particularly if the disease expands on a spatio-temporal scale such that the disease becomes a pandemic.

2.1.3. Epidemio-surveillance

The number of epidemics is multiplying very quickly which increases the challenge to overcome them (Cunniffe et al., 2014; Severns et al., 2019). To avoid the harmful effects of epidemics, it is vital to establish generic and adequate epidemio-surveillance plans in such ways that it can be adapted to different kinds of diseases. Surveillance can be defined as a continuous process of collecting, analyzing, interpreting, and communicating descriptive data for the monitoring of health events in a population. Epidemio-surveillance is the surveillance in an epidemiological context, it requires a careful lookout, and should be carefully designed in a spatial and temporal context.

Different generic options for epidemio-surveillance are defined by Kean et al. (2008):

- Early detection: detecting the disease sufficiently early, at the time of the invasion to a new spatial area, in order to thwart its spread;
- Delimitation: assessing the spatial extent and density of the pathogen in order to delimit the geographical infected area;
- Area freedom: demonstrating that the pathogen is absent from a given area.

All these options of surveillance share the same objective: minimizing the proliferation risk of the epidemic, but the most problematic one is the early detection because in this case, we don't have enough information about where, when, and how a new epidemic is likely to happen (Cunniffe et al., 2016; Parnell et al., 2015). Nonetheless, further investigation for this type of surveillance will definitely be helpful for reducing the consequent spread of diseases. For these reasons, we will be focusing on early detection surveillance.

2.1.3.1. Early detection of epidemics and their cost

Detecting epidemics in an early stage by adapting adequate surveillance strategies is essential for the control of dissemination and remains a challenge (Hashimoto et al., 2000). In a focal vision of epidemics, an epidemic spreads in a small area and gradually spreads widely (Van den Bosch et al., 1994). If the spread increases slowly, it will help to identify the sources of pathogens. This fact, at an early stage, will allow reducing the loss, since a small area is usually infected at this stage. Hence, it will lead to a decreased impact on the populations by raising public awareness about emerging disease threats (Heymann et al., 2001; Steele et al., 2016).

Generally, plant health managers are suffering from the lack of guidance concerning the appropriate level of early surveillance efforts. Hence, the surveillance effort is usually not sufficient and the epidemic is detected when it reaches high prevalence (Mastin et al., 2019b). However, in this case, the early detection seems to be excessive and the eradication of the hosts' plant population is less feasible (Parnell et al., 2015). In this context, trying to achieve early detection remains crucial to prevent the disease spread and to reduce its damages.

2.2. Networks

Networks are one of the approaches that are often used in the study of complex, dynamic, and interacting systems describing the relationships between entities or objects. Gradually, network theories are making major advances for the study of complex interacting systems in different research fields such as sociology (Carrington et al., 2005; Eagle et al., 2009), transportation (Gallotti and Barthelemy, 2015), biology (Ma and Gao, 2012; Yu et al., 2013), etc. This type of concept allows us to give a view of reality by quantifying the interaction through dynamic networks.

A network is defined as the portrayal of a graph with two essential components: nodes and vertices. Networks can be represented by vast interdisciplinary contexts such as nodes (as people, spatial regions, etc.) and edges will express the relationship between the nodes. It simplifies the mechanism without losing its essential characteristics.

2.2.1. Network or Graph?

Network and *graph* are often used as synonyms, but, in the scientific literature networks refer to the real systems (e.g. science collaboration networks, email networks, metabolic networks, etc.) whereas the graph refers to a mathematical representation (e.g. social graph, web graph, etc.). The technical terms are different, as it is shown in table 2.3 (Barabási et al., 2016).

Network Science	**Graph Theory**
Network	Graph
Link	Edge
Node	Vertex

Table 2.3. – Terminology used in *Network Science* and *Graph Theory*.

2.2.2. Graph theory

Graph theory was introduced by Leonhard Euler in 1735, alongside his famous problem "the bridges of the medieval city of Königsberg" (Alexanderson, 2006; Euler, 1953; Sachs et al., 1988). Königsberg is the historical name for the Prussian city that is now called Kaliningrad and located in Russia. This city is crossed by seven bridges as shown in figure 2.3a The intriguing question behind this city is the following: Is there a possible path that can cross all the bridges once? This question remains unanswered until it was solved through the mathematical proof of Euler. From the schematic illustration, Euler constructed a graph with four nodes $(1,2,3,4)$ and seven links as in fig 2.3b He proved that no continuous path can solve the problem by using the theory of graphs. The latter gives a manual approach of the problem by simplifying the representation and considering a unordered (ordered) pairs of set

of edges $E(G) = (e_1 \cdots e_M)$ connecting a non-empty set of components known as vertices $V(G) = (v_1 \cdots v_N)$ and forming a graph $G = (E, V)$ of order $N = \text{card}(V)$ and length $M = \text{card}(E)$. In the case of two nodes linked by edges, we can say that these nodes are adjacent. The resulting representation of the network is a graph with distinct points (vertices) linked by lines (edges). Mathematically, G is expressed by an adjacency matrix or connectivity matrix. It is a square matrix $N \times N$ given the index of connectivity between every couple of nodes in the network.

(a) (b)

Figure 2.3. – (a) The seven bridges of Königsberg during Euler's time, with the four corresponding patches of land $(1,2,3,4)$, (b) Graph constructed by Euler with the patches as nodes and seven links representing the bridges.

Different versions of graphs are defined (Newman, 2004):

- binary or weighted;
- undirected or directed.

If we assess the presence of a connection between two vertices and also the edges, then the connection between two nodes is evaluated by a binary variable. In this case, the adjacency matrix is made of binary terms and the term (i, j) $(i, j \in \{1, ...N\})$ is formulated as follows:

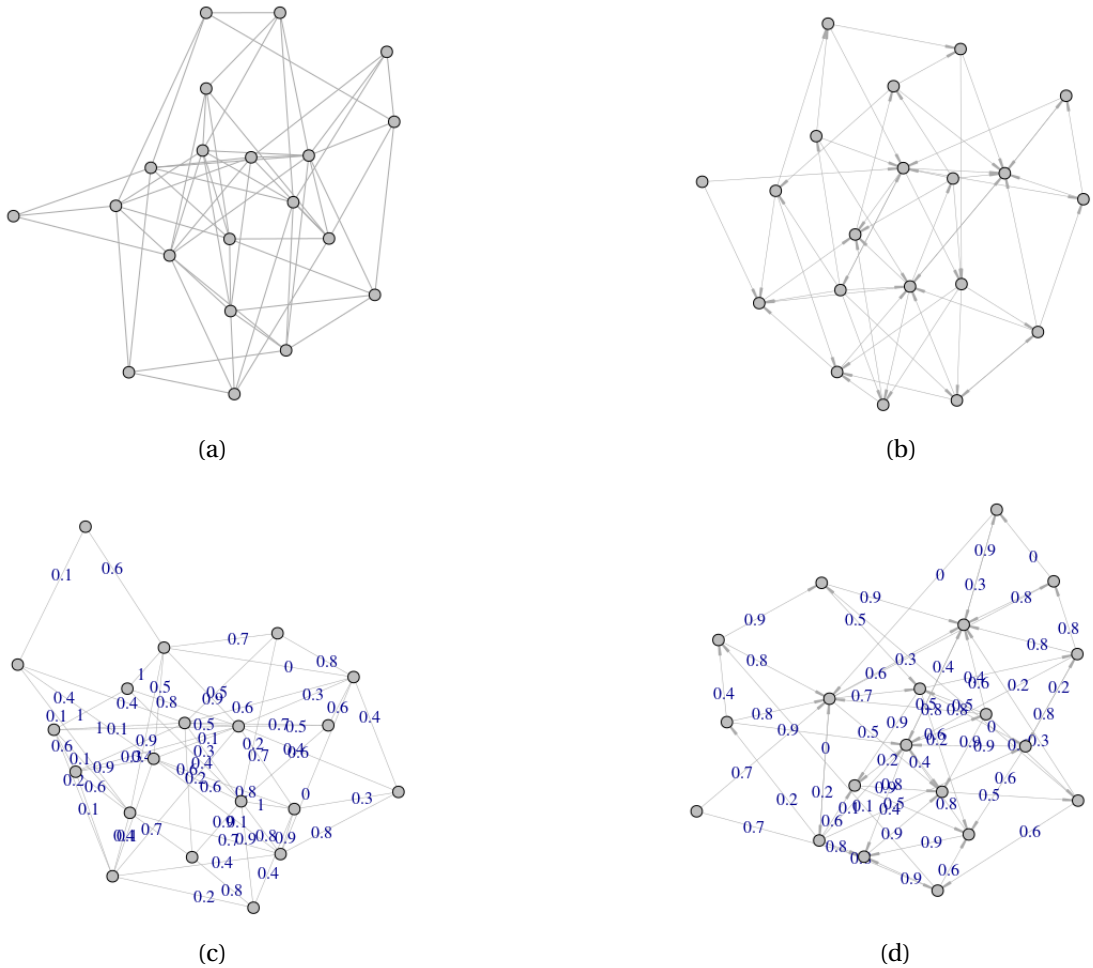
$$M_{ij} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

But if we assess the connection with a weight value, then the edge is defined as a 3-tuple $(v_1, v_2, w_{1,2})$, where $w_{1,2}$ is the weight of the connection between v_1 and v_2. The adjacency matrix for a weighted network is filled as follows:

$$M_{ij} = \begin{cases} a_{ij} & \text{if } \{v_i, v_j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

If v_1 and v_2 are linked by e_1, without a precise direction, then it is possible to go from v_1 to v_2 and vice versa. In this case, the network is undirected and the adjacency

matrix is symmetric ($a_{ij} = a_{ji}$). If a direction is imposed, then the network is directed and the matrix is non-symmetric. The direction is represented by an arrow.



(a) (b)

(c) (d)

Figure 2.4. – Different types of networks: (a) Binary and non-directed network, (b) binary and directed network, (c) weighted and non-directed network, and (d) weighted and directed network.

2.2.3. Classes of networks

The networks can be classified according to their structure in space and their agility and dynamism in time. In what follows, we recall some classes of networks that are then used in the heart of the thesis: spatial networks (Barthélemy, 2011), temporal networks (Holme and Saramäki, 2012) and spatio-temporal networks (George and Shekhar, 2008).

2.2.3.1. Spatial networks

A spatial network is defined as a graph where the nodes are located and embedded in a metric space as defined in Barthélemy (2011) and the metric space is equipped with a distance measure. Generally, space is the two-dimensional or three-dimensional Euclidean space with the usual Euclidean distance as a metric. Spatial networks are characterized by the topological information given by the adjacency matrix and the spatial one given by the position of the nodes.

These types of networks are popular and used in many fields to represent a spatial connection such as infrastructure networks, static transportation networks, computer networks, etc. (Barthélemy, 2011). Even though the literature on this class of networks is still relatively scarce, and there are still many unsolved questions related to these structures caused by the sometimes dichotomic information provided by the spatial embedding of nodes versus the network topology. Furthermore, many spatial statistical concepts, such as spatial autocorrelation, still need to be properly translated to the context of spatial networks.

2.2.3.2. Temporal networks

A temporal network is a network that changes its topology through time, by adding, removing, activating, or deactivating nodes and or edges (Holme and Saramäki, 2012). Each edge is characterized by its time of activation. If the network is weighted, the activation weight is defined by the weight at the activation time. The adjacency matrix will be a function of time $M(t)$ instead of a single static matrix. Since it changes through time, and at each time step, then, we can define this matrix by a dynamic adjacency matrix. The new ingredient of such type of networks is a variable, "*time*", treated as continuous or discrete according to the context of the study. They are characterized by their dynamicity across time (Holme and Saramäki, 2012).

These types of networks are popular in different fields such as disease spreading processes, information networks, neural networks, etc. Temporal networks are usually difficult to handle because of the amount of information that they involve, making their study rather challenging from a computational perspective (Kempe et al., 2002).

2.2.3.3. Spatio-temporal networks

Spatio-temporal networks are defined as a combination of spatial and temporal networks, having a spatial topology varying across time. At each temporal step, the snapshot of a spatio-temporal network (a.k.a. projected static network) is a spatial network.

These types of networks are important in the context of dynamic spatial networks such as traffic networks and communication networks (George and Shekhar, 2008; Williams and Musolesi, 2016). Of the three classes of networks presented here, this is by far the less studied in the literature and many research questions are still open.

2.2.4. Networks metrics

In order to analyze and compare the resulting networks, it is important to extract meaningful information from the complex networks. To unravel such information, that a network can contain, one can use network metrics. Two types of metrics can be identified: the ones which characterized the node in order to evaluate the relevant ones in a network by identifying the vital nodes (Lü et al., 2016), the second one is used mainly to characterize the interactions in the network (Boccaletti et al., 2006). Depending on the context, the interpretations of such metrics can differ.

In the following paragraphs, we summarize some of the most frequently used statistical metrics, in order to characterize and have a better understanding of this network (in terms of node importance and structure).

2.2.4.1. General network metrics

In order to evaluate and synthesize the whole topology of the network, different metrics have been defined in the literature such as the diameter (Boccaletti et al., 2006), the geodesic distance (Bouttier et al., 2003), the clustering coefficient (Serrano and Boguna, 2006), eigencentrality (Bonacich and Lloyd, 2001). These metrics are detailed in the table 2.4.

Metrics	**Description of the metrics**	**Relevance**	**Networks**
Diameter	Shortest distance between the two most distant vertices	Shortest the diameter, fastest the movements through the vertices in the network	*W* / *UW* *D* / *UD*
Geodesic distance	Minimum number of edges linking two nodes	Measure the path distance separating two vertices	*W* / *UW* *D* / *UD*
Clustering Coefficient, Transitivity	Measure of the degree to which nodes in a graph tend to cluster together	Higher the clustering coefficient, more the connections are dense	*W* / *UW* *D* / *UD*
Eigencentrality	Eigenvector of the adjacency matrix corresponding to the highest eigenvalue	Higher the eigencentrality, more the nodes is connectied to important nodes	*W* / *UW* *D* / *UD*

Table 2.4. – Definition and relevance of different network metrics. Networks refers to the type of network on which we can apply the metrics: *W* for weighted, *UW* for unweighted, *D* for directed and *UD* for undirected.

2.2.4.2. Node importance metrics

In large-scale complex networks, it is significant to extract the different characterization of the networks by identifying the vital nodes. Vital nodes and important nodes are the nodes with the greatest impact on the network structure (Bae and Kim, 2014;

Basaras et al., 2013; Boccaletti et al., 2006; Chen et al., 2012; Hou et al., 2012; Kitsak et al., 2010; Liu et al., 2013) and they are essentially characterized by centrality measures. It characterizes the nodes which have the potential to spread the information faster and vaster such as degree, strength (Lü et al., 2016), betweenness (Barthelemy, 2004; Brandes, 2001), closeness (Freeman, 1978), eigencentrality (Bonacich and Lloyd, 2001), voterank (Zhang et al., 2016), kshell (Kitsak et al., 2010). These centrality metrics are detailed in table 2.5.

Metrics	**Description of the metrics**	**Relevance**	**Networks**
Degree	Number of edges incident to the vertex	Higher the degree, the more central the node is	*W* / *UW* *D* / *UD*
In-Degree	Number of edges directed towards the vertex	Higher the in-degree, the more in-central the node is	*W* / *UW* *D*
Out-Degree	Number of edges directed away from the vertex	Higher the out-degree, the more out-central the node is	*W* / *UW* *D*
Strength	Sum of weights of edges connected to the vertex	Higher the strength, the more central the node is	*W* *D* / *UD*
In-Strength	Sum of inward edge weights	Higher the in-strength, the more in-central the node is	*W* *D*
Out-Strength	Sum of outward edge weights	Higher the out-strength, the more out-central the node is	*W* *D*
Betweenness	Number of shortest paths that pass through the vertex	Higher the score, more the vertex is in-between others	*W* / *UW* *D* / *UD*
Closeness	Sum of geodesic distances from the vertex to all others	Higher the score, shortest the distances to all other nodes	*W* / *UW* *D* / *UD*
VoteRank	Measure of the ability of each node to vote according to its neighbors	Identifying a set of decentralized spreaders with the best spreading ability	*W* / *UW* *D*
k-shell	Largest subgraph comprising nodes of degree at least k	Larger the k–shell index, more the node can act as a spreader in the network	*W* / *UW* *D* / *UD*

Table 2.5. – Definition and relevance of different node importance metrics. Networks refers to the type of network on which we can apply the metrics: *W* for weighted, *UW* for unweighted, *D* for directed and *UD* for undirected.

2.2.4.3. Degree distribution defining network's topology

As defined in table 2.5, the degree of a node is the number of connections with other nodes. So, the degree distribution is defined as the probability distribution of these degrees all over the nodes of the network. This metric provides a suitable

representation of the structure and apprises about the network topology. The network model is qualified according to which distribution it follows. Four general models of networks are defined in the literature: Regular, Random, Scale-free, and Small World networks.

- **Regular networks**: A regular network is two dimensional with non crossed links, each node is only connected to its adjacent nodes. In this type of network, all the vertices have the same degree, i.e. each edge has the same number of neighbors. The indegree and outdegree are as well equal. The regular network is characterized by a high clustering coefficient and high geodesic distance.
- **Random networks**: As its name implies, a random network as developed by Erdös and Rényi consists of a network construction that is truly random (Boccaletti et al., 2006). The distribution of the node's degree follows a bell-shaped and most of the nodes have approximately the same number of links (Barabási and Bonabeau, 2003). The degree distribution of this network is a binomial distribution and represents the probability that a node v_i has exactly k links (from $N-1$ potential links) that are independently connected or not with probability p or $(1-p)$; it is defined by:

  $$P(k) = \binom{n-1}{k} p^k (1-p)^{n-k}.$$

  This distribution is approximated as a Poisson distribution if the number of nodes is sufficiently large. The random network is characterized by a small clustering coefficient and small geodesic distance.
- **Small world networks**: The small world network (Watts and Strogatz, 1998) is defined by taking a regular network and rewiring randomly each link with a probability p (Newman, 2000). This will result in shortcuts which allow direct and rapid connection between distant nodes. Hence, we obtain a network with high clustering coefficients compared to a random graph and a short average path length (Watts and Strogatz, 1998). In chapter 3 of Barabási et al. (2016), it is predicted that the degree distribution of such networks will follow a Poisson-like bounded distribution. A small world network is characterized by a high clustering coefficient and small geodesic distance.
- **Scale free networks**: Scale-free networks are defined by Barabási and Albert (1999) and characterized by their highly heterogeneous degree distribution: most of the nodes have few link connections while few of the nodes have a high number of links. The distribution of the node's degree will follow a power-law distribution (Barabási and Bonabeau, 2003; Wang and Chen, 2003) represented as follows:

  $$P(k) = k^{-\gamma}$$

where γ is a positive parameter. Barabási and Albert (1999) specify that for a scale-free network the degree exponent is usually $2 \leq \gamma \leq 3$, while when $\gamma > 3$ the scale-free can be considered as a random network. These networks are characterized by the presence of large hubs.

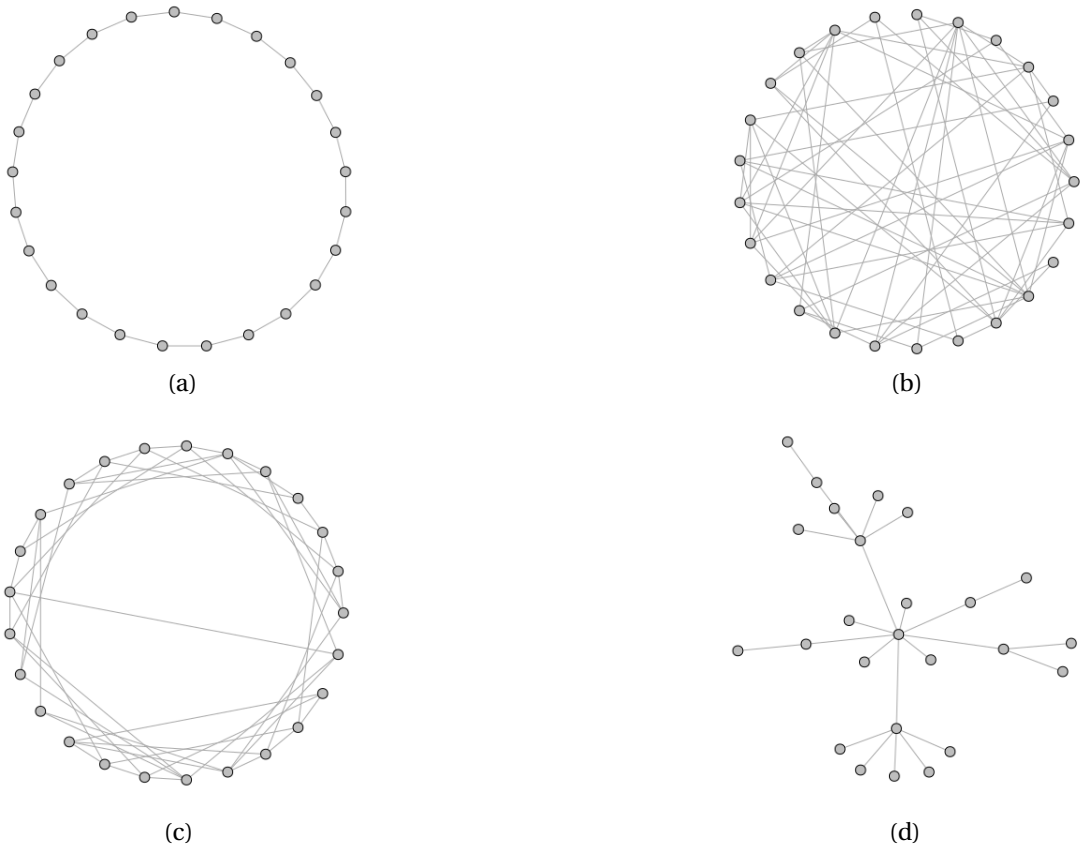


Figure 2.5. – Kinds of networks model: (a) Regular, (b) Random, (c) Small world and (d) Scale free.

2.2.5. Networks in plant epidemiology

Network approaches are becoming increasingly deployed in epidemiology. The interplay between these two fields provides a strong and flexible representation of the dynamic of the diseases within a population. By understanding and exploring the structure of this representation, one aims to monitor and control the invasion of diseases, as well as improve the predictions (Danon et al., 2011; Klovdahl, 1985; Miller and Kiss, 2014; Moslonka-Lefebvre et al., 2011; Pautasso and Jeger, 2014). This subject has been widely covered by the scientific literature, particularly in the context of the spread of infectious diseases between animals (Beaunée et al., 2015; Hamede

et al., 2009; Rautureau et al., 2011; Vernon and Keeling, 2009), humans (Balcan et al., 2009; Colizza and Vespignani, 2008; Wallinga et al., 1999) and plants (Jeger et al., 2007; Moslonka-Lefebvre et al., 2011; Pautasso and Jeger, 2014; Shaw and Pautasso, 2014). In what follows, I will restrict development to the latter context. Approaching networks in plant epidemiology can be done at different levels of plant pathology ranging from the gene interaction to the development of plant epidemics within a field or a landscape, and to the trade movements of infected plants across regions and even the world. According to the context, nodes could be proteins, cells, plants, nurseries, fields, or countries and the links generally define the potential transmission routes of the epidemics. The resulting network is commonly known as a contact network, representing the interactions between the individuals in a host population. One of the most central questions around these networks is about detecting the nodes that are the sources or the crucial nodes of the outbreaks (Eames et al., 2015; Pautasso and Jeger, 2014). To tackle this question, a considerable amount of work has already been done by using, for instance, the centrality metrics (Bell et al., 1999; Candeloro et al., 2016; Kitsak et al., 2010; Martinetti and Soubeyrand, 2019)and the network topology (Ganesh et al., 2005; Gang et al., 2005; Shirley and Rushton, 2005).

Take home messages

Advantages of graph theory:

- Networks derived by graph theory handle the computational problems of the relational aspect of the data in a practical, scalability, and effective manner.
- Networks are known for their ability to model different real-world situations (which could be complex with other tools) such as roadmap, World Wide Web, Facebook, Genetic interaction networks, etc.
- Through graph theory, we can represent the topology of networks in a natural way for many case studies and hence obtain easily interpretable results.
- To understand the networks, centrality measures are used as means for identifying the importance of the nodes in a network.

Limitations of graph theory:

- Time limitations: complex networks request a long time to find the neighboring nodes of a node of interest.
- Space limitations: complex networks request a large memory to store the graph.
- The different centrality measures, which are useful to evaluate the importance of the nodes, may disagree on the most important nodes in the network. Each measure quantifies a specific type of importance and, hence, ranking nodes may not be consistent.

2.3. HYSPLIT Trajectory Model

2.3.1. Description

The Hybrid Single-Particle Lagrangian Integrated Trajectory model (`HYSPLIT`) developed by the National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratory's (ARL) computes air parcel trajectories, complex dispersion, and deposition (Draxler and Hess, 1998; Stein et al., 2015). It has been evolving over the past 30 years; it began by estimating single simplified trajectories based on radiosonde observations. Nowadays, it is a system that computes multiple interactions between pollutants transported, dispersed, and deposited from local to global scales. This model is one of the most popular models in the field of atmospheric sciences primarily in atmospheric transportation and dispersion. A typical application of this

model is the analysis of backward air mass trajectories to track their sources and their relationship with the receptor. HYSPLIT can be used in different applications aiming to describe the dispersion and deposition of pollutants such as wildfire smoke, allergens, volcanic ash, windblown dust, etc.

The model calculation is a hybrid between the Lagrangian approach and the Eulerian one. The Lagrangian approach aims to follow the movement of a fluid parcel through space and time, so it can therefore be easily adapted to advection and diffusion calculation of the trajectories moving across space and time. While the Eulerian one consists of following the fluid flows through time in a specific spatial point, then it can be used to compute the pollutant air concentrations.

2.3.2. Applications

HYSPLIT is one of the most efficient models used for air parcel trajectory and dispersion calculations. Here are a few examples of applications taken from the literature: emission and transportion of pollen and allergens (Hernandez-Ceballos et al., 2014; Makra et al., 2010), forecasting volcanic ash (Chai et al., 2017; Hurst and Davis, 2017), transportation and deposition of air pollutants (Anastassopoulos et al., 2004; Shan et al., 2009), forecasting emissions of wind-blown dust events, transport, dispersion and deposition of such particles (Ashrafi et al., 2014; Wain et al., 2006), forecasting smoke wildfire (Rolph et al., 2009), transportation, dispersion and deposition of different types of radionuclides (Bowyer et al., 2013; Draxler and Rolph, 2012; Kinoshita et al., 2011).

2.3.3. Reconstruction of air mass trajectories

As stated above, the HYSPLIT model calculation is based on advection and diffusion concepts. The theoretical equations have been developed by Draxler et al. (1997), Draxler and Hess (1998) and Stein et al. (2015). In what follows, we will introduce the fundamental equations that have enabled the model to evolve.

Stein et al. (2015) detailed the calculation of a new position according to the advection by the wind at a time step $t + \Delta t$. The advection is calculated from the mean position of the tridimensional vector between the initial position $\mathbf{P}_{\text{mean}}(t)$ and the position of the first approximation $\mathbf{P'}_{\text{mean}}(t + \Delta t)$. This change in the position through time is defined by the following equations that are considered the basis of calculation of trajectories in HYSPLIT:

$$\mathbf{P}_{\text{mean}}(t + \Delta t) = \mathbf{P}_{\text{mean}}(t) + \frac{1}{2}\left[\mathbf{V}(\mathbf{P}_{\text{mean}}, t) + \mathbf{V}(\mathbf{P'}_{\text{mean}}, t + \Delta t)\right]\Delta t$$

where

$$\mathbf{P'}_{\text{mean}}(t + \Delta t) = \mathbf{P}_{\text{mean}}(t) + \mathbf{V}(\mathbf{P}_{\text{mean}}, t)\,\Delta t$$

and **V** is the average of the three-dimensional velocity vectors interpolated linearly in time and space (Draxler and Hess, 1998).

In order to calculate the concentration, it is important to know the distribution of the particles over the average travel of the trajectory. This can be done by adding a turbulent component to the advection velocity obtained from the meteorological data (Draxler et al., 1997; Fay et al., 1995; Stein et al., 2015). The dispersion of the particles horizontally and vertically may be represented as follows:

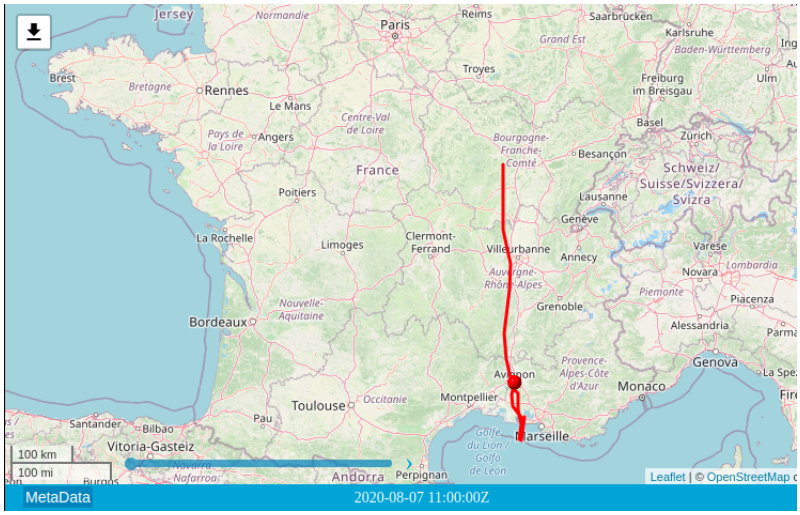
$$\mathbf{X}_{\text{final}} = \mathbf{X}_{\text{mean}}(t + \Delta t) + U'(t + \Delta t)\Delta t,$$

$$\mathbf{Y}_{\text{final}} = \mathbf{Y}_{\text{mean}}(t + \Delta t) + V'(t + \Delta t)\Delta t.$$

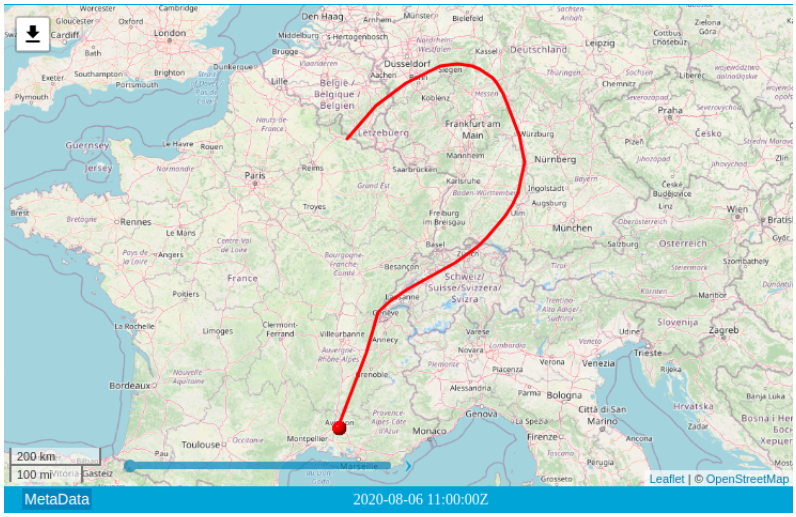
where $\mathbf{X}_{\text{mean}}$ and $\mathbf{Y}_{\text{mean}}$ are the mean components of the particle positions, $\mathbf{X}_{\text{final}}$ and $\mathbf{Y}_{\text{final}}$ give the final position in the horizontal and the vertical dimensions, U' and V' are the turbulent velocity components. Details of the theoretical calculation are provided in the supplementary material of Stein et al. (2015).

All these modeling equations have been converted to a set of different numerical algorithms coded in Fortran to form the software system `HYSPLIT` that takes as input the output of a meteorological model. In this work, we consider the `HYSPLIT` model applied to the meteorological data output from the Global Data Assimilation System files with a 0.5-degree spatial resolution (GDAS: `https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-data-assimilation-system-gdas`. This latter is used by the Global Forecast System (GFS) which is a weather forecast model developed by the National Centers for Environmental Prediction (NCEP).

A web version READY (Real-time Environmental Applications and Display System) has been developed to allow users to display meteorological data products. It also allows the HYSPLIT transport and dispersion model to be run on the NOAA Air Resources Laboratory's (ARL) server: `www.ready.noaa.gov`.



(a)



(b)

Figure 2.6. – 72-hours backward air mass trajectories, extracted from READY, arriving at Avignon on (a) August 6, 2020 and (b) August 7, 2020 at 11am.

2.4. Mathematical modeling in epidemiology

Over the past decades, mathematical models have become a paramount component of the understanding of the spread of infectious diseases in populations of hosts. Through the developed frameworks in this field of biology, science has progressed to the point where mathematical models facilitate development, allow construction, and the test of theories, also allowing ones to compare, plan, implement and evaluate various detection, prevention, therapy, and control programs for infectious diseases

(Hethcote, 2000). Epidemiologically speaking, and with the help of mathematical models, we can depict the spreading of pathogens among the hosts based on, e.g., type of contacts between the susceptible and infectious individuals of the populations; latency period during which the individual is infected and not infectious; duration of the infectious period; acquired immunity following infection, etc. After outlining these factors which are considered to be the parameters of the model, we would be able to estimate the expected number of infected individuals at each time step represented by an epidemic curve, a duration of the epidemic, and additional epidemic features (Grassly and Fraser, 2008; Huppert and Katriel, 2013; Roddam, 2001).

One often refers to the work presented by Ross in 1911 as a starting point of the epidemic modeling: Ross indeed proposed a framework based on differential equations to model the malaria transmission (Ross, 1911). Later on, based on several developments, Kermack and McKendrick introduced a mathematical paradigmatic and rigorous framework for modeling epidemics by introducing the compartmental disease models (Kermack and McKendrick, 1932). Thereafter, mathematical epidemiology grew exponentially, and to this day numerous models are formulated and analysed mathematically and then applied to various infectious diseases as reviewed in Becker (1979), Wickwire (1977), Hethcote (2000), Hethcote (1994) and Britton (2010). In what follows, we will review some fundamental concepts for such mathematical modeling.

2.4.1. Compartmental disease models

Mathematical modeling for describing the dynamics of infectious disease relies on the compartmental models that have been developed during the early 20th century by the famous work of Kermack and McKendrick (Kermack and McKendrick, 1932). The population is divided into compartments, which cluster the individuals having the same status according to the disease. Different types of compartmental models are defined based on the flow patterns between the individuals of the population, we can mention SI, SIS, SIR, SIRS, SEI, SEIS, SEIR, SEIRS, MSEIR, and MSEIRS as the most frequent models. They are generally represented by graphs as in Figure 2.7 where the nodes are labeled by the compartment: M, S, E, I, and/or R:

- Compartment M: individuals with passive immunity;
- Compartment S: susceptible individuals without passive; immunity;
- Compartment E: exposed individuals, which means infected but not yet infectious;
- Compartment I: infected individuals who have the capacity to transmit the infection;
- Compartment R: recovered or removed individuals.

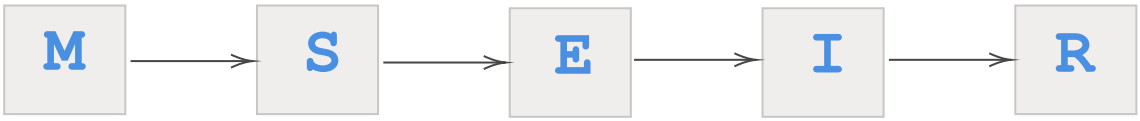


Figure 2.7. – Typical diagram of the connections between the different compartments in an MSEIR compartmental model. The arrows between the nodes represent the flow pattern between the compartments.

The compartments M and E are often omitted, because they may be considered, according to the objectives of the study, as non-fundamental compartments for the susceptible-infective interaction (Hethcote, 2000). The decision of including or omitting some compartments depends on the study, the characteristics of the disease as well as on the available data. We can identify two types of compartmental models: deterministic and stochastic, which will be described in detail below.

2.4.1.1. Deterministic *vs* stochastic

- A deterministic epidemic model is a model whose predictions are determined by its initial conditions and its fixed parameters. The transition between the compartments in this model is defined mathematically by derivatives (if the model is continuous in time), which results in a model based on differential equations. The population size in such models is assumed to be differentiable with respect to time. If the model is discrete in time, the differential equations are replaced by difference equations.
- A stochastic model depends on random variables or random processes, it is essentially used for phenomena that appear and vary in a random manner or that depend on processes which cannot be fully described in a deterministic manner (in the aim of keeping the model concise or handling our ignorance). The probability distribution of such a model depends on the random variation through the inputs over time, which expresses the dynamicity of the disease spread. Mathematically, these models may be based, for instance, on Markov chains or processes (Britton, 2010), semi-Markov processes (Soubeyrand, 2016) and piecewise-deterministic Markov processes (Abboud et al., 2018).

2.4.1.2. Deterministic compartmental model

The SIR model is one of the simplest and most used models, it is represented by the graph in Figure 2.8 (Capasso and Serio, 1978; Roberts et al., 2015). At any time, an individual of the population can be either in S, I, or R compartment. In the simplest form of the SIR model, the population is said to be *closed*, i.e. no birth, death, or immigration of individuals can occur during the defined period of study, the infected individuals can infect the susceptible ones, and then they recover until the end of the study period and become completely immune (Britton, 2010). Mathematically

speaking, this model is defined by a set of differential equations determining the size of each compartment over time:

$$\begin{cases} \frac{dS}{dt}(t) = -\frac{\beta I(t)S(t)}{N} \\ \frac{dI}{dt}(t) = -\frac{\beta I(t)S(t)}{N} - \gamma I(t) \\ \frac{dR}{dt}(t) = \gamma I(t) \end{cases} \tag{2.1}$$

where $t \in \mathbb{R}^+$, the derivatives $\frac{dS}{dt}(t)$, $\frac{dI}{dt}(t)$ and $\frac{dR}{dt}(t)$ are the rates of change of $S(t)$, $I(t)$, and $R(t)$, which are respectively the community fractions of susceptible, infective and recovered individuals such that case $S(t) + I(t) + R(t) = 1$. The starting configuration of this set of equations is generally defined by:

$$\begin{cases} S(0) = 1 - \epsilon \\ I(0) = \epsilon \\ R(0) = 0 \end{cases}$$

where $\epsilon > 0$ is the fraction of infected individuals at step $t = 0$, it is positive and generally small. At $t = 0$, $R(0) = 0$ is due to the fact that initially the individuals are assumed to be susceptible and not immune. More advanced developments on deterministic compartmental models can be found in Hethcote (2000), Daley and Gani (2001) and Anderson (2013).

Figure 2.8. – Diagram of a simple form of the SIR model.

To get infected, susceptible individuals must be in contact with infective individuals. Based on this assumption the non-linear term $\beta I(t)S(t)$ is derived. $S(t)$ decreases monotonically to $S(\infty)$, R(t) increases monotonically to $R(\infty)$ as represented in Figure 2.9.

To measure the capacity of the infectious disease to spread, an important variable is defined, namely the Basic reproduction number (Van den Driessche and Watmough, 2002):

$$R_0 = \frac{\beta}{\gamma}.$$

It gives the expected mean number of cases generated by one infected individual introduced in a susceptible population (Bailey et al., 1975). This variable is generally compared to the threshold value 1:

- $R_0 = 1$: an infectious individual can infect an average of one individual;
- $R_0 < 1$: on average, an infected individual can infect less than one susceptible individual; the outbreak is minor and cannot grow,
- $R_0 > 1$: on average, an infected individual can infect more than one susceptible individual; the outbreak is major until an important decrease of the susceptible component of the population; the higher R_0, the faster the spread of the disease.

The value of R_0 depends on the type and the contagion of the diseases and is determined by the system of equations 2.1 or its analogs in more complex settings. It also depends on the homogeneity of the community and the uniformity of the mixing between individuals.

Despite the possibility to include sophisticated assumptions in deterministic models, these models may not always fit the dynamics of the diseases of interest, in particular for small populations, when R_0 is low or close to one, when the initial phase of the outbreak is of interest, or when the changes in sanitary status at the individual level are of interest. In such situations, incorporating randomness into the model may be relevant. In the following section, we introduce such models known as stochastic models (Allen and Burgin, 2000; Brauer, 2008; Britton, 2010).

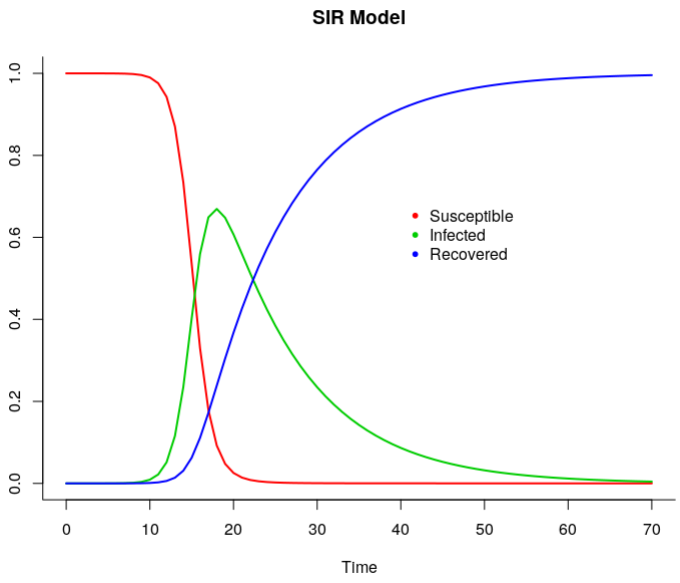


Figure 2.9. – Epidemic curves produced by the SIR model with $\beta = 1$ and $\gamma = 0.1$.

2.4.1.3. Stochastic compartmental model

A stochastic compartmental model shares the same concept with the deterministic one, but incorporates random effects. Here, we briefly introduce a classical stochastic SIR model. Let $S(t)$, $I(t)$ and $R(t)$ respectively represent the number of susceptible, infective and recovered individuals in a population of size N at time t such that $S(t) + I(t) + R(t) = N$. We suppose that at time $t = 0$, we have:

$$\begin{cases} S(0) = N - n \\ I(0) = n \\ R(0) = 0 \end{cases}$$

where n is the number of infected individuals at time $t = 0$. The dynamics of this model is defined through random transitions illustrated in Table 2.6:

- Susceptible-to-infectious transition: infectious individuals spread the infection to a susceptible individual with a constant rate parameter β,
- Infectious-to-recovered transition: an infected individual remains infectious for a random duration, i.e. the infectious period, after which it recovers and become immune to the disease, with a constant rate parameter γ.

Stage	Operation	Rate
Infection	$S_{t-1} \rightarrow S_t - 1$ $I_{t-1} \rightarrow I_t + 1$	$\beta \frac{S_t I_t}{N}$
End of infectious stage	$I_{t-1} \rightarrow I_t - 1$ $R_{t-1} \rightarrow R_t + 1$	γI_t

Table 2.6. – Transitions for a simple stochastic SIR model.

At the time $t = 0$, the starting time of the epidemic, a (typically small) number n of individuals are infected. As time goes by, the epidemic evolves according to the transitions defined above: susceptible individuals get infected and then recover until the end time of the epidemic period, i.e. until there are no more infected individuals in the population, which means that the epidemic has stopped.

Britton (2010) reviews two types of stochastic models according to the distribution of the infectious period. If this distribution is an exponential law with parameter λ, the model is called "stochastic general epidemic model" and the process $(S(t), I(t), R(t))$ is actually a Markov process whose jump-intensities are closely related to the deterministic epidemic model defined by the system of equations 2.1. The states in such a model are discrete and the time t is continuous. The second type of stochastic models reviewed by Britton (2010) corresponds to the case where the infectious period

is non-random and leads to the continuous-time version of the Reed–Frost model. In this latter model defined by Reed and Frost in 1928, each infected individual at a given *epidemic generation* infects independently each susceptible individual in the population with a probability p. The newly infected individuals constitute a new generation and the previous is removed from the epidemic process.

2.4.1.4. Epidemics on networks

Through networks, especially contact-networks, the connections between individuals can be explicitly represented. Out of that, networks are considered as a substrate where the epidemic spreads. So, the epidemic network models have been proved to be an efficient tool to understand the role of contact patterns between individuals in an epidemic context (Karrer and Newman, 2011; Keeling and Eames, 2005; Newman, 2002; Volz and Meyers, 2007; Youssef and Scoglio, 2011).

Newman (2000) proposed a formalism for epidemic models conditional on complex networks by defining the probability of transmission T_{ij} from an infected node i to a susceptible one j during a continuous-time through the definition of its complementary probability which is defined by:

$$1 - T_{ij} = \lim_{\delta t \to 0} \left(1 - r_{ij}\delta t\right)^{\frac{\tau_i}{\delta t}} = \exp^{-r_{ij}\tau_i}, \tag{2.2}$$

where r_{ij} is the average disease contact rate, τ_i the infectious period duration and δt the time-step. For discrete-time versions of the models, which are often considered for numerical simulations, Equation 2.2 is simplified by considering time step $\delta t = 1$ and, therefore:

$$1 - T_{ij} = \left(1 - r_{ij}\right)^{\tau_i} \tag{2.3}$$

where τ is measured in time-steps.

The quantities r_{ij} and τ_i are dependent on several disease-specific characteristics, the topology of the network, and the demography of the population. One generally needs to consider simplifications for r_{ij} and τ_i (in order to establish theoretical results or handle numerical issues) such as considering r_{ij} and τ_i as independent and identically distributed (Newman, 2002). From this starting point, a significant amount of work was developed within the framework of modeling disease epidemics on networks. Grabowski and Kosiński (2004); Yuan and Chen (2008); Zhou et al. (2015) work on the epidemic spread over social networks, Mishra and Jha (2010); Tang (2011); Yang and Yang (2014) presented different epidemic model over computer networks, Keeling (1999); Lang et al. (2018); Murray (2003) presented epidemic models on networks with spatial diffusion of the disease and Holme (2016); Holme and Saramäki (2012); Rocha et al. (2013) presented epidemic models in dynamic temporal networks. Beyond the modeling aspect of epidemics on networks, authors may be interested in networks metrics (such as those presented in Table 2.4) related to the impact of network topology on the diffusion of the disease, related to the detection of the influencing spreaders

(e.g., by using the centrality measures presented in Table 2.5), or related to the basic reproduction number that is dependent on the mean of the degree of all the nodes.

Take home messages

Advantages of mathematical modeling in epidemiology:

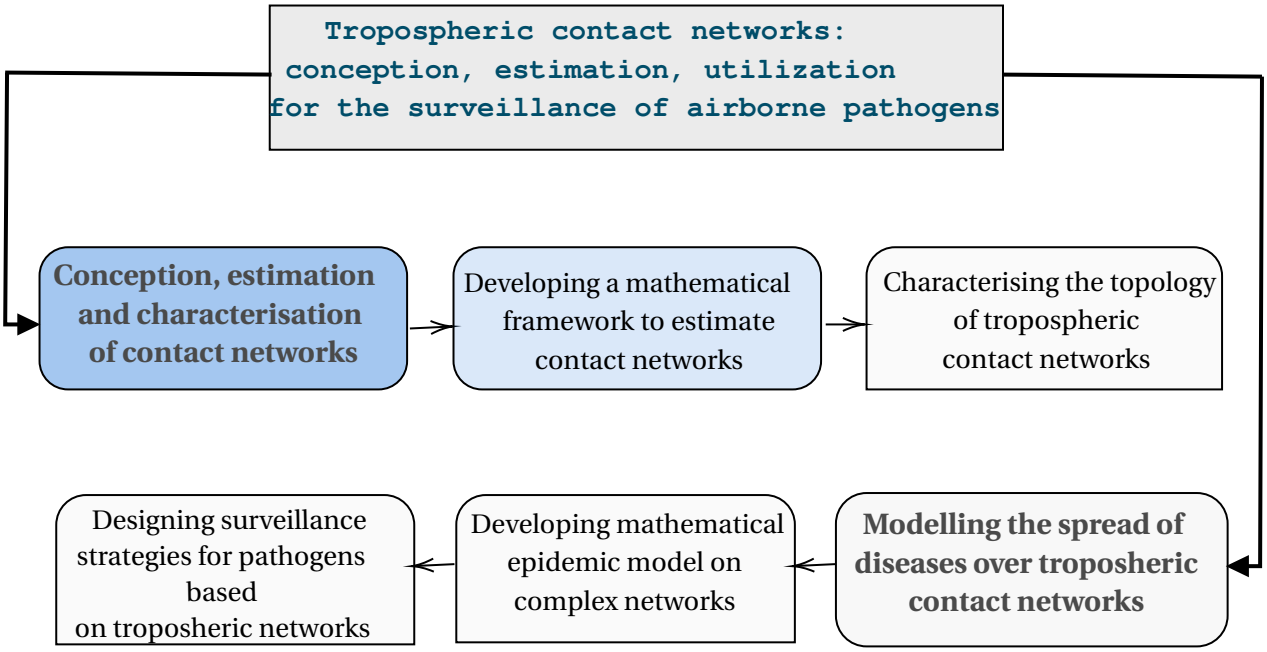
- Numerous mathematical models of the dynamics of infectious diseases have been proposed, grounded on diverse mathematical formalization.
- Mathematical models are used, in particular, to nowcast and forecast epidemics, as well as to evaluate, compare, and optimize theoretically the detection, the therapy, and the control programs.
- Mathematical models allow the estimation of epidemic measures such as epidemic threshold, reproduction number, and herd immunity threshold.
- Today, the output of mathematical models are widely used as support for decision making by worldwide authorities, as illustrated in the current COVID-19 pandemic.

Limitations of mathematical modeling in epidemiology:

- These models depend on different parameters, which can be arbitrarily fixed, taken from the literature or estimated from data, and model output and their uncertainty are obviously sensitive to the parameter values.
- Deterministic epidemiological models may not be appropriate when the infected population is small.
- The theoretical analysis of stochastic epidemiological models is complex when sophisticated components are included in the model, and the estimation of their parameters may be relatively difficult if a latent process is incorporated into the model.

3. Spatio-temporal large-scale networks shaped by air mass movements

Table of Contents



3.1. Abstract

The movement of atmospheric air masses can be seen as a continuous and generally complex flow of particles hovering over our planet. It can however be locally simplified by considering three-dimensional trajectories of air masses connecting distant areas. The connection can extend to thousands of kilometers, by crossing countries and continents and by carrying viral, bacterial, and even fungal diseases, insect pests, pollutants, etc. Thus, to survey and control airborne pathogens, pests, and pollutants, it is essential to study the pattern of air mass trajectories.

In this chapter, we present a mathematical framework based on the concept of graph theory whose aim is to conceive spatial and spatio-temporal networks where the nodes are the subsets of a partition of a geographical area, and the links between these nodes are inferred from sampled trajectories. The links, so-called integrated connectivities, between subsets are defined as integrals that cannot be analytically calculated but can be estimated from a sample of trajectories. We propose different estimators of link intensities relying on different bio-physical hypotheses and covering adjustable periods.

This approach leads to a new class of spatio-temporal networks characterized by adjacency matrices giving, e.g., the probability of connection via air mass movement between distant areas during a specified period. We use the term *tropospheric networks* to qualify these networks. We apply our representation framework to two real geographical contexts: the watersheds of the French region Provence-Alpes-Côte d'Azur and the coastline of the Mediterranean Sea. Then we estimate the corresponding spatio-temporal networks and perform an analysis of their properties. The key

output of this analysis allowed us to identify a marked seasonal pattern in air mass movements in the two study areas.

The networks constructed from air mass trajectories can be used to investigate issues, e.g., in aerobiology and epidemiology of airborne plant pathogens. Similar networks could be estimated based on other types of trajectories, such as animal trajectories, to characterize connectivity between different components of the landscape where the animals live.

3.1.1. Graphical outline

In this chapter, we answer the following questions by following the structure presented in the graphical outline 3.1:

- How to model, construct, and estimate contact networks from trajectory data?
- How can a tropospheric network be characterized?

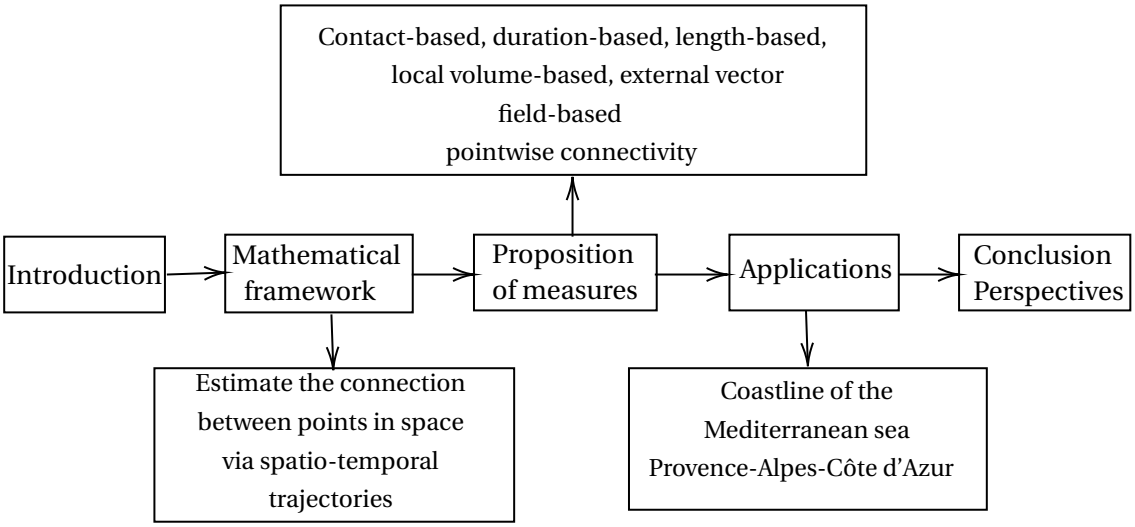


Figure 3.1. – Graphical outline of Chapter 3.

3.1.2. Status of the chapter

Research Article

- This article has been submitted.
- Authors of this article: Maria Choufany, Davide Martinetti, Cindy E. Morris, Rachid Senoussi and Samuel Soubeyrand.

3.2. Introduction

Atmospheric air masses are volumes of air with a defined temperature and water vapor content that have long been known to rule fundamental atmospheric phenomena like weather and air currents. Their composition is mostly inert gases, but both organic and inorganic particles have been found to linger in high-altitude air as a consequence of the constant interaction of air masses with the earth's surface below them. A non-exhaustive list includes gases and minerals like wildfire smoke, radioactive material, dust, sand, volcanic ash, and sea salt, but also living organisms such as pollen, fungal spores, bacteria, virus, and small insects. Despite the relative sparse density of these particles with respect to the volume of an air mass, their presence and transportation across the planet has proven to have strong effects on many phenomena impacting human health and safety (pollen Bogawski et al. (2019); Mahura et al. (2007); Šauliene and Veriankaite (2006), dust concentrations Aciego et al. (2017); Khaniabadi et al. (2017), nuclear byproducts Moroz et al. (2010); Rolph et al. (2014), human, animal and plant epidemics Aylor (1990); Hiraoka et al. (2017); Leyronas et al. (2018); Mundt et al. (2009); Sadyś et al. (2014); Wang et al. (2010), air pollution Liu et al. (2018a,c); Talbi et al. (2018), and rainfall Armon et al. (2018); Chen and Luo (2018); Rabinowitz et al. (2019)).

The rise in the number of publications on these subjects suggests a growing interest of the scientific community on the effects of air-mass movements on the biosphere, that has surely been boosted by recent available developments, such as the Hybrid Single-Particle Lagrangian Integrated Trajectory model (`HYSPLIT`, Stein et al. (2015)), allowing reconstruction of actual air-mass movements at rather fine geographical and temporal scales and with a global cover.

The vast majority of studies focused on isolated events, such as dust storms or peaks of air pollutants, that are rather concentrated in time (from few hours to few weeks) and/or space (just a few locations such as cities). Nonetheless, the movement of air masses is expected to have impacts on a broader spatio-temporal scale, as reviewed in recent studies Leyronas et al. (2018); Margosian et al. (2009). The purpose of the present paper is then to propose a mathematical framework for studying air-mass movements on large spatio-temporal scales, under the hypothesis that these movements can create stable and recurrent connections between distant portions of a territory. The very nature of these connections will be further specified throughout the manuscript, but as a general rule we will consider that any pair of points (or areas) in space can have a certain degree of connection, regardless of their geographic distance, provided that there are recurrent air-mass trajectories that connect the two points (or areas). The direction and strength of these connections will be estimated by looking at the trajectories linking every pair of points/areas and weighting them according to appropriate measures. In this perspective, it seems natural to resort to graph and network theory, since the formalism of nodes and edges provides an adequate environment for describing complex connections and can further be used

to deepen into the topology of the constructed networks in order to infer interesting properties of the graphs, such as the presence of hubs. From a generic statistical point of view, we aim to (i) estimate the weighted and directed edges of a graph using a sample of trajectories of *individuals* traveling through the space formed by the nodes of the graph, and (ii) characterize the estimated graph based on relevant statistics.

In the following sections, we first introduce the definitions and properties that will allow us to describe and then estimate connections between points/areas in space via spatio-temporal trajectories. Then, we propose several types of measures to model diverse types of connections. The expected output consists of a spatio-temporal graph describing the network of links induced by trajectories. It's worth noting that our approach is meant to infer connectivity induced by air-mass movements and it is readily applicable to `HYSPLIT`-type data, but we have maintained a sufficient level of generality to be applied to other phenomena, provided that trajectory data are available (e.g. animal trajectories). Finally, we apply our method to two case studies concerning the coastline of the Mediterranean sea and the French region of Provence-Alpes-Côte d'Azur. The two case studies have different spatio-temporal granularities and they will be used to provide examples of application of the proposed methodology.

3.3. Framework for the definition of trajectory-based networks

In this section, we show how a set of trajectories evolving within space during a finite time interval can be used to construct pertinent spatio-temporal networks. We first recall some basic definitions related to networks (Section 3.3.1) and then propose a statistical methodology to infer the network structure from a data set of trajectories (Section 3.3).

3.3.1. Network theory

Network theory (a.k.a. graph theory) is a mathematical formalism introduced by Leonhard Euler to describe the famous Königsberg bridge problem Newman (2003); Strogatz (2001); West et al. (1996). The two basic components of a network are a set of *nodes* linked by a set of edges. Nodes can represent a variety of things, such as persons, regions, computers, neurons, etc., while edges are used to describe the connections between those nodes. Formally, a *network* $G = (V, E)$ is defined as a set of *nodes* (or vertices) $V = \{v_1, v_2, \ldots, v_N\}$ connected by a set of *edges* $E = \{e_{ij}\}_{i,j \in \{1,\ldots,N\}}$. A natural way of representing a network is given by means of a $N \times N$ square matrix M, usually referred to as an *adjacency matrix*, whose term (i, j), M_{ij}, is non-zero an edge exists between i and j. By convention, adjacency matrices are defined to have an empty diagonal (i.e. $M_{ii} = 0, i \in \{1, \ldots, N\}$), meaning that nodes cannot be self connected. If M is symmetrical (i.e. $M_{ij} = M_{ji}, i, j \in \{1, \ldots, N\}$), then the network is said to be

undirected, and *directed* otherwise. If $M_{ij} \in \{0,1\}$, the network is said to be *binary*, meaning that an edge between two nodes i and j either exists or does not. Otherwise, if $M_{ij} \in \mathbb{R}$, the network is said to be *weighted*, meaning that the edge between nodes i and j are more or less connected.

In this paper, a network is said to be *spatial* Barthélemy (2014) when nodes correspond to geographic locations, while we use the term *temporal* Holme and Saramäki (2012) to refer to networks where edge values can change over time. Finally, we will use the term *spatio-temporal network* to refer to networks that are simultaneously spatial and temporal, under the constraint that nodes cannot change position, neither appear nor disappear over time. The networks considered in this paper also fall into the rather generic definition of spatio-temporal networks. If the spatial qualifier means that the nodes of the networks represent fixed geographical locations, the temporal qualifier is more complex. Indeed, temporal networks are generally divided in the literature into two main classes, namely contact graphs or interval graphs Holme and Saramäki (2012). The former type refers to networks where edges represent instantaneous contacts between nodes (Figure 3.2(a)), while in the second type edges are active over time intervals instead of instants of time (Figure 3.2(b)). In this paper, we propose a new definition of spatio-temporal networks where nodes correspond to disjoint regions of the space and edges are computed as a function of the flow of trajectories linking these nodes (Figure 3.2(c)), as it will be explained in the rest of the current section.

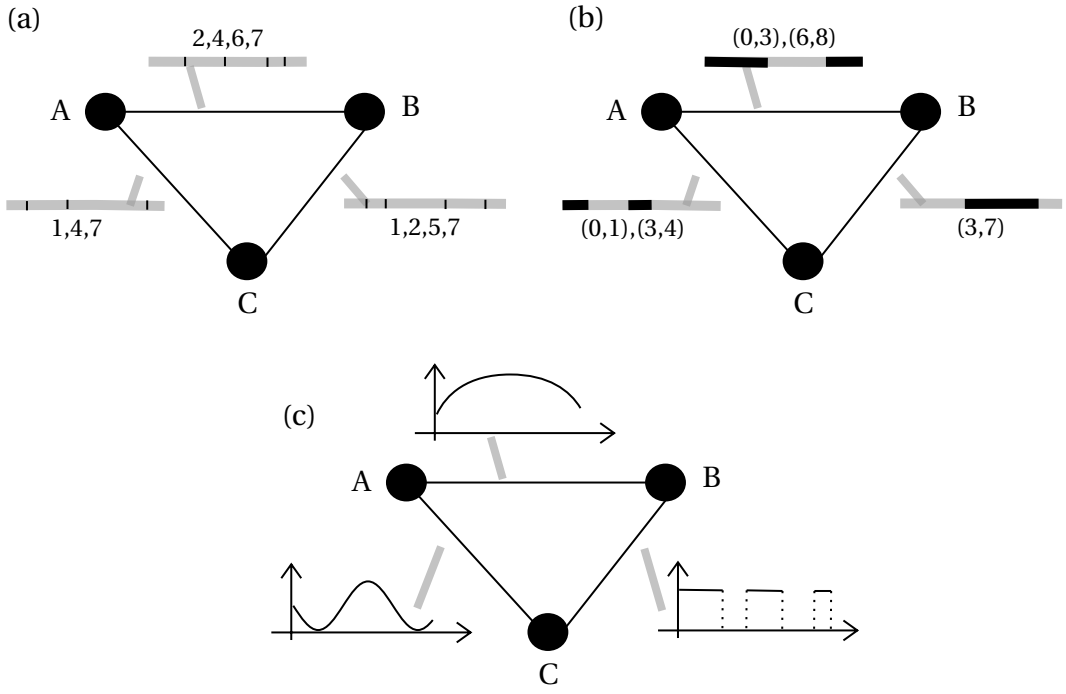


Figure 3.2. – Types of temporal networks. The time of activation is indicated within the grey bar next to the edges (ranging between 0 and 8). For contact networks (a), edges activate only for one instant at the time and are marked with black vertical lines inside the grey bars. For example, in panel (a), the edge between nodes A and B is only active at instants 2, 4, 6, and 7. For interval networks (b), edges can be activated during an interval of time. For example, the edge between A and B in panel (b) is active during the time intervals (0,3) and (6,8). For contact networks (c), the edges are quantitatively more or less active across time, and the quantity of activity of any edge is described by a temporal function.

3.3.2. Flows and trajectory segments

We consider a function $\Phi : \mathbb{R} \times \mathbb{R} \times \Omega \to \Omega$, usually called flow on the spatial domain Ω of $\mathbb{R}^d$, satisfying the following properties:

$$\begin{cases} \Phi(t,s,x) &= \Phi(t,t',\Phi(t',s,x)) \\ \Phi^{-1}(t,s,.) &= \Phi(s,t,.), \end{cases} \tag{3.1}$$

where $s,t,t' \in \mathbb{R}$ and $x \in \Omega$. For fixed t and s, the flow $\Phi(t,s,\cdot)$ is a spatial transformation. For fixed x and varying s or t, the function gives a forward or backward trajectory of a particle over Ω between times $t \wedge s = \inf(t,s)$ and $t \vee s = \sup(t,s)$. If $s \leq t$, $y = \Phi(t,s,x)$ gives the future location at time t of the particle presently located at x at time s. Contrarily, if $s \geq t$, $y = \Phi(t,s,x)$ gives the location at past time t that

was occupied by the particle located at x at present time s. $\Phi(t,s,.)$ is assumed to be a bijective mapping meaning that particles following distinct trajectories cannot be at the same location at the same time.

In general, a flow is defined with respect to a possibly time-dependent vector field F over $\mathbb{R}\times\Omega$, as the solution $u:\mathbb{R}\to\Omega$ of an ordinary differential equation (see e.g. Hamilton's equations in classical mechanics) with specified initial condition at a specified time s:

$$\begin{cases} \dfrac{du}{dt}(t) &= F(t,u(t)) \\ u(s) &= x, \end{cases} \tag{3.2}$$

where F is continuous and Lipschitzian over $\mathbb{R}\times\Omega$. In the setting introduced above, $\Phi(t,s,x)=u(t)$ with $\Phi(s,s,x)$=u(s)=x. The solution u represents the trajectory of the particle located at x at time s. Varying the initial condition in System (3.2), i.e. varying s and x, leads to consider pieces of trajectories of all particles which dynamics are governed by the vector field F. In this article, the vector field F will not be made explicit, but we will consider samples of trajectory segments (defined below) for constructing trajectory-based networks.

Definition 3.3.1 *The trajectory segment associated to the flow Φ over the time interval $\Delta_{ts}=[t\wedge s, t\vee s]$, $s,t\in\mathbb{R}$, for a particle located at $x\in\Omega$ at time s is defined as follows:*

$$\Gamma(t,s,x) \quad = \{(t',\Phi(t',s,x)) : t'\in\Delta_{ts}\}. \tag{3.3}$$

If $s<t$ (resp. $s>t$), $\Gamma(t,s,x)$ is a forward (resp. backward) trajectory segment. In this article, we are mainly interested in backward trajectories, but the framework presented here encompasses forward trajectories as well.

Example 3.3.1 *The notions of flow and trajectory segment can be adapted to cope with air mass trajectories over the Earth's surface. In this case, the spatial domain Ω representing the Earth's surface is the sphere $\mathbb{S}^2$ in $\mathbb{R}^3$. If in addition, air masses are characterized by altitude and temperature evolving in space and time, then $\Omega=\mathbb{S}^2\times\mathbb{R}_+\times\mathbb{R}$, where $\mathbb{R}_+$ (resp. $\mathbb{R}$) is the domain of the altitude (resp. temperature) coordinate.*

Example 3.3.2 *Animal movements and behavior activities can also be represented with the notions of flows and trajectory segments, providing, for instance, the animal locations and the covariate value indicating whether animals are feeding or not. In this case, $\Omega=\mathbb{R}^2\times\{0,1\}$, where 1 stands for 'the animal is feeding' and 0 otherwise. The use of a binary variable for describing the feeding activity may require the use of stochastic processes or generalized functions undergoing dynamic analog to the System (3.2) for constructing the flow if it is defined with respect to a vector field F.*

3.3.3. Pointwise and integrated connectivities

Trajectory-based networks are grounded on the notion of *connectivity* used as a quantitative, directed measurement of edges between graph nodes. In this aim, we first define the *pointwise connectivity* as a measure (or submeasure), in the mathematical sense, of the connectivity between a subset A and a point x of Ω induced by the trajectory segments $\Gamma(t,s,x)$ of a particle located at x at time s. Then, we use the pointwise connectivity to define the *integrated connectivity* between two subsets A and B of Ω over a temporal domain Δ of $\mathbb{R}$ (Δ can be the union of disjoint intervals).

Definition 3.3.2 *Let $x \in \Omega$ and $A \in \mathscr{B}(\Omega)$, where $\mathscr{B}(\Omega)$ is a σ-algebra of subsets of Ω. The pointwise connectivity associated to the flow Φ is defined as a real valued function Ψ on $\mathscr{B}(\Omega) \times \mathbb{R} \times \mathbb{R} \times \Omega$, conveniently denoted by $\Psi(A \mid t,s,x)$, where $A \mapsto \Psi(A \mid t,s,x)$ is a measure or a submeasure on Ω for each t,s,x*

Diverse types of pointwise connectivity can be constructed, either using trajectory segments generated by Φ, or directly using Φ. Specific pointwise connectivities can include environmental covariates and even covariates associated with very the movements of particles. Below, we give several examples of such specifications. Some of these examples are graphically represented in Figure 3.3. Most examples are particularly relevant when Ω is a simple geographic domain and when Φ defines movements of *individuals* (e.g., air masses, animals, or particles) within Ω.

Example 3.3.3 *The contact-based pointwise connectivity is defined by:*

$$\Psi_C(A \mid t,s,x) = \mathbb{1}_{\{\mathscr{A}_{ts} \cap \Gamma(t,s,x) \neq \emptyset\}}, \tag{3.4}$$

where $\mathscr{A}_{ts} = \Delta_{ts} \times A$ and $\mathbb{1}$ denotes the indicator function. $\Psi_C(A \mid t,s,x)$ indicates whether or not the particle whose movement in Ω is governed by $\Phi(\cdot,s,x)$ hit A during the time interval Δ_{ts}. Note that $A \mapsto \Psi_C(A \mid t,s,x)$ is only a submeasure on Ω since $\Psi_C(A \cup A' \mid t,s,x) \leq \Psi_C(A \mid t,s,x) + \Psi_C(A' \mid t,s,x)$ for disjoint sets A and A' of $\mathscr{B}(\Omega)$.

Remark 1 *This example based on the simple contact between sets can be considered as too strict from a statistical and measure-theory perspective since the length or the duration of a contact may be null. Instead, a positive constraint on contact length for example can be used to define another version of the contact-based pointwise connectivity: Equation (3.4) could then be replaced by*

$$\Psi_{\tilde{C}}(A \mid t,s,x) = \mathbb{1}_{\{\mathscr{L}(\mathscr{A}_{ts} \cap \Gamma(t,s,x)) > 0\}},$$

where $\mathscr{L}(\mathscr{A}_{ts} \cap \Gamma(t,s,x))$ denotes the length of the curve Γ within A. The length operator $\mathscr{L}$ will be made explicit in Example 3.3.5.

Example 3.3.4 *The duration-based pointwise connectivity is defined by:*

$$\Psi_D(A \mid t,s,x) = \int_{\Delta_{ts}} \mathbb{1}_{\{\Phi(v,s,x)\in A\}}\, dv, \tag{3.5}$$

to measure the duration spent by the particle in A during Δ_{ts}.

Example 3.3.5 *The length-based pointwise connectivity is defined by:*

$$\Psi_L(A \mid t,s,x) = \int_{\Delta_{ts}} \mathbb{1}_{\{\Phi(v,s,x)\in A\}} ||\nabla_v \Phi(v,s,x)||\, dv, \tag{3.6}$$

where $\nabla_v \Phi(v,s,x)$ stands for the gradient of the flow Φ with respect to the time variable v and $||\cdot||$ denotes the Euclidean norm. $\Psi_L(A \mid t,s,x)$ measures the distance traveled within A by the particle during Δ_{ts}.

Example 3.3.6 *The pointwise connectivity based on local volume is defined by:*

$$\Psi_V(A \mid t,s,x) = \int_{\Delta_{ts}} \mathbb{1}_{(\Phi(v,s,x)\in A)} |\det(J^x_\Phi(v,s)|\, dv \tag{3.7}$$

where $\det(J^x_\Phi(v,s))$ is the determinant of the Jacobian matrix (with respect to x) of the spatial transformation $\Phi(v,s,\cdot)$. The absolute value $|\det(J^x_\Phi(v,s)|$ of the Jacobian determinant at x gives the ratio by which the function $\Phi(v,s,\cdot)$ expands/shrinks infinitesimal volumes around location x into infinitesimal volumes around location $\Phi(v,s,x)$.

In other words, $\Psi_V(A \mid t,s,x)$ assesses how particle density increases or decreases from x to A along the time interval Δ_{ts}. Intuitively, if n particles are initially in A and if the infinitesimal volume around any of these particles tends to shrink from A to x, then one expects high concentration of particles in a fixed volume around x and, therefore, a high connectivity from A to x. Conversely, if the infinitesimal volume around a particle tends to expand from A to x, then one expects a lower concentration of particles in the same fixed volume around x and, therefore, lower connectivity from A to x.

More sophisticated specifications of the pointwise connectivity can be proposed by incorporating spatio-temporal covariates in its formulation, like in the following examples.

Example 3.3.7 *Let G denote a time-varying vector field defined over $\mathbb{R}\times\Omega$. The pointwise connectivity based on the external vector field G is defined by:*

$$\Psi_G(A \mid t,s,x) = \int_{\Delta_{ts}} \mathbb{1}_{(\Phi(v,s,x)\in A)} \mid <\nabla_v \Phi(v,s,x), G(v,\Phi(v,s,x))> \mid dv$$

where $<\nabla_v \Phi(v,s,x), G(v,\Phi(v,s,x))>$ is the scalar product between the gradient with respect to the time variable v of the flow Φ and the vector field G. Larger the average

collinearity in A between the instantaneous movement of the particle and the simultaneous direction of the vector field G, higher the connectivity between A and x. For instance, if Φ gives the movement of air masses and G provides the intensity and the direction of a continuous release of specific particles, then the connectivity will be high (resp. low) if the movement of the air in A and the movement of particles released in A are approximately collinear (resp. orthogonal).

Example 3.3.8 *Let Z and $\tilde{Z}$ be positive real valued spatio-temporal functions defined over $\mathbb{R} \times \Omega$. The pointwise connectivity based on Z and $\tilde{Z}$ is defined by:*

$$\Psi_{Z,\tilde{Z}}(A \mid t,s,x) = Z(s,x) \int_{\Delta_{ts}} \mathbb{1}_{(\Phi(v,s,x)\in A)} \tilde{Z}(v,\Phi(v,s,x))\, dv. \tag{3.8}$$

This form of pointwise connectivity may represent, for example, (i) the negative effect of the altitude of the air mass when it is above A on the recruitment of specific particles from the ground, and (ii) the positive effect of rainfall at (s,x) on the deposition of particles from the air mass to the ground (see Figure 3.4). Thus, the lower the average altitude of the air mass above A and more intense the rainfall at (s,x), the larger the contribution to the connectivity between A and x. This is expressed in Equation (3.8) as follows: (i) $\tilde{Z}$ is defined as the binary function indicating whether or not the altitude of the air mass (located at x at time s) is lower than a threshold h when it is located at $\Phi(v,s,x)$ at time v; (ii) Z is a function of the local rainfall intensity at (s,x).

Remark 2 *If in Example 3.3.8, the altitude of the air mass is incorporated as the third coordinate of Φ and A is a 3D-domain vertically limited by the threshold value h, then, Equation (3.8) is simplify reduced to Equation (3.5).*

Remark 3 *Example 3.3.8 could be generalized by considering a measure, say μ, over $\mathbb{R}$, to handle the potential contribution of discrete-time events to the pointwise connectivity:*

$$\Psi_{Z,\mu}(A \mid t,s,x) = Z(s,x) \int_{\Delta_{ts}} \mathbb{1}_{(\Phi(v,s,x)\in A)} \tilde{Z}(v,\Phi(v,s,x))\, d\mu(v). \tag{3.9}$$

Remark 4 *In the same vein, Example 3.3.8 can also be modified by adding within the integral the term $||\nabla_v \Phi(v,s,x)||$ arising in Equation (3.6) to account for a supplementary effect of the distance travelled within A on the pointwise connectivity.*

Each pointwise connectivity defined above can be used for defining the integrated connectivity, which measures the quantitative directional link between two subsets A and B of $\mathscr{B}(\Omega)$ generated by trajectories of particles located in B at times belonging to the temporal domain T.

Definition 3.3.3 *Let A and B be two sets of $\mathscr{B}(\Omega)$ and T a subset of the temporal domain $\mathbb{R}$. The δ-lag integrated connectivity linking B to A over T is defined by:*

$$\Psi^{(2)}_{\nu,\delta}(B \times A \mid T) = \int_{T\times B} \Psi(A \mid s+\delta, s, x)\nu(ds, dx), \tag{3.10}$$

where $\delta \in \mathbb{R}$ and ν is a measure on $\mathbb{R} \times \Omega$.

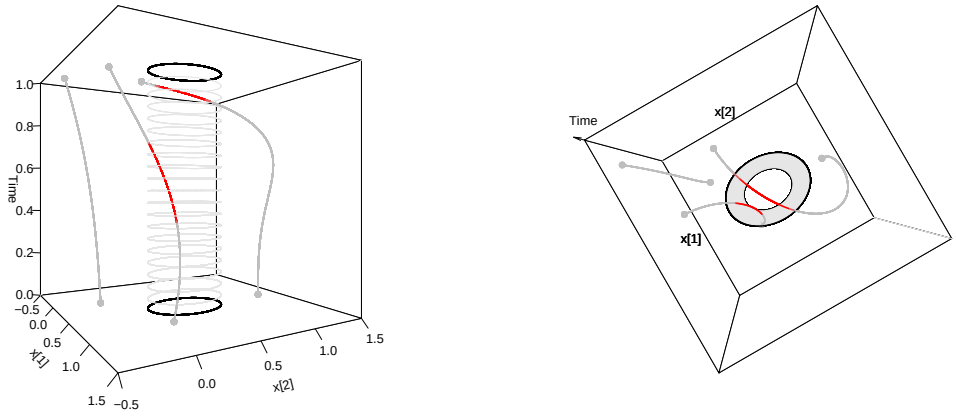
Definition 3.3.3 encompasses connectivities generated by either forward or backward trajectories, depending on the sign of δ. The use of a unique duration $|\delta|$ could be relaxed to account for space-time heterogeneities in the duration of trajectories. It could even be infinite by introducing a measure over time like in Equation (3.9).

The measure ν in Definition 3.3.3 can be continuous, discrete or hybrid over $\mathbb{R} \times \Omega$. Indeed, if particles of interest are air masses, then B can be considered as continuously filled in space and time. Conversely, if particles of interest are animals of a specific species, then animals occupy only punctual locations in B at each time and the measure $x \mapsto \nu(ds, dx)$, given s, is discrete in Ω, whereas the temporal component of ν is continuous. Another example occurs when the time s corresponds to death times of animals, then ν is both discrete in space and time with a mass only at a countable collection of space-time points.

3.3.4. Trajectory-based network

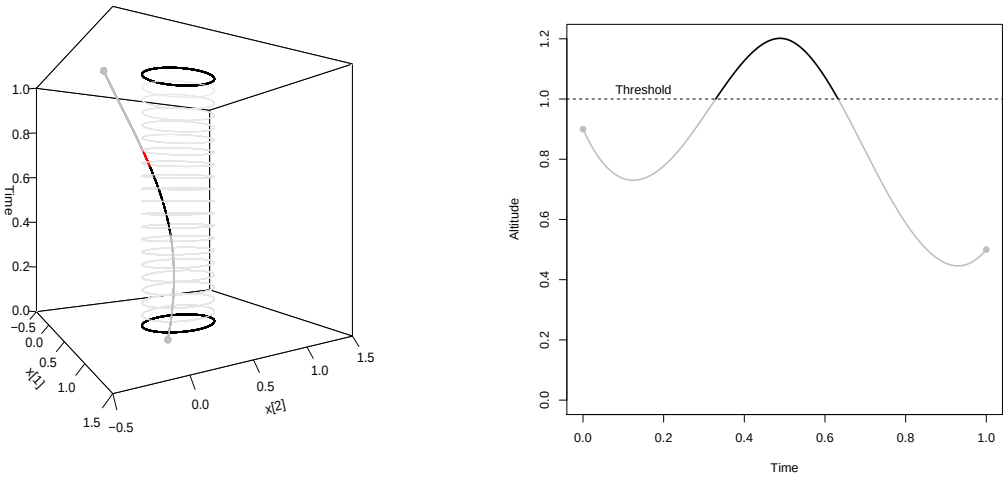
Definition 3.3.4 *A trajectory-based network generated by $\Psi^{(2)}_{\nu,\delta}$ (given by Equation (3.10)) over the temporal domain $T \subset \mathbb{R}$, is a graph whose nodes $A_i, i = \{1,..,I\}$, are disjoint sets of Ω in $\mathscr{B}(\Omega)$ and whose directed edges are weighted by integrated connectivities $M_{ij} = \Psi^{(2)}_{\nu,\delta}(A_i \times A_j \mid T)$, $1 \le i, j \le I$ and $i \ne j$.*

Definition 3.3.4 corresponds to a spatial trajectory-based network evaluated over the fixed temporal domain T. It can be extended in different ways to obtain spatio-temporal analogs. For example, if $T_1, \ldots, T_K$ denote K disjoint but successive time intervals with equal lengths, then the sequence of trajectory-based networks generated by $\Psi^{(2)}_{\nu,\delta}(\cdot \times \cdot \mid T_k)$, $k = 1, \ldots, K$, forms a spatio-temporal trajectory-based network that can be analyzed to assess how connectivities across space are changing with time. This is one of the issues considered in Section 3.5.2.



(a) (b)

Figure 3.3. – Illustration of contact-based, duration-based and length-based pointwise connectivities (resp. Ψ_C, Ψ_D and Ψ_L) between the elliptic spatial domain $A \subset \mathbb{R}^2$ and different spatial points x at time $s = 1$, for $\Delta_{ts} = [0,1]$. The left curve on panel (a) never enters the domain A. The middle curve on panel (a) enters A (red part of the curve) over a relatively long duration (as shown by panel (a)) but a short distance (as shown by panel (b)). The right curve on panel (a) enters A over a shorter duration but a longer distance. Thus, $\Psi_C(A \mid t,s,x)$, $\Psi_D(A \mid t,s,x)$ and $\Psi_L(A \mid t,s,x)$ are zero for the left curve; $\Psi_C(A \mid t,s,x) = 1$ for the two other curves; $\Psi_D(A \mid t,s,x)$ is larger for the middle curve than for the right one, whereas $\Psi_L(A \mid t,s,x)$ is larger for the right curve than for the middle one.



(a) (b)

Figure 3.4. – Illustration of pointwise connectivity based on a covariate measured along the trajectory (see Example 3.3.8). In this illustration, the passage of the particle in the elliptic spatial domain A contributes to the pointwise connectivity (red part of the curve in panel (a)) only when the particle is at an altitude lower than a threshold value (the grey part of the curve in panel (b)).

3.4. Estimation of integrated connectivities

In practice, the integral defining the integrated connectivities between subsets of Ω (Definition 3.3.3) cannot be analytically computed in general, but can be estimated from a sample of trajectories. For instance, the integrated connectivity $\Psi^{(2)}_{\nu,\delta}(B \times A \mid T)$ can be estimated by its empirical counterpart obtained by importance sampling, say $\widehat{\Psi}^{(2)}_{\nu,\delta}(B \times A \mid T)$:

$$\widehat{\Psi}^{(2)}_{\nu,\delta}(B \times A \mid T) = \frac{|T|}{N}\frac{|B|}{N'}\sum_{k=1}^{N}\sum_{l=1}^{N'}\Psi(A \mid s_k+\delta, s_k, b_l), \tag{3.11}$$

where $s_1,\dots,s_N \in T$ and $b_1,\dots,b_{N'} \in B$ are times and locations, respectively, randomly drawn under the measure ν restricted to $T \times B$, $|T|$ and $|B|$ are the length and area of T and B, respectively, and $\Psi(A \mid s_k+\delta, s_k, b_l)$ is the pointwise connectivity associated to the trajectory of the particle located at b_l at time s_k and observed over $\Delta_{s_k+\delta,s_k} = [s_k \wedge s_k+\delta, s_k \vee s_k+\delta]$.

If ν is constant, other classical numerical approaches can be applied to approximate the integral, such as a hybrid approach in which the mid-point rule is applied in time

and a regular point process is used in space. In such a case, the integrated connectivity estimator is also given by Equation (3.11).

Example 3.4.1 *Using Equation* (3.11)*, the contact-based pointwise connectivity in Example 3.3.3 is estimated by:*

$$\widehat{\Psi}^{(2)}_{C,\delta}(B \times A \mid T) = \frac{|T|}{N}\frac{|B|}{N'}\sum_{k=1}^{N}\sum_{l=1}^{N'} \mathbb{1}_{\{\mathscr{A}_{s_k+\delta,s_k} \cap \Gamma(s_k+\delta,s_k,b_l) \neq \emptyset\}}, \tag{3.12}$$

where $\mathscr{A}_{s_k+\delta,s_k} = \Delta_{s_k+\delta,s_k} \times A$*. Thus,* $\widehat{\Psi}^{(2)}_{C,\delta}(B \times A \mid T)$ *is simply the proportion of sampled trajectories intersecting A, multiplied by* $|T||B|$*.*

Example 3.4.2 *Using Equation* (3.11)*, the duration-based pointwise connectivity in Example 3.3.4 is estimated by the average duration of the intersections between the sampled trajectories and A, multiplied by* $|T||B|$*.*

Example 3.4.3 *Using Equation* (3.11)*, the length-based pointwise connectivity in Example 3.3.5 is estimated by the average length of the intersections between the sampled trajectories and A, multiplied by* $|T||B|$*.*

3.5. Applications

In this section, we applied our general framework to the flow of air mass movements. Indeed, these movements compiled over years were used to characterize climatic patterns Hondula et al. (2010) and to describe the transport of pollutants Pérez et al. (2015). We show now how to deploy our approach for constructing air-mass movement networks in two real geographical contexts, namely the coastline of the Mediterranean Sea and the French region of Provence-Alpes-Côte d'Azur. These two examples have been chosen in order to prove the flexibility of our approach to different situations and geographical scales.

3.5.1. Case study regions and network construction

The first study region corresponds to the coast of the Mediterranean Sea, ranging approximately 1,600 km from north to south and 3,860 km from east to west. The temperate climate of the chosen region is strongly influenced by the presence of the Mediterranean Sea, with mild winters, hot summers, and relatively scarce precipitations events. The landscape is characterized by coastal vegetation, typically shrubs and pines, and densely populated areas with intensive crop production of wheat, barley, vegetables, and fruits, especially olive, grapes, and citrus. In this paper, we characterize recurrent movements of air masses through the Mediterranean region by defining a grid with mesh size 74 km covering the coastline from 5 km up to 250

km inland from the coast, including the four largest islands (namely Sicily, Sardinia, Cyprus, and Corsica). Thus, we divide the region into 604 cells, where the centroids of the cells will be used as arrival locations of air-mass trajectories and will correspond to the nodes of the constructed network.

The second study region corresponds to the French region of Provence-Alpes-Côte d'Azur (PACA, hereafter), located in the south-eastern part of France and characterized by a rather complex landscape formed by a densely-populated coastline, agricultural lands (high-value-crops with fruit and olive orchards, vineyards, vegetable cultivation, and horticulture), and natural mostly-alpines areas. The choice of this particular region is justified in the context of a research project aimed at assessing the potential long-distance dissemination of phytopathogenic bacterial populations that are known to be transported by air currents. The bacteria of interest (e.g., *Pseudomonas syringae*) can be lifted in to the air from a source location and then be passively transported by air masses until they are deposited back to land onto a different, far away sink location. Since the life cycles of the considered species of bacteria are strongly linked to the water cycle Morris et al. (2008), we naturally partitioned the study area in a way that fit this assumption. Hence, we considered the 294 watersheds of the PACA region to define the sites that will later constitute the nodes of the constructed network. Since watersheds have irregular shapes and varying sizes, we selected a certain number of arrival locations per watershed (between 1 and 10 and proportionally to the watershed area) in order to cover the watersheds consistently according to the relative importance of their size and estimate the integrated connectivities. In total, a set of 833 arrival locations for air-mass trajectories was generated.

Once the arrival points for the two study regions have been established, we turned to the computation of air-mass trajectories arriving at the prescribed locations using the Hybrid Single-Particle Lagrangian Integrated Trajectory model (`HYSPLIT`, Stein et al. (2015)). The `HYSPLIT` model can be fed with meteorological data from the Global Data Assimilation System files with a 0.5-degree spatial resolution (GDAS [1]) and was tuned by us to return 48-hours backward air-mass trajectories arriving at the prescribed locations at an altitude of 500 m above mean sea level. A single trajectory consists of a vector containing the hourly positions (longitude, latitude, and altitude) visited by the air mass before arriving at the specified location and time. Air-mass trajectories have been computed for every arrival location (604 for the Mediterranean region and 833 for the PACA region) and for every day between January 1, 2011, and December 31, 2017 (arrival hour is 12:00 GMT). The total number of computed trajectories is 1,543,220 for the Mediterranean region and 2,128,315 for the PACA region.

The final step for the construction of the networks is the estimation of the adjacency matrices of the networks, based on the methodology presented in the previous sections. To do that, for each pair of subsets of the spatial domain, we used the daily 48-hours backward trajectories arriving at the locations sampled within the receptor

1. https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-data-assimilation-system-gdas

subset and computed the contact-based estimator (see Example 3.4.1). The subsets of the spatial domain are the watersheds for PACA and circular buffers of radius 20 km for the Mediterranean region, as in Leyronas et al. (2018)

In this work we will consider networks corresponding to three temporal contexts: (i) the spatial networks obtained when T is the entire period 2011-2017, (ii) the yearly spatio-temporal networks formed by the seven spatial networks obtained when T_1 encompasses the year 2011, T_2 encompasses 2012 and so on, and (iii) monthly spatio-temporal networks formed by the twelve spatial networks obtained when T_1 represents every January from 2011 to 2017, T_2 every February from 2011 to 2017, and so on. In all these cases, we consider that the length of the time interval was 1 to easily compare the inferred networks (i.e., $|T| = 1$ in Equations (3.11) and (3.12)).

3.5.2. Network analysis

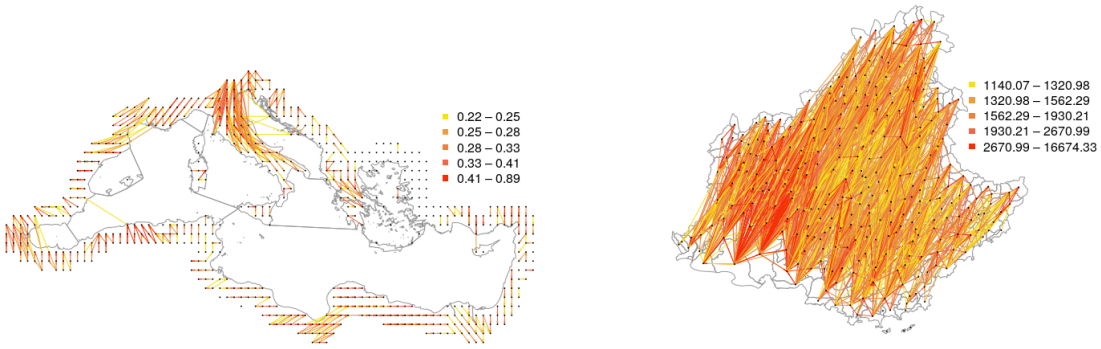
The constructed networks are directed and weighted by contact-based connectivities generated by air mass trajectories. They are inherently complex by the sheer amount of spatial and temporal information that they encompass. Hence, there is no easy way of representing the results either graphically or numerically, without compromising the original complexity of the networks. While a comprehensive physical study of the spatio-temporal properties of these networks goes beyond the scope of the paper, we explore the estimated trajectory-based networks by looking at some generic properties through the following indices:

- Diameter: the longest of all possible shortest paths between any two pairs of nodes.
- Density: the ratio between the sum of all edge weights and the number of all possible edges Liu et al. (2009).
- Transitivity (also known as clustering): the equivalent definition of density, but applied to triplets of nodes instead of pairs of nodes Opsahl and Panzarasa (2009).
- Shortest path: characterized by the average and standard deviation of the computed shortest path between any possible pair of different nodes Newman (2001).
- Small worldness: refers to the property of a network of being highly clustered and having relatively short shortest paths. It is computed as the ratio between the normalized clustering and the normalized average shortest path distance Colon-Perez et al. (2016); Li et al. (2007).
- Scale-free property: The degree of a node in terms of the total number of edges entering and exiting from it, and for directed networks, it can be decomposed in the incoming and outgoing degree, respectively. The degree distribution is the empirical distribution of the degree of a network and it said to be scale-free when it approximately follows a power-law distribution, i.e. $P(k) \sim k^{(-\alpha)}$, where $P(k)$ represents the probability of a node having degree equal to k Barabási and

Albert (1999); Barabási and Bonabeau (2003). Some authors impose that the α parameter of the power-law distribution has to fall within the interval $[2,3]$ Barabási et al. (2016). Thus, a network is scale-free when most of its nodes have a low degree, while the probability of having nodes with a very high degree is not negligible (fat right tail of the distribution). Nodes with very high degrees play a crucial role in dynamics conditional on networks and are often referred as hubs Liu et al. (2011).

- Degree correlation: in directed networks, it accounts for the correlation between the incoming and the outgoing degree of a node. Networks with positive (resp. negative) degree correlation foster (resp. hamper) epidemic spread Pautasso et al. (2010).

3.5.3. Results



(a) (b)

Figure 3.5. – Networks weighted by contact-based connectivities generated by air mass trajectories between (a) the 604 sampled circular areas within the Mediterranean basin and (b) the 294 watersheds of PACA. Edges with weights lower than 0.3 for (a) and 2×10^3 for (b) are not drawn. The cuts of the intervals in the two legends are chosen in such a way that each interval contains 20% of the observed data. The differences in the values taken by the connectivities in (a) and (b) are due to different measures of the area $|B|$ in Equations (3.11) and (3.12): $|B| = 1$ for each node in (a), whereas B is the actual area (expressed in km^2) of each whatershed in (b).

The two spatial trajectory-based networks representing the strength of tropospheric connections in the Mediterranean region and PACA during the entire period 2011 to 2017 are represented in Figure 3.5. In order to highlight the edges that represent strong connections, we depicted them with darker shades of color, while we did not draw the connections that had a weight of less than 0.3 for the Mediterranean and 2×10^3 for PACA. It can be seen that the strongest connections tend to link nodes that are

geographically close, but nonetheless, moderate connections also exist between rather distant nodes (see also Figure .5). This is confirmed by small values of the average shortest path distance 8.20×10^{-4} for the Mediterranean and 2.57×10^{-4} for PACA, and high values of the transitivity index (0.74 for the Mediterranean region and 0.99 for PACA), as shown in the first lines of Tables 3.1 and 3.2. The connectivities in PACA are mostly oriented from North-East to South-West, which corresponds to the direction of the prevailing wind in this region. For the Mediterranean basin, the direction of the connectivities depends on the region and does not have a fixed direction (see Figure .6). An interesting additional difference between the two networks is that the one for PACA has a very negative degree correlation (-0.85), meaning that nodes having a high incoming degree will have a low outgoing degree, and vice versa. On the other hand, for the Mediterranean network, the value of the degree correlation is moderately positive (0.31), meaning that nodes having high an incoming degree tend to have also high an outgoing degree.

Mediterranean region								
	Diam	Dens	Trans	S P (mean)	S P (sd)	S W	S F	D C
2011-2017	3.12×10^{-3}	2.96×10^{-4}	0.74	8.20×10^{-4}	2.19×10^{-4}	18242	11.6	0.31
2011	0.03	3.0×10^{-4}	0.67	6.17×10^{-3}	1.92×10^{-3}	2222	11.5	-0.04
2012	0.02	2.9×10^{-4}	0.67	6.09×10^{-3}	1.87×10^{-3}	2236	10.4	-0.06
2013	0.01	3.0×10^{-4}	0.66	6.06×10^{-3}	1.86×10^{-3}	2228	13.9	0.20
2014	0.08	2.9×10^{-4}	0.65	6.52×10^{-3}	2.11×10^{-3}	2044	9.0	0.30
2015	0.06	3.0×10^{-4}	0.66	6.22×10^{-3}	1.93×10^{-3}	2147	8.2	0.26
2016	0.55	3.0×10^{-4}	0.66	6.32×10^{-3}	1.97×10^{-3}	2110	8.8	0.25
2017	0.12	3.0×10^{-4}	0.65	6.08×10^{-3}	1.83×10^{-3}	2186	9.2	0.12
January	0.23	3.1×10^{-4}	0.63	1.15×10^{-2}	3.81×10^{-3}	1118	14.9	0.14
February	0.67	3.0×10^{-4}	0.63	1.15×10^{-2}	3.92×10^{-3}	1106	14.9	0.24
March	0.29	3.0×10^{-4}	0.62	1.15×10^{-2}	3.87×10^{-3}	1104	11.9	0.26
April	0.70	3.1×10^{-4}	0.64	1.13×10^{-2}	3.91×10^{-3}	1137	15.4	0.18
May	1.02	3.2×10^{-4}	0.63	1.20×10^{-2}	4.43×10^{-3}	1064	12.4	0.21
June	1.14	3.2×10^{-4}	0.61	1.19×10^{-2}	4.18×10^{-3}	1041	12.9	0.15
July	1.41	3.1×10^{-4}	0.60	1.19×10^{-2}	4.01×10^{-3}	1017	28.2	-0.07
August	1.87	2.9×10^{-4}	0.59	1.22×10^{-2}	4.14×10^{-3}	987	11.9	-0.02
September	1.51	2.9×10^{-4}	0.60	1.23×10^{-2}	4.18×10^{-3}	994	12.8	0.13
October	0.12	2.8×10^{-4}	0.62	1.26×10^{-2}	4.46×10^{-3}	998	14.9	0.12
November	0.10	2.7×10^{-4}	0.62	1.27×10^{-2}	4.90×10^{-3}	997	14.3	0.41
December	1.07	3.0×10^{-4}	0.61	1.16×10^{-2}	3.89×10^{-3}	1079	14.1	0.32

Table 3.1. – Network indices (Diameter, density, transitivity, shortest path (mean and standard deviation), small worldness, scale-free property, degree correlation) calculated from the networks covering the Mediterranean region and estimated in three temporal contexts: the entire period 2011-2017, yearly time periods from 2011 to 2017 and monthly time periods.

PACA								
	Diam	Dens	Trans	S P (mean)	S P (sd)	S W	S F	D C
2011-2017	6.06×10^{-3}	2.51×10^{-3}	0.99	2.57×10^{-4}	1.00×10^{-4}	7813	19.2	-0.85
2012	2.22×10^{-3}	9.9×10^{-4}	0.97	6×10^{-4}	1.7×10^{-4}	3287.37	21.50	-0.88
2013	2.42×10^{-3}	1.01×10^{-3}	0.98	6.5×10^{-4}	1.8×10^{-4}	3069.76	23.71	-0.91
2014	4.23×10^{-3}	10^{-3}	0.98	7.3×10^{-4}	2.2×10^{-4}	2717.58	24.56	-0.92
2015	2.92×10^{-3}	1.02×10^{-3}	0.99	7.2×10^{-4}	2.2×10^{-4}	2789.31	27.09	-0.92
2016	4.18×10^{-3}	1.01×10^{-3}	0.98	8×10^{-4}	2.5×10^{-4}	2470.12	19.29	-0.92
2017	2.46×10^{-3}	1.01×10^{-3}	0.98	5.9×10^{-4}	1.6×10^{-4}	3348.41	17.51	-0.89
January	2.76	1.8×10^{-3}	0.69	3.03×10^{-3}	9.2×10^{-4}	462.54	6.84	-0.66
February	4.13	1.64×10^{-3}	0.73	3.35×10^{-3}	1.09×10^{-3}	439.54	8.23	-0.6
March	0.57	1.71×10^{-3}	0.7	2.18×10^{-3}	6.9×10^{-4}	651.63	22.3	-0.84
April	1.34×10^{-2}	1.69×10^{-3}	0.99	2.81×10^{-4}	8.8×10^{-4}	712.56	43.44	-0.98
May	0.37	1.86×10^{-3}	0.84	1.73×10^{-3}	5.7×10^{-4}	984.96	46.82	-0.93
June	5.05	1.79×10^{-3}	0.7	5.28×10^{-3}	2.04×10^{-3}	270.61	6.96	-0.64
July	5.02	1.84×10^{-3}	0.71	6.85×10^{-3}	2.58×10^{-3}	211.02	7.92	-0.64
August	5.1	1.84×10^{-3}	0.71	5.25×10^{-3}	2.21×10^{-3}	274.57	6.48	-0.66
September	5.1	1.84×10^{-3}	0.64	2.57×10^{-3}	7.6×10^{-4}	505.03	6.46	-0.73
October	1.11×10^{-2}	1.87×10^{-3}	0.99	1.83×10^{-3}	5.3×10^{-4}	1101.63	25.46	-0.92
November	8.23×10^{-3}	1.72×10^{-3}	0.95	1.68×10^{-3}	5.1×10^{-4}	1147.04	29.41	-0.89
December	1.55	1.8×10^{-3}	0.73	2.69×10^{-3}	8.1×10^{-4}	548.04	7.08	-0.7

Table 3.2. – Network indices (Diameter, density, transitivity, shortest path (mean and standard deviation), small worldness, scale-free property, degree correlation) calculated from the networks covering PACA and estimated in three temporal contexts: the entire period 2011-2017, yearly time periods from 2011 to 2017 and monthly time periods.

Qualitatively, the indices provided in Tables 3.1 and 3.2 are overall more variable for the monthly spatio-temporal trajectory-based networks than for the yearly ones. Thus, focusing on what follows on the monthly networks, we investigate possible seasonal patterns by using the complete-linkage hierarchical clustering method Ferreira and Hitchcock (2009). We applied the clustering using the Euclidean distance over the 8-dimensional space formed by the 8 indices provided in Tables 3.1 and 3.2.

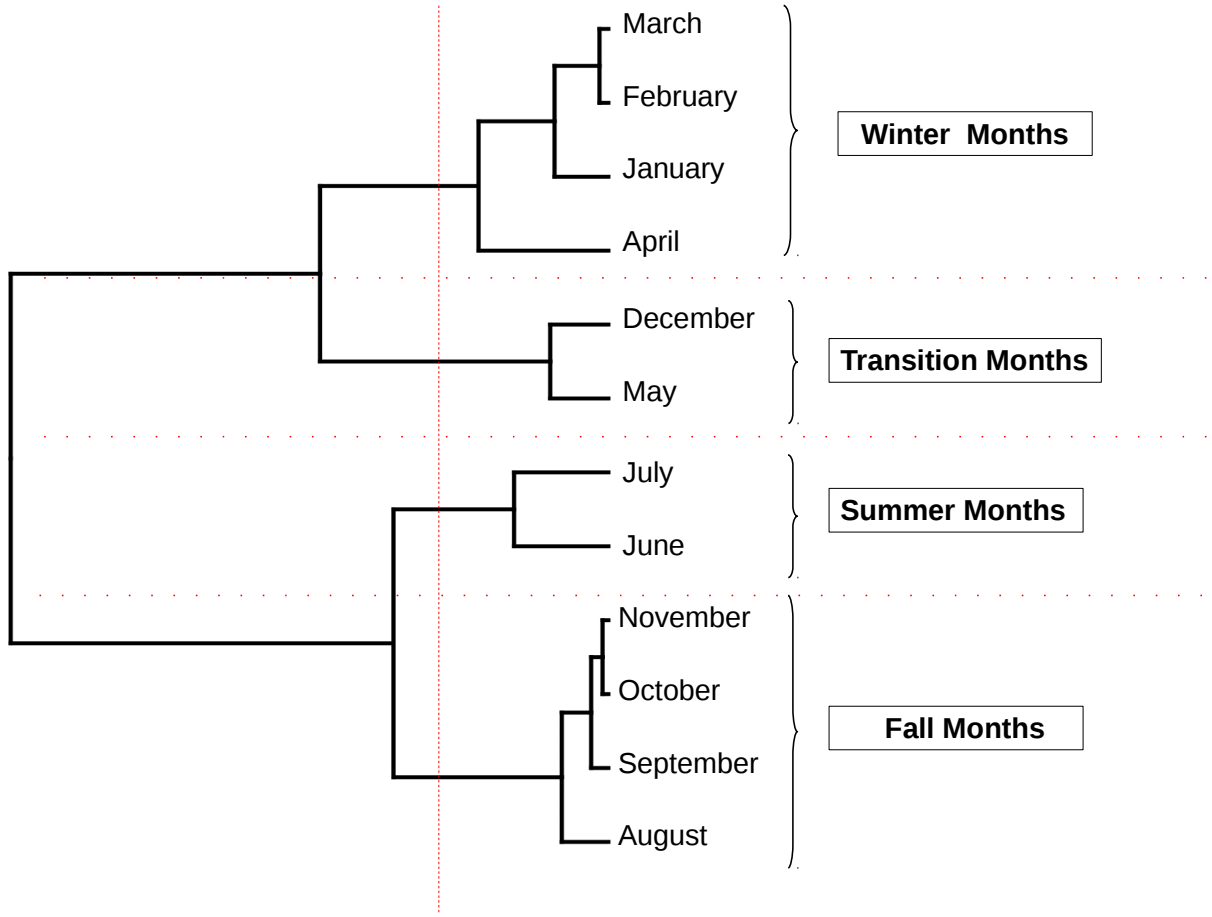
For the Mediterranean region, the dendrogram in Figure 3.6(a) can be used to identify four distinctive periods: summer months (June and July), winter months (January, February, March, April), fall months (August, September, October, November) and a

set of transition months (May and December) surrounding the winter months. The spatial networks derived from this clustering are shown in Figure 3.6(b-e), which displays clear differences in the connectivity patterns even if one observes similarities between the networks for winter months and the surrounding transition months (winter and transition months are precisely in the same dendrogram cluster if one increases the cut-off). The main differences are observed in the northwestern part of the Mediterranean basin with, in particular, increased connectivities in the North of Italy in Winter, in the South of France and the East of Spain in Summer, between Spain and Algeria / Morocco in Summer, and along the eastern Mediterranean coast in Summer.

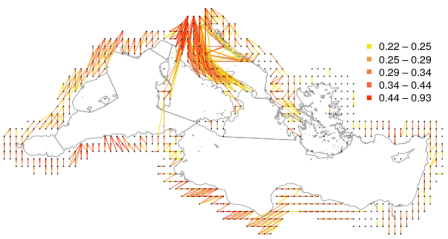
As shown by Figure 3.7(a), summer months are characterized by high diameter and density, low values of transitivity, small-worldness, and the lowest values of degree correlation. Winter months have a lower diameter and show high values of small-worldness due to its high values of clustering and low values of average shortest path distances. Fall months have lower values of density and small-worldness due to its low values of clustering and high values of average shortest path distances. Finally, the group of transition months shows high values of density and degree correlation.

For PACA, the dendrogram in Figure 3.8(a) can be used to identify three distinctive periods: summer months (from June to August), winter and spring months (from December to April, plus September that can be considered as an outlier from a chronological viewpoint) and a set of transition months between the two previous periods (May, October, and November).

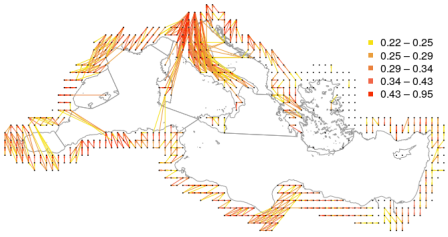
Figure 3.8(b-d) illustrates the differences between the networks derived from this clustering. The summer network is largely more connected than the two other networks, and the transition months, surprisingly, do not lead to intermediate connectivities but to the lowest connectivities. Based on Figure 3.7(b), the group of summer months is characterized by a high diameter, density and average shortest path, low values of transitivity, small-worldness, and the lowest degree correlations (yet still negative). Winter and spring months have a significantly lower diameter, density, and average shortest path distances. Finally, the group of transition months shows the highest values of small-worldness, due to their high values of clustering and low values of average shortest path distances.



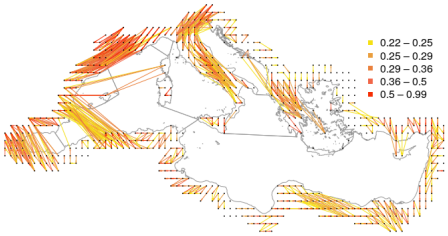
(a)



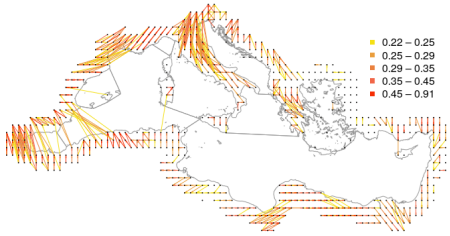
(b) Winter Months



(c) Transition Months

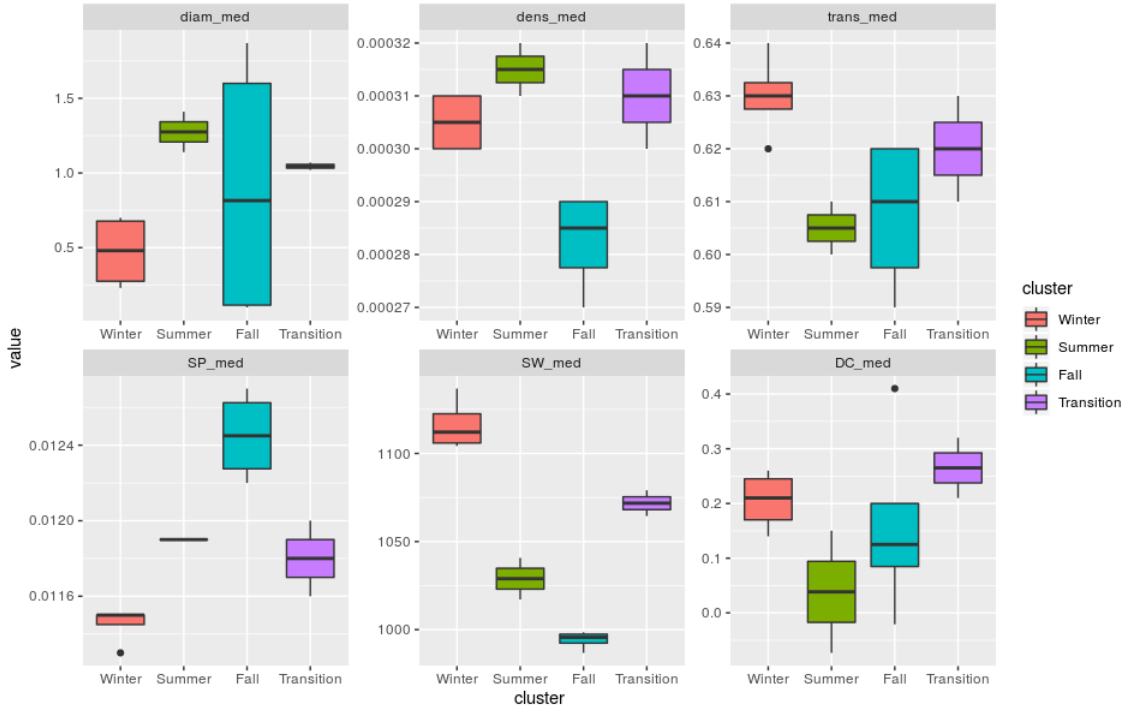


(d) Summer Months

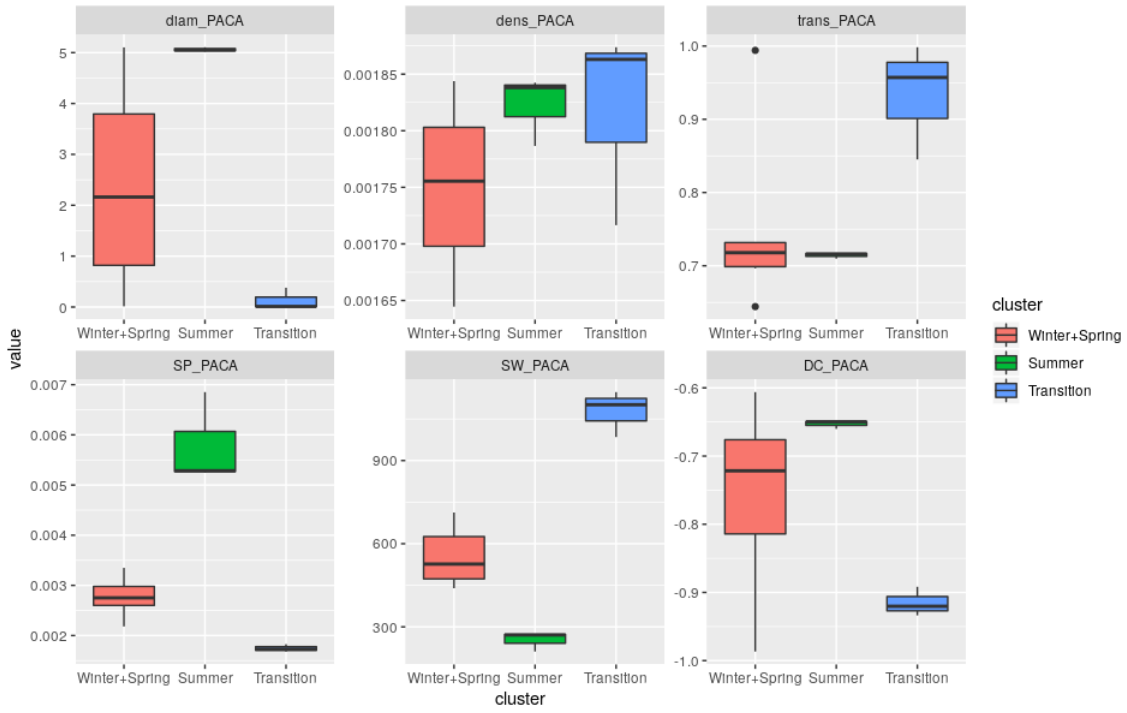


(e) Fall months

Figure 3.6. – (a): Dendrogram of the months obtained from a hierarchical cluster analysis of the Mediterranean spatio-temporal network based on the monthly dissimilarities of the indices presented in Table 3.1. (b), (c), (d), and (e): Networks corresponding to the four identified clusters where one displayed only the edges between the nodes connected more than 10 days per month via the air mass trajectories.

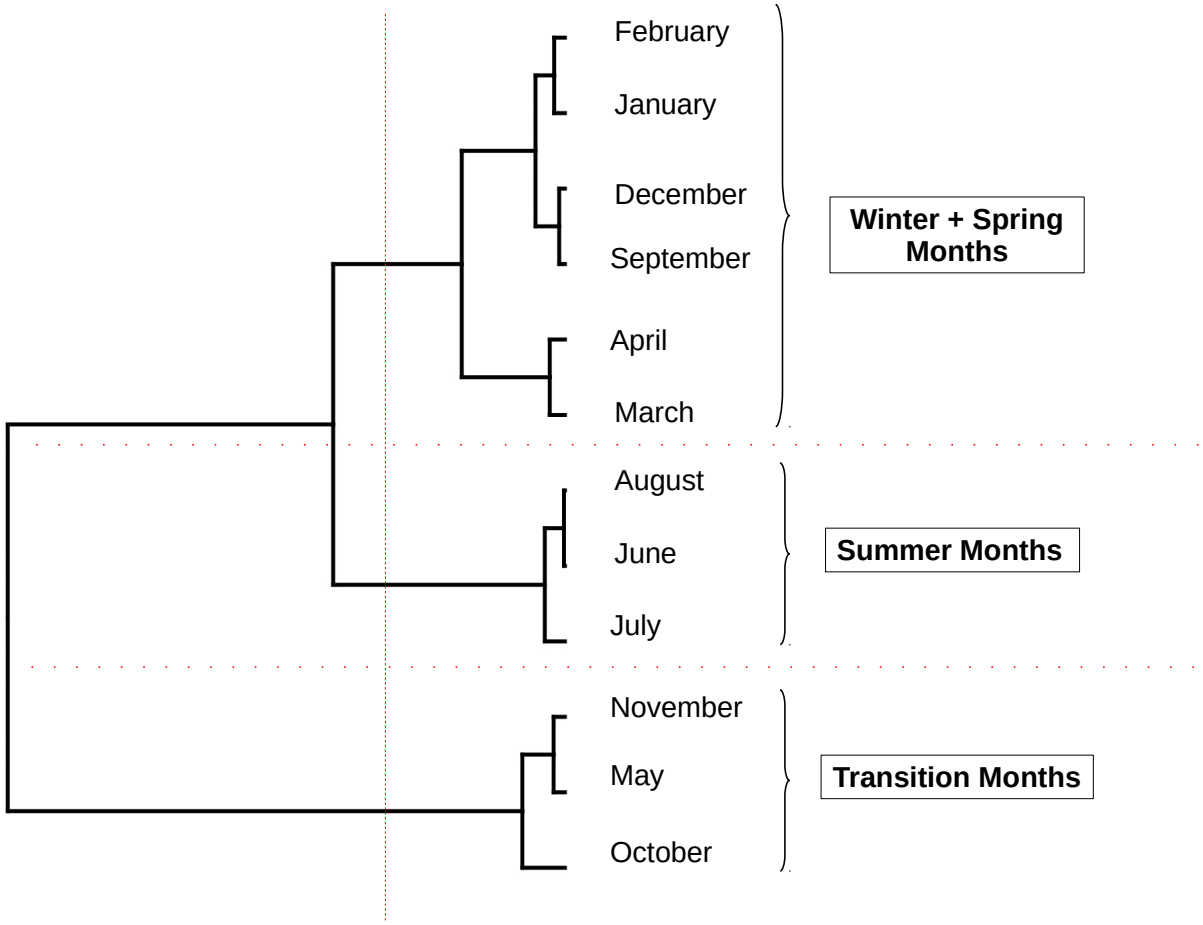


(a) Mediterranean region

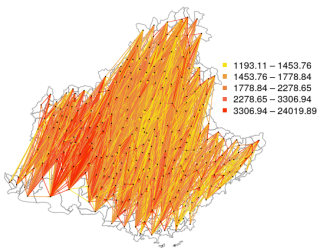


(b) Mediterranean region

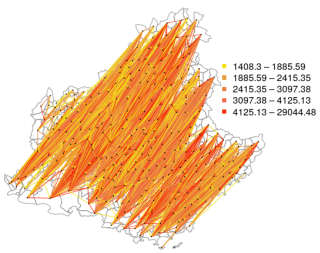
Figure 3.7. – Boxplot for the different indices (Diameter, density, transitivity, shortest path (mean), small worldness, degree correlation) obtained from (a) the four clusters identified for the Mediterranean region (see Figure 3.6) and (b) the three clusters for PACA (see Figure 3.8).



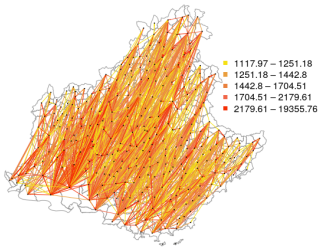
(a)



(b) Winter + Spring months



(c) Summer months



(d) Transition months

Figure 3.8. – (a): Dendrogram of the months obtained from a hierarchical cluster analysis of the PACA spatio-temporal network based on the monthly dissimilarities of the indices presented in Table 3.2. (b), (c) and (d): Networks corresponding to the three identified clusters where one displayed only the edges between the nodes connected more than 10 days per month via the air mass trajectories.

3.6. Discussion

We presented a framework for estimating and characterizing spatial and spatio-temporal networks generated by trajectory data. The development of this framework was motivated by the study of networks resulting from the movement of air masses sampled over long time periods and large spatial scales. Thus, in the application, we investigated the tropospheric connectivities across the Mediterranean basin and the French region PACA, and their variations through years and months. Our approach could be applied to diverse phenomena, from which trajectories can be observed. For instance, one could estimate networks generated by the movement of animals on the landscape scale based on animal trajectories observed with GPS devices Bastille-Rousseau et al. (2018). This would allow the characterization of connectivity between different landscape components. Sampled trajectories of humans, sampled transports of specific goods (such as plant material), and sampled trajectories of knowledge in social communities (that cannot be exhaustively observed) could also be used to estimate networks in other applied settings.

In Section 3.3.3, we proposed diverse measures of connectivity with different underlying (physical or biological) interpretations. Thus, the analyst can adapt the connectivity measure to the mechanistic processes he investigates. In the application section, we only used the contact-based connectivity. Comparisons of contact-based, length-based, and duration-based connectivities, not shown in this manuscript, led to little variations for the two case studies considered in this article. However, the use of covariates such as local rainfall and air-mass altitude for defining connectivities, as proposed in Section 3.3.3, is expected to potentially impact the inferred networks and deserves to be explored. This perspective would be particularly relevant in the context of aerobiology: e.g., the airborne transport of organic particles, such as bacteria and fungal spores, can be influenced by rainfall favoring the deposition of these particles Morris et al. (2017).

In statistics, we are not only interested in point estimation, but also in the assessment of estimation uncertainties. In this paper, we, however, focused on connectivity estimation, even if quantifying the estimation variance could have been useful for more rigorously investigating temporal variation in connectivities. Formally, the connectivity measures that we defined are integrals. Hence, results on integral numerical approximations (e.g., midpoint, trapezoidal, or Monte Carlo integration) can be exploited to assess errors or variances of the connectivity estimates Caflisch (1998); Davis and Rabinowitz (2007); Geweke (1996). However, for this assessment, one should ideally take into account dependencies between connectivity estimates for different pairs of nodes, which is not trivial. Further in-depth methodological developments are required to tackle this issue.

To more finely estimate connectivity, and its uncertainty, one could also take into account, if relevant, the uncertainty about the trajectories themselves. For example, when observed trajectories are smoothed versions of actual trajectories (as it is likely

the case for air-mass trajectories calculated with HYSPLIT) or when the trajectories are partially observed and rather erratic, (i) a probabilistic model grounded on, for instance, a stochastic differential equation, could be used to reconstruct probable trajectories and (ii) the connectivity would be estimated from these reconstructed trajectories. Obviously, step (ii) should incorporate the uncertainty about the trajectory reconstruction impacted by an eventual preliminary step consisting of estimating the parameters of the above-mentioned probabilistic model.

Concerning the application treated in this article, we observed distinct seasonal patterns in the temporal variation of the networks covering the Mediterranean coastline and PACA. In the former case, the networks corresponding to the four clusters shown in Figure 3.6(b-e) exhibit different spatial patterns of hubs (in terms of location and size) and different trends in the main connectivity directions. In the latter case, the differences between the three networks identified with the clustering approach are mostly related to connectivity amplitude. It would be interesting to explore whether this observation made at two very different spatial scales and resolutions generally holds by studying regions of a size similar to PACA all along the Mediterranean coastline.

In the long-term context of our applied research projects connected to aerobiology, the construction and exploration of networks generated by air-mass movements are a way to unravel epidemiological dynamics (and the resulting genetic patterns) of microbial pathogens disseminated at a long distance via air movements in the troposphere (see Leyronas et al., 2018, for a proof of concept). Indeed, even if the pathogen is not explicitly taken into account by the framework proposed in this article, the description of connectivities that it offers provides us a proxy of airborne pathogen movements over long temporal terms and large spatial scales. This proxy is a means to understand pathogen transportation and to anticipate its long-distance dissemination. Specifically, network indices such as those calculated in this article can be associated with particular epidemiological properties such as the probability of long-distance transport of pathogens Jeger et al. (2007); Moslonka-Lefebvre et al. (2011); Pautasso and Jeger (2014). For instance, for plant pathogens, recent studies Aho et al. (2019); Bowers et al. (2013b); Nicolaisen et al. (2017a) showed that airborne populations of bacteria and fungi are rather constant across the years, while higher diversity can be observed in different seasons. This statement resonates with our analyses where we observed clear seasonal signals in the estimated monthly spatio-temporal networks in Section 3.5.3 whereas the yearly signals were less obvious.

Finally, the networks estimated using our approach could be a basis for developing epidemiological models (explicitly handling the pathogen) incorporating long-distance dissemination conditional on recurrent air-mass movements. Such models could be exploited to set up surveillance strategies for early warning and epidemic anticipation in order to help reduce the impacts of airborne pathogens on human health, agricultural production, and ecosystem functioning Mundt et al. (2009).

Acknowledgments

This research was funded by the SPREE project from the French National Research Agency (grant n° ANR-17-CE32-0004-01) and the PHYTOSENTINEL project (grant n° IB-2019-SPE). The authors thank Loïc Houde for his technical assistance in the calculation of trajectories with `HYSPLIT`.

3.7. Conclusion notes

Take home messages

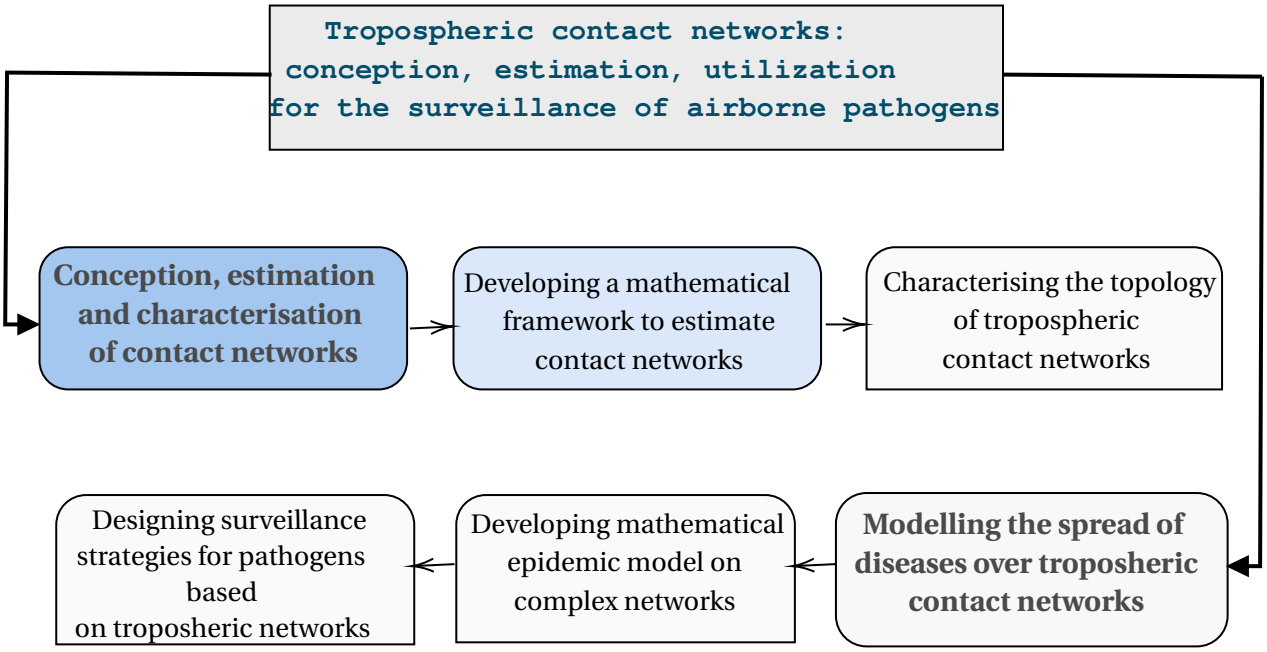
- We propose a mathematical framework for defining the links between geographical areas, which result in spatial and spatio-temporal networks, based on trajectories.
- This framework is used to estimate tropospheric networks based on air mass trajectory data.
- Generally, air mass connectivity is related to the geographic distance between the nodes: the strongest connections tend to link geographically close nodes, while moderate connections also exist between rather distant nodes. However, when looking at the details, in particular by comparing the seasons, the distance is only a rough proxy.
- Using generic indicators used for the analysis of networks, we observe an obvious heterogeneity of the network along the months of the year, but only a subtle variation across years.

Perspectives

- The connectivities in the network are estimated via approximations of integrals taking the form of empirical means. From a statistical perspective, it is crucial to assess the accuracy of the estimation. Since the estimation accuracy depends on how trajectories are sampled to approximate the above-mentioned integrals, one can investigate the impact of the sampling scheme connectivity estimates. **This perspective is the key point of Chapter 4**.
- Different estimators have been proposed in the methodological section of this work: e.g., contact-based, length-based, and duration-based connectivities, whereas in the application section, we only consider the contact-based measure. Hence, it might be worthwhile to take into consideration the other ones in the same context of application to design comparison between the different resulting networks.
- The estimation of the connectivities results in spatio-temporal networks, whose properties can be explored with a multitude of approaches, in particular network centrality measures to identify the relevance of some geographic areas and general network metrics to assess the network topology. **This perspective is the key point of Chapter 5**.
- A spatio-temporal tropospheric network can be viewed as a foundation for representing aerial pathogen dispersal between spatial units. This task can be accomplished by constructing an epidemic model conditional on the tropospheric network. **This perspective is the key point for Chapter 6**.

4. Impact of the sampling scheme on the uncertainty estimation of trajectory-based connectivity between network nodes

Table of Contents



4.1. Abstract

The general framework developed in Chapter 3 consists of an estimation of the weighted and directed spatial connectivity generated by *individuals* traveling through space. It is based on the concept of graph theory where the nodes represent spatial domains and the estimated connectivity characterize the edges linking these domains. Accordingly, when referring to estimation it is essential to evaluate the accuracy of the estimators.

Here, we present a theoretical and numerical approach for studying the accuracy of the connectivity estimators introduced in Chapter 3. Since the estimation of the connectivity between two nodes depends on two essential variables (the spatial arrival point, and the temporal arrival date), we consider three different sampling schemes for these variables: a classical Monte Carlo sampling scheme, a random sampling scheme with dependencies, and a deterministic sampling scheme. We compare their effectiveness by computing some measures of their uncertainty (e.g., the variance and the bias for the stochastic schemes) in a theoretical way. To illustrate the theoretical results, we perform a numerical comparison between the two stochastic schemes. The latter comparison has been designed from the application presented in Chapter 3, which consists of the estimation of the flow of the air mass movements between the watersheds in the French region of Provence-Alpes-Côte d'Azur. Numerical results match the theoretical ones, and both types of results stress the significant impact of the sampling scheme on the estimation of trajectory-based connectivity between network nodes.

4.1.1. Graphical outline

In this chapter, we answer the following problem by following the structure presented in the graphical outline 4.1:

- How to evaluate the accuracy of the air-mass connectivity estimators according to the spatio-temporal sampling scheme?

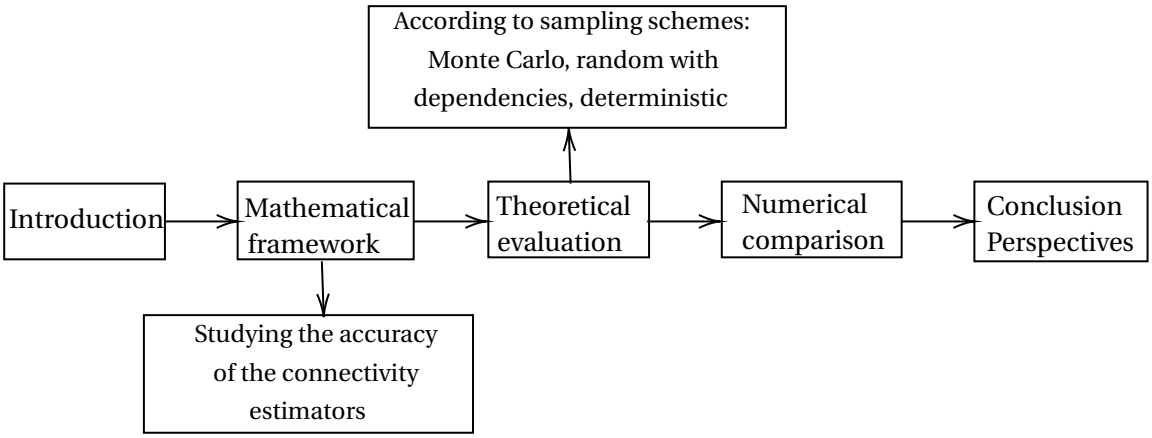


Figure 4.1. – Graphical outline of Chapter 4

4.1.2. Status of the chapter

Research Article

- This article is in progress.
- Authors of this article: Maria Choufany, Rachid Senoussi and Samuel Soubeyrand.

4.2. Introduction

Networks are nowadays used in many fields, e.g., in computer science (Deo, 2017), economics (Koenig and Battiston, 2009), sociology (Harary and Norman, 1953; Scott, 1988), genetics (Staples et al., 2013), epidemiology (Allen et al., 2008; Moslonka-Lefebvre et al., 2011) and ecology (Bunn et al., 2000; Urban and Keitt, 2001). A network is a collection of nodes/vertices linked to each other by means of edges. The way the edges between the nodes are constructed depends on the application field and the available data. Here, we are interested in spatial networks whose directed and weighted edges can be estimated from trajectory data (Choufany et al., 2019a). For instance, if we are interested in quantifying how specific spots in a forest (e.g., forest

clearings) are connected by animal movements, we may follow some animal trajectories tracked with a positioning system and assess how frequent the spots *are linked* by the tracked animals. A second example is related to the estimation of the connectivity between the basins of production of a specific crop via the movements of air (which can carry airborne pathogens): this connectivity can be evaluated by assessing how frequent the trajectories of air masses *go through* different pairs (or sets) of basins of production. Choufany et al. (2019a) proposed a mathematical framework to build a spatial network where the nodes are well-identified disjoint geographic domains, and the edges between nodes are inferred from trajectories sampled in a space encompassing the above-mentioned geographic domains. Therein, the value of a directed edge $(A \rightarrow B)$ between nodes A and B is called *connectivity from domain A to domain B* and is expressed as a multiple integral (over a given temporal window and domain B) of a specific function depending on the mathematical formalization of trajectories hitting both A and B. Hence, the connectivity from A to B can be simply estimated via numerical approximation techniques of integrals such as midpoint rule, trapezoidal rule, sparse grids, Monte Carlo integration, or Bayesian quadrature (Caflisch, 1998; Davis and Rabinowitz, 2007; Evans and Swartz, 2000; Geweke, 1996). These references form only a small snapshot of the huge literature on integral approximation with both deterministic and stochastic approaches. Hence, the more advanced integral approximation techniques could be applied to our problem. However, practical and operational constraints depending on the application field can lead to consider very specific (and eventually far-from-optimal) sampling schemes. In this chapter, we provide a short overview of the impact of the sampling scheme on the uncertainty of the connectivity estimation. In this aim, we consider approximations (of the integrals measuring the connectivity) expressed as empirical means, and we investigate the accuracy of the approximations with respect to the sampling scheme that is used (i.e., how are drawn the trajectories and the locations of observations/measures within these trajectories). We consider three sampling schemes: a classical Monte Carlo sampling scheme, a random sampling scheme with dependencies, and a deterministic sampling scheme.

In what follows, we provide theoretical features (including bias and variance) of the connectivity estimator obtained with each of the three above-mentioned sampling schemes. We also perform a numerical comparison of their efficiency.

4.3. Description of the problem

The directed connectivity $\Psi(A \rightarrow B)$ from domain $A \subset \mathbb{R}^d$ to domain $B \subset \mathbb{R}^d$ $(d \in \mathbb{N}^*)$ over an interval of time $T \subset \mathbb{R}$ is theoretically defined in Choufany et al. (2019a) by:

$$\Psi(A) = \int_{T \times B} \psi(A|s,x)\, \nu(ds, dx) \tag{4.1}$$

where $\Psi(A)$ is a short notation for $\Psi(A \to B)$, $\nu(ds, dx)$ is a probability measure on $T \times B$, and ψ is a scalar operator called *pointwise connectivity* by Choufany et al. (2019a, Definition 2.2). Unless drastic restrictions, the connectivity cannot be analytically computed but can be empirically estimated. Here, we consider the following empirical estimator:

$$\widehat{\Psi}(A) = \frac{1}{|I|} \sum_{i \in I} \psi(A|s_i, x_i) \tag{4.2}$$

where the sample $(s_i, x_i)_{i \in I}$ is drawn from a measure m over $(T \times B)^{|I|}$, which can correspond to either a random or a deterministic process, I is the set of sample labels and $|I|$ is the sample size. Depending on the chosen sampling measure m, the statistical properties of $\widehat{\Psi}(A)$ may significantly differ. We precisely aim to determine the statistical properties of $\widehat{\Psi}(A)$ for specific forms of the measure m and to identify the appropriate link between m and ν.

For the form of m, we consider three specific sampling schemes: namely, the set of vectors $(x_i, s_i)_{i \in I}$ correspond to one of the following cases:

- $[S_1:]$ Independent and identically distributed (i.i.d.) random variables (which corresponds to a Monte-Carlo approach);
- $[S_2:]$ Identically distributed but dependent (i.d.d.) random variables;
- $[S_3:]$ Deterministically fixed variables.

In the first two cases, we are interested in the expression of the bias, moments and variances of $\widehat{\Psi}(A)$:

$$\begin{aligned}
\mathbb{E}_m\left[\widehat{\Psi}(A)\right] &= \frac{1}{|I|} \sum_{i \in I} \mathbb{E}_m\left[\psi(A|s_i, x_i)\right] \\
\mathbb{E}_m\left[\left(\widehat{\Psi}(A)\right)^2\right] &= \frac{1}{|I|^2} \sum_{i \in I, j \in I} \mathbb{E}_m\left[\psi(A|s_i, x_i)\, \psi\left(A|s_j, x_j\right)\right] \\
\mathrm{Var}_m\left[\widehat{\Psi}(A)\right] &= \frac{1}{|I|^2} \sum_{i,j} \left(\mathbb{E}_m\left[\psi(A|s_i, x_i)\, \psi\left(A|s_j, x_j\right)\right] - \mathbb{E}_m\left[\psi(A|s_i, x_i)\right] \mathbb{E}_m\left[\psi\left(A|s_j, x_j\right)\right]\right).
\end{aligned}$$

In the third case, we are interested in bounding the error between $\widehat{\Psi}(A)$ and $\Psi(A)$.

Remark: For any random measure μ over $T \times B$ with appropriate properties, the Monte-Carlo approximation of integral (4.1) is, in the general case:

$$\frac{1}{|I|} \sum_{i \in I} \psi(A|s_i, x_i) \frac{\nu(ds_i, dx_i)}{\mu(ds_i, dx_i)},$$

where (s_i, x_i) are i.i.d. from μ and, in general, the closer μ and ν, the better. In this chapter, we voluntarily omit the ratio $\nu(ds_i, dx_i)/\mu(ds_i, dx_i)$, we consider a joint distribution m for the whole sample (instead of μ) and, as mentioned above, we identify the appropriate link between m and ν under the constraint given by the sampling scheme S_1, S_2 or S_3. In the case of S_1 treated in Section 4.4.1, we will

obviously retrieve the classical result that m should coincide with a product of ν.

4.4. Theoretical formulas

Thereafter, for the sake of simplicity, we use the following notation: $\mathbb{E}_\mu\left[\psi(A|s,x)\right]$, $\mathrm{Var}_\mu\left[\psi(A|s,x)\right]$, considering that (s,x) is a random variable distributed from μ.

4.4.1. S_1: independent and identically distributed (i.i.d.) random variables

Suppose that vectors (s_i,x_i), $i\in I$, are independent and identically distributed from the same individual distribution $\mu(ds,dx)$, and let $m=\mu^{\otimes(I)}$. In this case:

$$\begin{aligned}\mathbb{E}_m\left[\widehat{\Psi}(A)\right] &= \mathbb{E}_m\left[\frac{1}{|I|}\sum_{i\in I}\psi(A|s_i,x_i)\right]\\ &= \frac{1}{|I|}\sum_{i\in I}\mathbb{E}_m\left[\psi(A|s_i,x_i)\right]\\ &= \frac{|I|}{|I|}\mathbb{E}_\mu\left[\psi(A|s_i,x_i)\right]\\ &= \mathbb{E}_\mu\left[\psi(A|s_i,x_i)\right]\\ &= \int_{T\times B}\psi(A|s,x)\,\mu(ds,dx),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_m\left[\widehat{\Psi}(A)^2\right] &= \mathbb{E}_m\left[\frac{1}{|I|^2}\left(\sum_{i\in I}\psi(A|s_i,x_i)\right)^2\right]\\ &= \frac{1}{|I|^2}\left[\sum_{i=j}\mathbb{E}_m\left[\psi(A|s_i,x_i)^2\right]+\sum_{i\neq j}\mathbb{E}_m\left[\psi(A|s_i,x_i)\psi(A|s_j,x_j)\right]\right]\\ &= \frac{|I|}{|I|^2}\mathbb{E}_\mu\left[\psi(A|s_i,x_i)^2\right]+\frac{|I|^2-|I|}{|I|^2}\left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2\\ &= \frac{1}{|I|}\left(\mathbb{E}_\mu\left[\psi(A|s,x)^2\right]+(|I|-1)\left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2\right).\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}_m\left[\widehat{\Psi}(A)\right] &= \mathbb{E}_m\left[\widehat{\Psi}(A|s,x)^2\right] - \left(\mathbb{E}_m\left[\widehat{\Psi}(A|s,x)\right]\right)^2 \\
&= \frac{1}{|I|}\mathbb{E}_\mu\left[\psi(A|s,x)^2\right] + \frac{|I|-1}{|I|}\left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 \\
&= \frac{1}{|I|}\left[\mathbb{E}_\mu\left[\psi(A|s,x)^2\right] - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2\right] \\
&= \frac{1}{|I|}\left[\int_{T\times B}\psi(A|s,x)^2\,\mu(ds,dx) - \left(\int_{T\times B}\psi(A|s,x)\,\mu(ds,dx)\right)^2\right] \\
&= \frac{1}{|I|}\mathrm{Var}_\mu\left[\psi(A|s,x)\right]
\end{aligned} \tag{4.3}$$

Consequently,

- $\widehat{\Psi}(A)$ is unbiased if $\mu(ds,dx) = \nu(ds,dx)$ (i.e., if $m = \nu^{\otimes(I)}$),
- Moreover, if $\mu = \nu$ and if we make the regularity assumptions necessary for the almost sure (a.s.) convergence theorem and the classical central limit theorem (CLT) to hold, we get:

  $$\widehat{\Psi}(A) \xrightarrow{a.s.} \Psi(A)$$

  and

  $$\sqrt{I} \times \left(\widehat{\Psi}(A) - \Psi(A)\right) \xrightarrow{weak} \mathcal{N}\left(0, \sigma^2\right),$$

  where, $\sigma^2 = \mathrm{Var}_\mu\left[\psi(A|s,x)\right]$. Note that σ^2 can be estimated by

  $$\hat{\sigma}^2 = \frac{1}{|I|}\sum_{i\in I}\left(\Psi(A|s_i,x_i) - \widehat{\Psi}(A)\right)^2.$$

To sum up, for the i.i.d. sampling scheme S_1, the estimator is unbiased and has variance $\frac{1}{|I|}\mathrm{Var}_\nu\left[\psi(A|s,x)\right]$ if $\mu = \nu$.

4.4.2. S_2: Identically distributed but dependent (i.d.d.) random variables

For the sampling scheme S_2, we consider the following design: Let ν_S and ν_X denote the marginal distributions of ν with respect to s and x, respectively. The sampling consists in drawing N_1 i.i.d. random variables $\widetilde{s}_1, \cdots, \widetilde{s}_{N_1}$ from $\nu_S(ds)$ and N_2 i.i.d. random variables $\widetilde{x}_1 \cdots \widetilde{x}_{N_2}$ from $\nu_X(dx)$, and then to cross combine them. Thus, the size of I is $|I| = N_1 \times N_2$, and for each $i = (l,k) \in I$, with $l = 1, \cdots, N_1$, $k = 1, \cdots, N_2$, we define the sampling item $(s_i, x_i) = (\widetilde{s}_l, \widetilde{x}_k)$. Any combination $(\widetilde{s}_l, \widetilde{x}_k)$ is therefore distributed from $\mu(ds,dx) = \nu_S(ds) \times \nu_X(dx)$, which is distinct from $\nu(ds,dx)$ in general. Even if this sampling scheme lacks important statistical properties because s and x can be more or less dependant, it is frequently used because the spatio-

temporal sample forms a grid with repetitions in time and repetitions in space.

Using the setting described above, the measure m on $(T \times B)^{|I|}$ can be described as the image measure of $\left(\widetilde{s}_1, \cdots \widetilde{s}_{N_1}, \widetilde{x}_1, \cdots \widetilde{x}_{N_2}\right)$ by the function $F\colon\ T^{N_1} \times B^{N_2} \to (T \times B)^{|I|}$ whose u-th component ($u \in \{1, \cdots, N_1 \times N_2\}$) is

$$F_u\left(\widetilde{s}_1, \cdots \widetilde{s}_{N_1}, \widetilde{x}_1, \cdots \widetilde{x}_{N_1}\right) = \left(\widetilde{s}_{j_1(u)}, \widetilde{x}_{j_2(u)}\right)$$

where $j(u) = (j_1(u), j_2(u))$, $u \in \{1, \ldots, N_1 \times N_2\}$, is the unique solution $(l,k) \in \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$ of the Euclidean division $u = (l-1)N_2 + k$.

One can express the image measure m as follows:

$$\begin{aligned} m\left(\prod_{i \in I}(ds_i, dx_i)\right) = \nu_S(ds_{j(1)})\nu_X(dx_{j(1)})\left(\prod_{k=2}^{N_2}\delta_{s_{j(1)}}(s_{j(k)})\nu_X(dx_{j(k)})\right) \\ \times \prod_{l=2}^{N_1}\Bigg[\nu_S(ds_{j((l-1)N_2+1)})\delta_{x_{j(k)}}(x_{j((l-1)N_2+1)}) \\ \prod_{k=2}^{N_2}\delta_{s_{j((l-1)N_2+1)}}(s_{j((l-1)N_2+k)})\delta_{x_{j(k)}}(x_{j((l-1)N_2+k)})\Bigg], \end{aligned}$$

where $\delta_a(\cdot)$ is the Dirac measure supported by $\{a\}$, $s_{j(u)} = \widetilde{s}_{j_1(u)}$ and $x_{j(u)} = \widetilde{x}_{j_2(u)}$. Hence, the pairs (s_i, x_i), $i \in I$ are identically distributed from the product of marginals $\nu_S(ds) \times \nu_X(dx)$, which differs in general from $\nu(ds, dx)$, but are not independent. The dependencies result from the fact that, for example, $(\widetilde{s}_1, \widetilde{x}_1)$ is neither independent from $(\widetilde{s}_1, \widetilde{x}_2)$ nor from $(\widetilde{s}_2, \widetilde{x}_1)$, but is independent from $(\widetilde{s}_2, \widetilde{x}_2)$. More generally, $(\widetilde{s}_l, \widetilde{x}_k)$ and $(\widetilde{s}_{l'}, \widetilde{x}_{k'})$ are independent if and only if $l \neq l'$ and $k \neq k'$. This translates as follows for the sampled times and locations: Let i be the solution of $u = (l-1)N_2 + k$ with respect to (l,k), let i' be the solution of $u' = (l-1)N_2 + k$ with respect to (l,k), (s_i, x_i) and $(s_{i'}, x_{i'})$ are independent if and only if $\lfloor u/N_2 \rfloor \neq \lfloor u'/N_2 \rfloor$, where $\lfloor a \rfloor$ is the integer part of a.

Under the sampling scheme S_2 specified above, the expression of the estimator of $\Psi(A)$ is:

$$\widehat{\Psi}(A) = \frac{1}{N_1 N_2}\sum_{l,k}\psi(A|\widetilde{s}_l, \widetilde{x}_k),$$

whose expectation and bias satisfy:

$$\begin{aligned} \mathbb{E}_m\left[\widehat{\Psi}(A)\right] &= \int_B\int_T \psi(A|s,x)\mu(ds,dx) \\ &= \mathbb{E}_\mu\left[\psi(A|s,x)\right] \end{aligned} \tag{4.4}$$

$$\mathrm{Bias}_m\left[\widehat{\Psi}(A)\right] = \int_B\int_T \psi(A|s,x)\left[\mu(ds,dx) - \nu(ds,dx)\right]. \tag{4.5}$$

The second order moment of $\psi(A|\tilde{s}_l, \tilde{x}_k)$, is:

$$\begin{aligned}\mathbb{E}_m\left[\psi(A|\tilde{s}_l,\tilde{x}_k)^2\right] &= \int_{T\times B}\psi^2(A|s,x)\mu(ds,dx)\\ &= \mathbb{E}_\mu\left[\psi(A|s,x)^2\right].\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}_m\left[\widehat{\Psi}(A)^2\right] &= \frac{1}{(N_1N_2)^2}\sum_{l_1,k_1}\sum_{l_2,k_2}\mathbb{E}_m\left[\psi(A|\widetilde{s}_{l_1},\widetilde{x}_{k_1})\times\psi(A|\widetilde{s}_{l_2},\widetilde{x}_{k_2})\right]\\ &= \frac{1}{(N_1N_2)^2}\left[\sum_{l_1=l_2}\sum_{k_1=k_2}\mathbb{E}_m\left[\psi(A|\widetilde{s}_{l_1},\widetilde{x}_{k_1})^2\right]\right.\\ &\qquad + \sum_{l_1\neq l_2}\sum_{k_1\neq k_2}\mathbb{E}_m\left[\psi(A|\widetilde{s}_{l_1},\widetilde{x}_{k_1})\times\psi(A|\widetilde{s}_{l_2},\widetilde{x}_{k_2})\right]\\ &\qquad + \sum_{l_1\neq l_2}\sum_{k_1=k_2}\mathbb{E}_m\left[\psi(A|\widetilde{s}_{l_1},\widetilde{x}_{k_1})\times\psi(A|\widetilde{s}_{l_2},\widetilde{x}_{k_1})\right]\\ &\qquad \left. + \sum_{l_1=l_2}\sum_{k_1\neq k_2}\mathbb{E}_m\left[\psi(A|\widetilde{s}_{l_1},\widetilde{x}_{k_1})\times\psi(A|\widetilde{s}_{l_1},\widetilde{x}_{k_2})\right]\right]\\ &= M_1+M_2+M_3+M_4,\end{aligned}$$

where

$$
\begin{aligned}
M_1 &= \frac{1}{(N_1 N_2)} \mathbb{E}_\mu \left[\psi(A|s,x)^2 \right] \\
M_2 &= \frac{(N_1-1)(N_2-1)}{(N_1 N_2)} \mathbb{E}_m \left[\psi(A|\widetilde{s}_{l_1}, \widetilde{x}_{k_1}) \times \psi(A|\widetilde{s}_{l_2}, \widetilde{x}_{k_2}) \right] \\
&= \frac{(N_1-1)(N_2-1)}{(N_1 N_2)} \left(\mathbb{E}_\mu \left[\psi(A|\widetilde{s}_{l_1}, \widetilde{x}_{k_1}) \right] \right)^2 \\
&= \frac{(N_1-1)(N_2-1)}{(N_1 N_2)} \left(\mathbb{E}_\mu \left[\psi(A|s,x) \right] \right)^2 \\
M_3 &= \frac{N_1-1}{(N_1 N_2)} \mathbb{E}_m \left[\psi(A|\widetilde{s}_{l_1}, \widetilde{x}_{k_1}) \times \psi(A|\widetilde{s}_{l_2}, \widetilde{x}_{k_1}) \right] \\
&= \frac{N_1-1}{(N_1 N_2)} \int_{T^2} \int_B \psi(A|s_1,x) \times \psi(A|s_2,x) \nu_S(d_{s_1}) \nu_S(d_{s_2}) \nu_X(d_x) \\
&= \frac{N_1-1}{(N_1 N_2)} \int_B \left[\int_T \psi(A|s,x) \nu_S(d_s) \right]^2 \nu_X(d_x) \\
M_4 &= \frac{N_2-1}{(N_1 N_2)} \mathbb{E}_m \left[\psi(A|\widetilde{s}_{l_1}, \widetilde{x}_{k_1}) \times \psi(A|\widetilde{s}_{l_1}, \widetilde{x}_{k_2}) \right] \\
&= \frac{N_2-1}{(N_1 N_2)} \int_{B^2} \int_T \psi(A|s,x_1) \times \psi(A|s,x_2) \nu_S(d_s) \nu_X(d_{x_1}) \nu_X(d_{x_2}) \\
&= \frac{N_2-1}{(N_1 N_2)} \int_T \left[\int_B \psi(A|s,x) \nu_X(d_x) \right]^2 \nu_S(d_s).
\end{aligned}
$$

In the expressions of M_3 and M_4, we encounter the following terms:

- The spatial mean of the squared temporal conditional mean of $\psi(A|s,x)$, which is also the second-order moment with respect to space of the temporal mean of $\psi(A|s,x)$:

$$
\mathbb{E}_{\nu_X} \left[(\mathbb{E}_{\nu_S}[\psi(A|s,x)])^2 \right] = \int_B \left[\int_T \psi(A|s,x) \nu_S(ds) \right]^2 \nu_X(dx),
$$

- The temporal mean of the squared spatial conditional mean of $\psi(A|s,x)$:

$$
\mathbb{E}_{\nu_S} \left[(\mathbb{E}_{\nu_X}[\psi(A|s,x)])^2 \right] = \int_T \left[\int_B \psi(A|s,x) \nu_X(dx) \right]^2 \nu_S(ds).
$$

This leads to:

$$\begin{aligned}
\mathrm{Var}_m\left[\widehat{\Psi}(A)\right] &= \mathbb{E}_m\left[\left(\widehat{\Psi}(A)\right)^2\right] - \left(\mathbb{E}_m\left[\widehat{\Psi}(A)\right]\right)^2 \\
&= \frac{1}{N_1 N_2}\Bigg[\mathbb{E}_\mu\left[\psi(A|s,x)^2\right] + (N_1-1)(N_2-1)\left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 \\
&\qquad + (N_1-1)\mathbb{E}_{\nu_X}\left[(\mathbb{E}_{\nu_S}[\psi(A|s,x)])^2\right] + (N_2-1)\mathbb{E}_{\nu_S}\left[(\mathbb{E}_{\nu_X}[\psi(A|s,x)])^2\right]\Bigg] \\
&\quad - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 \\
&= \frac{1}{N_1 N_2}\Bigg[\mathbb{E}_\mu\left[\psi(A|s,x)^2\right] - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 \\
&\qquad + (N_1-1)\left\{\mathbb{E}_{\nu_X}\left[\left(\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right)^2\right] - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2\right\} \\
&\qquad + (N_2-1)\left\{\mathbb{E}_{\nu_S}\left[\left(\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right)^2\right] - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2\right\}\Bigg]
\end{aligned}$$

Since $\mu = \nu_S \times \nu_X$,

$$\mathbb{E}_\mu\left[\psi(A|s,x)\right] = \mathbb{E}_{\nu_S}\left[\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right] = \mathbb{E}_{\nu_X}\left[\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right]$$

and, consequently,

$$\begin{aligned}
\mathbb{E}_{\nu_X}\left[\left(\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right)^2\right] - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 &= \mathbb{E}_{\nu_X}\left[\left(\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right)^2\right] - \left(\mathbb{E}_{\nu_X}\left[\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right]\right)^2 \\
&= \mathrm{Var}_{\nu_X}\left[\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right].
\end{aligned}$$

Similarly

$$\begin{aligned}
\mathbb{E}_{\nu_S}\left[\left(\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right)^2\right] - \left(\mathbb{E}_\mu\left[\psi(A|s,x)\right]\right)^2 &= \mathbb{E}_{\nu_S}\left[\left(\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right)^2\right] - \left(\mathbb{E}_{\nu_S}\left[\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right]\right)^2 \\
&= \mathrm{Var}_{\nu_S}\left[\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right].
\end{aligned}$$

Hence, $\mathrm{Var}_m\left[\widehat{\Psi}(A)\right]$ can be written as follows:

$$\begin{aligned}
\mathrm{Var}_m\left[\widehat{\Psi}(A)\right] = \frac{1}{N_1 N_2}\Big[&\mathrm{Var}_\mu\left[\psi(A|s,x)\right] + (N_1-1)\mathrm{Var}_{\nu_X}\left[\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right] \\
&+ (N_2-1)\mathrm{Var}_{\nu_S}\left[\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right]\Big],
\end{aligned} \tag{4.6}$$

where we remember that $|I| = N_1 N_2$.

Remarks. If $\nu_S(d_s) = ds/|T|$ (where $|T|$ is the duration of T), $\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]$ can be interpreted as the proportion of times in T at which the location x contributes to the connectivity $B \to A$, and $\mathrm{Var}_{\nu_X}\left[\mathbb{E}_{\nu_S}\left[\psi(A|s,x)\right]\right]$ is the variance of this proportion when x varies in B. Similarly, if $\nu_X(d_x) = dx/|B|$ (where $|B|$ is the volume of B), $\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]$ can be interpreted as the proportion of the volume B contributing to

the connectivity $B \to A$ at time s, and $\mathrm{Var}_{\nu_S}\left[\mathbb{E}_{\nu_X}\left[\psi(A|s,x)\right]\right]$ is the variance of this proportion when s varies in T.

Equations (4.3-4.6) yields the following properties of $\widehat{\Psi}(A)$:

- A bias is generally expected with the sampling procedure S_2. However, the bias is null if ν equalizes the product of the marginals ν_S and ν_X, i.e., if $\nu = \mu = \nu_S \times \nu_X$.
- The variance of $\widehat{\Psi}(A)$ obtained with the sampling scheme S_2 is strictly larger than the variance obtained with the i.i.d. sampling scheme S_1, even when $\nu = \nu_S \times \nu_X$, except if $(s,x) \mapsto \psi(A|s,x)$ is constant over $T \times B$, of course.
- The same statement holds for the quadratic error satisfying:

$$\begin{aligned}\mathbb{E}_m\left[\left(\widehat{\Psi}(A) - \Psi(A)\right)^2\right] &= \mathbb{E}_m\left[\left(\widehat{\Psi}(A) - \mathbb{E}_m\left(\widehat{\Psi}(A)\right)\right)^2\right] + \left(\mathbb{E}_m\left[\left(\widehat{\Psi}(A)\right) - \psi(A)\right]\right)^2 \\ &= \mathrm{Var}_m\left[\widehat{\Psi}(A)\right] + \mathrm{Bias}_m\left[\widehat{\Psi}(A)\right]^2\end{aligned}$$

4.4.3. S_3: Deterministically fixed variables

Here, we consider an irregular sampling grid $\{(s_i,x_i) : i \in I\} = \{(\widetilde{s}_l, \widetilde{x}_k) \in T \times B : l = 1,\dots,N_1, k = 1,\dots,N_2\}$, like in Section 4.4.2, but we assume that it is deterministically fixed. Hence, instead of investigating moments of $\widehat{\Psi}(A)$, we are interested in bounding the error $\widehat{\Psi}(A) - \Psi(A)$.

We further assume that $T \times B$ is bounded, ψ is differentiable and $\nu(ds,dx)$ has a differentiable density $f(s,x)$. Thus,

$$\begin{aligned}\Psi(A) &= \int_T \int_B \alpha(s,x)\, ds dx \\ \alpha(s,x) &= \psi(A|s,x) f(s,x)\end{aligned}$$

and we consider the following estimator of $\Psi(A)$:

$$\widehat{\Psi}(A) = \sum_{l,k} \alpha(\widetilde{s}_l, \widetilde{x}_k)\, \Delta_l^S \Delta_k^X,$$

where T is partitioned into real-valued intervals, $T = \bigcup_{l=1}^{N_1} (\delta_l, \delta_{l+1}] = \bigcup_{l=1}^{N_1} T_l$, and B into cells of $\mathbb{R}^d$, $B = \bigcup_{k=1}^{N_2} C_k$. In addition, let $\widetilde{x}_k = (\widetilde{x}_{k1} \cdots \widetilde{x}_{kd}) \in C_k \subset \mathbb{R}^d$, $\Delta_l^S = \delta_{l+1} - \delta_l$ (interval length) and $\Delta_k^X = |C_k|$ (cell volume). The error of the approximation $\widehat{\Psi}(A)$ is equal to:

$$\mathrm{Error}\left[\widehat{\Psi}(A)\right] = \Psi(A) - \widehat{\Psi}(A) = \sum_{l,k}\left[\int_{T_l}\int_{C_k} \left(\alpha(s,x) - \alpha(\widetilde{s}_l, \widetilde{x}_k)\right) ds dx\right].$$

Since α is differentiable, we define for $(s,x) \in (\delta_l, \delta_{l+1}] \times C_k$, the first order approxi-

mation:

$$\alpha(s,x) = \alpha(\widetilde{s}_l,\widetilde{x}_k)+(s-\widetilde{s}_l)\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)+\sum_{j=1}^{d}\left(x_{\cdot j}-x_{kj}\right)\partial_{x|j}\alpha(\widetilde{s}_l,\widetilde{x}_k)+\mathscr{O}\left(|s-\widetilde{s}_l|^2+|x-\widetilde{x}_k|^2\right),$$

where $x=(x_{\cdot 1},\ldots,x_{\cdot d})$, ∂_s denotes the derivative operator with respect to s, and $\partial_{x|j}$ denotes the derivative operator with respect to j-th component of x. Thus,

$$\begin{aligned}\text{Error}\left[\widehat{\Psi}(A)\right] = \sum_{l,k}\Bigg[\int_{T_l}\int_{C_k}\Big\{&(s-\widetilde{s}_l)\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)+\sum_{j=1}^{d}\left(x_{\cdot j}-x_{kj}\right)\partial_{x|j}\alpha(\widetilde{s}_l,\widetilde{x}_k)\\ &+\mathscr{O}\left(|s-\widetilde{s}_l|^2+|x-\widetilde{x}_k|^2\right)\Big\}\,ds\,dx\Bigg],\end{aligned}$$

where $|a|$ denotes the Euclidean distance when a is a point in $\mathbb{R}$ or $\mathbb{R}^d$

We now analyze each of the terms arising in the expression of the error:

$$\begin{aligned}T_1 &= \sum_{l,k}\int_{T_l}\int_{C_k}(s-\widetilde{s}_l)\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\,ds\,dx\\ T_2 &= \sum_{l,k}\int_{T_l}\int_{C_k}\sum_{j=1}^{d}\left(x_{\cdot j}-x_{kj}\right)\partial_{x|j}\alpha(\widetilde{s}_l,\widetilde{x}_k)\,ds\,dx\\ T_3 &= \sum_{l,k}\int_{T_l}\int_{C_k}\mathscr{O}\left(|s-\widetilde{s}_l|^2+|x-\widetilde{x}_k|^2\right)ds\,dx.\end{aligned}$$

- Term 1: Since $\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)$ does not depend on (s,x), and using the variable change $w=s-\widetilde{s}_l$, we get:

$$\begin{aligned}T_1 &= \sum_{l,k}\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\int_{T_l}\int_{C_k}(s-\widetilde{s}_l)\,ds\,dx\\ &= \sum_{l,k}\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\,\Delta_k^X\int_{T_l}(s-\widetilde{s}_l)\,ds\\ &= \sum_{l,k}\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\,\Delta_k^X\int_{\delta_l-\widetilde{s}_l}^{\delta_{l+1}-\widetilde{s}_l}s\,ds\\ &= \frac{1}{2}\sum_{l,k}\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\,\Delta_k^X\left((\delta_{l+1}-\widetilde{s}_l)^2-(\delta_l-\widetilde{s}_l)^2\right)\\ &= \frac{1}{2}\sum_{l,k}\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\,\Delta_k^X\left((\delta_{l+1}-\delta_l)(\delta_{l+1}+\delta_l)-2\widetilde{s}_l(\delta_{l+1}-\delta_l)\right)\\ &= \frac{1}{2}\sum_{l,k}\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)\,\Delta_k^X\Delta_l^S(\delta_{l+1}+\delta_l-2\widetilde{s}_l)\end{aligned}$$

and, consequently,

$$|T_1| \leq \frac{1}{2} \sum_{l,k} \left(\Delta_l^S\right)^2 \Delta_k^X \, |\partial_s \alpha \, (\widetilde{s}_l, \widetilde{x}_k) \, |. \tag{4.7}$$

- Term 2:

$$\begin{aligned} T_2 &= \sum_{l,k} \int_{T_l} ds \sum_{j=1}^{d} \partial_{x|j} \alpha \, (\widetilde{s}_l, \widetilde{x}_k) \int_{C_k} \left(x_{\cdot j} - x_{kj}\right) dx \\ &= \sum_{l,k} \Delta_l^S \sum_{j=1}^{d} \partial_{x|j} \alpha \, (\widetilde{s}_l, \widetilde{x}_k) \int_{C_k} \left(x_{\cdot j} - x_{kj}\right) dx_{\cdot 1} \cdots dx_{\cdot d}. \end{aligned}$$

Each spatial cell C_k is included in the following hyper-rectangle R_k:

$$R_k = \left[r_{k1}, r_{k1} + \Delta_{k1}^X\right] \times \left[r_{k2}, r_{k2} + \Delta_{k2}^X\right] \times \cdots \times \left[r_{kd}, r_{kd} + \Delta_{kd}^X\right],$$

where $r_{kj} = \min\{x_{\cdot j} : x \in C_k\}$ and $r_{kj} + \Delta_{kj}^X = \max\{x_{\cdot j} : x \in C_k\}$. Hence, similarly to T_1, we get:

$$\begin{aligned} |T_2| &\leq \sum_{l,k} \Delta_l^S \sum_{j=1}^{d} \partial_{x|j} \alpha \, (\widetilde{s}_l, \widetilde{x}_k) \int_{R_k} \left(x_{\cdot j} - x_{kj}\right) dx_{\cdot 1} \cdots dx_{\cdot d} \\ &\leq \frac{1}{2} \sum_{l,k} \Delta_l^S \sum_{j=1}^{d} \Delta_{kj}^X \, |R_k| \, |\partial_{x|j} \alpha(\widetilde{s}_l, \widetilde{x}_k)| \end{aligned} \tag{4.8}$$

where $|R_k| = \prod_{j=1}^{d} \Delta_{kj}^X$ is the volume of R_k (note that $|R_k| \geq |C_k| = \Delta_k^X$).

- Term 3: Assuming that α is twice differentiable and using the inequality of Taylor-Lagrange's for several variables, we get:

$$|T_3| \leq \sum_{l,k} \int_{T_l} \int_{C_k} \frac{1}{2} M_{lk} ||h_{lk}||^2 ds dx,$$

where M_{lk} is the maximum absolute value of the second-order derivatives of α in $[\delta_l, \delta_{l+1}] \times C_k$, $h_{lk} = (s, x) - (\widetilde{s}_l, \widetilde{x}_k)$ and $||h_{lk}||$ is the sum of the absolute values of the components of h_{lk}. Assume that $M_{lk} \leq M < \infty$ for all (l, k),

$$\begin{aligned} |T_3| &\leq \frac{M}{2} \sum_{l,k} \int_{T_l} \int_{C_k} \left(|s - \widetilde{s}_l| + \sum_{j=1}^{d} |x_{\cdot j} - \widetilde{x}_{kj}| \right)^2 ds dx \\ &\leq \frac{M}{2} \sum_{l,k} \int_{T_l} \int_{C_k} \left((s - \widetilde{s}_l)^2 + \sum_{j=1}^{d} |s - \widetilde{s}_l| \, |x_{\cdot j} - \widetilde{x}_{kj}| + \sum_{j,j'=1}^{d} |x_{\cdot j} - \widetilde{x}_{kj}| \, |x_{\cdot j'} - \widetilde{x}_{kj'}| \right)^2 ds dx. \end{aligned}$$

Using the following inequalities,

$$\int_{T_l} |s-\widetilde{s}_l| ds \leq \frac{1}{2}(\Delta_l^S)^2$$
$$\int_{T_l} (s-\widetilde{s}_l)^2 ds \leq \frac{1}{3}(\Delta_l^S)^3$$
$$\int_{C_k} |x_{.j}-\widetilde{x}_{kj}| dx \leq \frac{1}{2}\Delta_{kj}^X |R_k|$$
$$\int_{C_k} |x_{.j}-\widetilde{x}_{kj}|\,|x_{.j'}-\widetilde{x}_{kj'}| dx \leq \frac{1}{4}\Delta_{kj}^X\Delta_{kj'}^X |R_k|, \qquad \text{for } j \neq j'$$
$$\int_{C_k} (x_{.j}-\widetilde{x}_{kj})^2 dx \leq \frac{1}{3}(\Delta_{kj}^X)^2 |R_k|,$$

we obtain:

$$\begin{aligned} |T_3| &\leq \frac{M}{2}\sum_{l,k}\left(\frac{1}{3}(\Delta_l^S)^3\Delta_k^X + \frac{1}{4}(\Delta_l^S)^2|R_k|\sum_{j=1}^{d}\Delta_{kj}^X + \frac{1}{4}\Delta_l^S|R_k|\sum_{j,j'=1}^{d}\Delta_{kj}^X\Delta_{kj'}^X \right. \\ &\qquad \left. + \frac{1}{3}\Delta_l^S|R_k|\sum_{j=1}^{d}(\Delta_{kj}^X)^2\right) \qquad (4.9) \\ &\leq \frac{M}{2}\sum_{l,k}\Delta_l^S|R_k|\left(\frac{1}{3}(\Delta_l^S)^2 + \frac{1}{4}\Delta_l^S\sum_{j=1}^{d}\Delta_{kj}^X + \frac{1}{4}\sum_{j,j'=1}^{d}\Delta_{kj}^X\Delta_{kj'}^X + \frac{1}{3}\sum_{j=1}^{d}(\Delta_{kj}^X)^2\right) \end{aligned}$$

We now use inequalities (4.7–4.9) and the notation $\rho_S = \max_{l,k,j}\Delta_l^S$, $\rho_X = \max_{k,j}\Delta_{kj}^X$, $\beta_S = \max_{l,k}|\partial_s\alpha(\widetilde{s}_l,\widetilde{x}_k)|$, $\beta_X = \max_{l,k,j}|\partial_{x|j}\alpha(\widetilde{s}_l,\widetilde{x}_k)|$, for bounding the error:

$$\begin{aligned} \left|\text{Error}\left[\widehat{\Psi}(A)\right]\right| &\leq |T_1| + |T_2| + |T_3| \\ &\leq \frac{N_1N_2}{2}\rho_S^2\rho_X^d\beta_S + \frac{N_1N_2}{2}\rho_S\rho_X^{d+1}d\beta_X \\ &\quad + \frac{N_1N_2}{2}\rho_S\rho_X^d(\rho_S^2/3 + \rho_S\rho_X d/4 + \rho_X^2 d(d-1)/4) + \rho_X^2 d/3)M. \end{aligned}$$

To simplify the bounding of the error, let $\rho = \max\{\rho_S,\rho_X\}$ and $\beta = \max\{\beta_S,\beta_X\}$, then

$$\left|\text{Error}\left[\widehat{\Psi}(A)\right]\right| \leq \frac{N_1N_2}{2}\rho^{d+2}\left(2\beta + \rho\left(\frac{d^2}{4} + \frac{d+1}{3}\right)M\right).$$

4.5. Numerical study

4.5.1. Simulation and estimation settings

In this section, we use for $\psi(A|s,x)$ the contact-based pointwise connectivity proposed by Choufany et al. (2019a, Equation (4)). It is a binary function indicating

whether or not the trajectory of the individual located at x at time s hit the domain A during the lifespan of the trajectory. Here, we consider that the lifespan of the trajectory is a time period preceding s with a fixed duration. Moreover, we set $\nu(ds, dx) = ds \times dx/(|T| \times |B|)$ (product of the normalized Lebesgue measures with $|T|$ the duration of T and $|B|$ the area of B) such that, theoretically, both S_1 and S_2 yield unbiased estimators. Thus, the integrated connectivity $\Psi(A \to B) = \Psi(A) = \int_{T \times B} \psi(A|s, x)\, \nu(ds, dx)$ is the expected proportion of trajectories hitting the domain A before hitting B.

For the numerical study, we randomly chose 10 watersheds (nodes) among the 294 ones of the PACA region (the grey watersheds in Figure 4.2, which are sorted with respect to their areas) and we aim to estimate the sub-network restricted to these nodes and generated by the trajectories of air-masses (the movement of air-masses are supposed to *connect* the distant watersheds). We consider the time period T from January 1, 2011, to December 31, 2017.

Figure 4.2. – 294 watersheds in the PACA region. In grey, the watersheds randomly selected among the 294 watersheds for the uncertainty analysis of the edge estimation.

For each of the two stochastic sampling schemes S_1 and S_2 and each node B among the 10 selected nodes, we independently repeated the following computation 100 times:

- Drawing 100 time-space points $(s_i, x_i) \in T \times B$ from the specified sampling scheme;
- Using the `HYSPLIT` software to reconstruct for each i the backward trajectory (over 48 hours) arriving in location $x_i \in B$ at time s_i and at altitude 500m above the ground level;
- For each of the nine nodes A different from B, computing $\psi(A|s_i, x_i) \in \{0, 1\}$ for all i and estimating $\Psi(A)$ with the empirical mean given by Equation (4.2):

$\widehat{\Psi}(A) = (1/100)\sum_{i=1}^{100} \psi(A|s_i, x_i)$.

Using this algorithm, one obtains, for each link $A \rightarrow B$ and each sampling scheme, 100 estimates of the directed edge and one can then compute its variance.

4.5.2. Results

Figures 4.3–4.4 provide the empirical distributions of the connectivity estimates between nodes for both schemes. We first remark the asymmetry of these *matrices* of distributions with respect to their first diagonals; this asymmetry illustrates the directed nature of the network ($A \rightarrow B \neq B \rightarrow A$). Secondly, we generally observe wider supports and flatter distributions with scheme S_2 than with scheme S_1, as theoretically expected from variance formulas (4.3) and (4.6). This observation clearly holds for significantly connected pairs of nodes, but is not systematic for pairs of nodes that are weakly connected (e.g., see row 1 – column 8 of Figures 4.3–4.4). The latter case is certainly due to the effect of rare events, i.e., rare trajectories hitting the concerning source and receiver watersheds.

Figure 4.5 gives the ratio of empirical variances of connectivity estimates obtained with S_1 and S_2. From the theoretical formulas, one expects ratios lower than 1. This is true in general, except for the weakly connected pair previously identified (i.e., row 1 – column 8) where the ratio is around 1.5.

Moreover, the bigger the watershed, the more it is connected with the other ones, the more the variance is non zero: the ratio for the 4 largest watersheds is approximately between 0 and 1. For the other watersheds, the pattern of the ratio is not homogeneous since we have for a different couple of nodes a zero ratio that corresponds to the absence of connectivity between the node in one of the schemes or both. For the watershed 6 and 7, we note that there are more than 6 white cases which means that they are the most unconnected since they have a small area and are located along the border of the region.

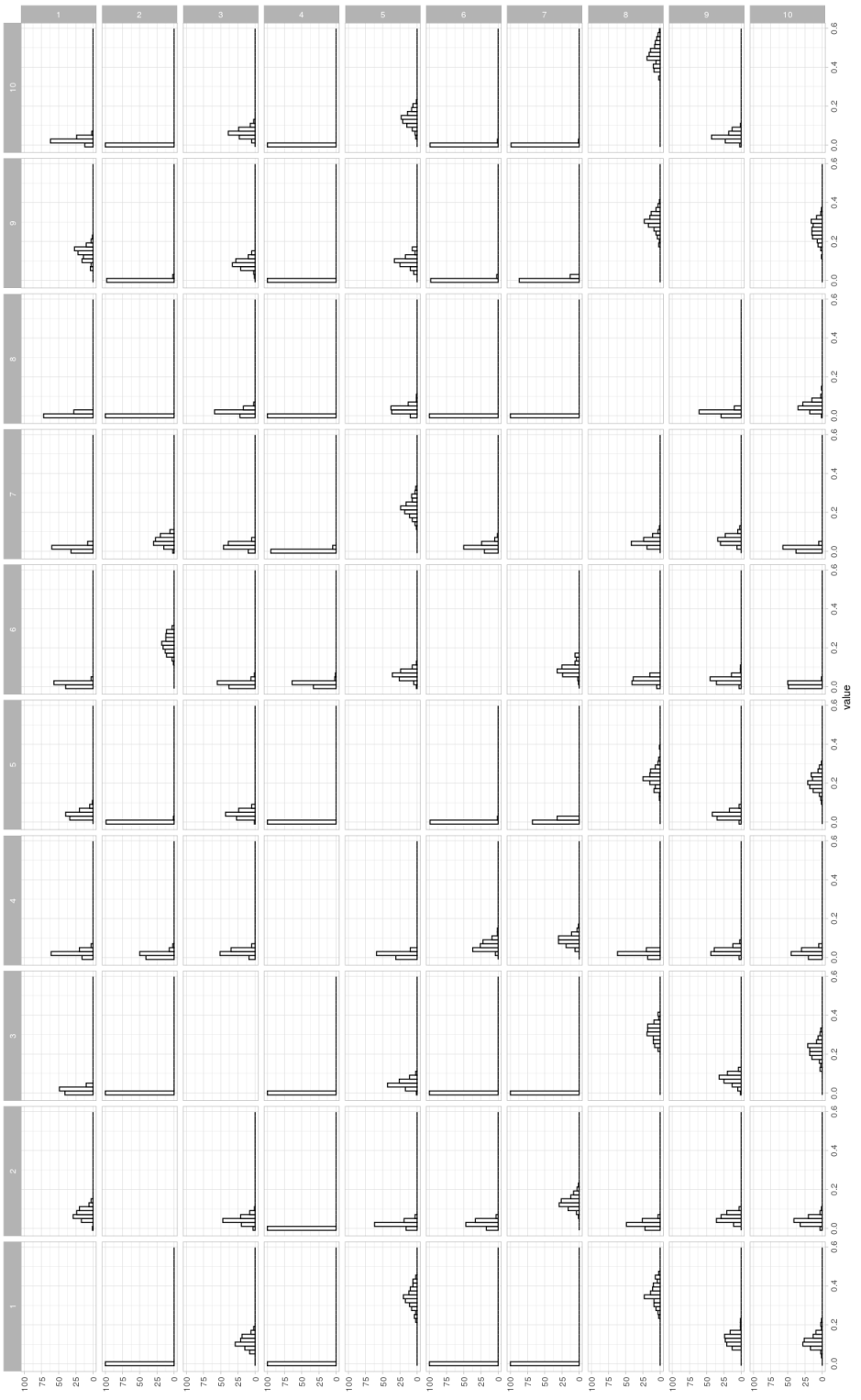


Figure 4.3. – Histograms representing the empirical distributions of connectivity estimates for the 10 selected watersheds in PACA, using sampling scheme S_1 (i.i.d. sampling).

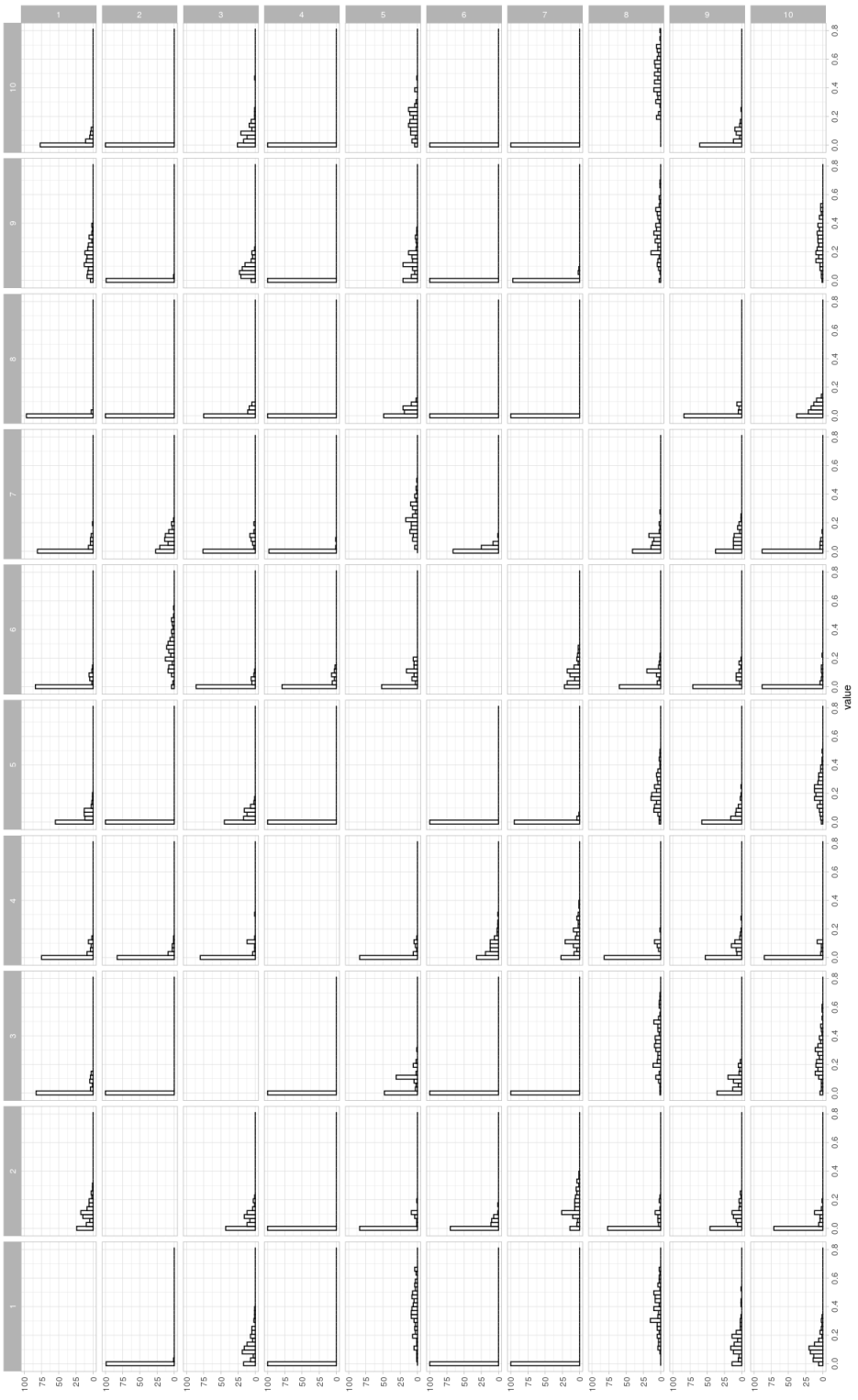


Figure 4.4. – Histograms representing the empirical distributions of connectivity estimates for the 10 selected watersheds in PACA, using sampling scheme S_2 (identically distributed but not independent sampling).

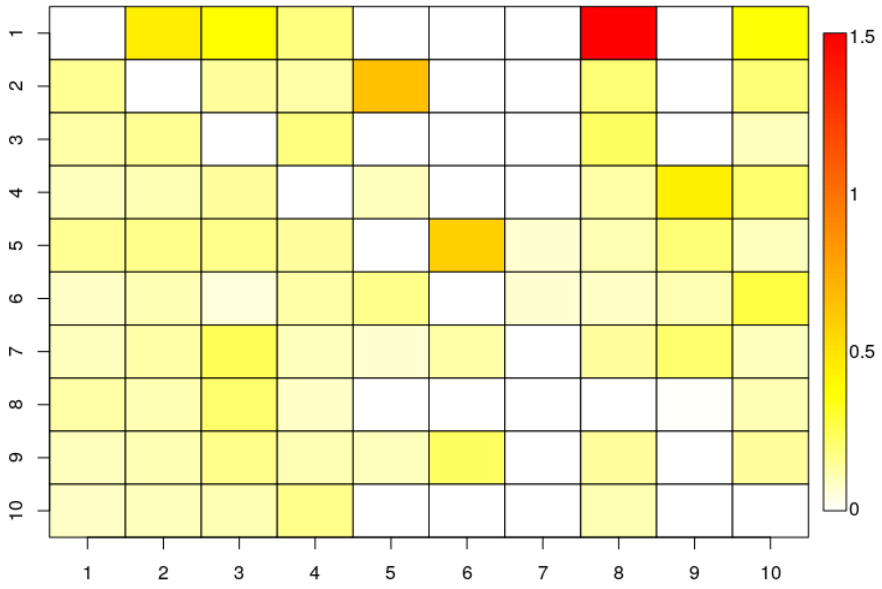


Figure 4.5. – Ratios of the empirical variances of the connectivity estimates obtained with sampling scheme S_1 and sampling scheme (S_2), for the 10 selected watersheds in PACA.

4.6. Conclusion and perspectives

This chapter addresses the issue of the efficiency of the edge estimation in the case of networks generated by the dynamical transport of particles or the movement of individuals. In this context, like in many statistical contexts, estimation is grounded on a sample, the sample is grounded on a design, and the design generally impacts the properties of the estimator. Our study focuses on three very common sampling schemes in spatio-temporal settings, namely a classical Monte Carlo sampling scheme, a random sampling scheme with dependencies built as a random grid, and a deterministic sampling scheme built as an irregular grid. In the first stochastic scheme, the space-time sampling points are independent and identically distributed random vectors. In the second stochastic scheme, a set of independent sampling times and a set of independent sampling locations are independently generated, then the sampling space-time points consist of all the combinations of the sampling times and the sampling locations (and form a random grid). Using appropriate distributions, these schemes lead to unbiased estimators whose uncertainties are assessed by the variance, and we demonstrate that the variance is larger with the 2nd scheme. The de-

terministic scheme is obtained by combining sampling times and sampling locations like for the 2nd sampling scheme, but the sampling times and the sampling locations are assumed to be *a priori* fixed. In this case, we provide an upper bound for the error made in the assessment of the connectivity.

In further work, the performance of other sampling schemes could be investigated such as Quasi Monte Carlo method (Niederreiter, 1988; Woźniakowski, 2000), Latin hypercube sampling (Helton and Davis, 2003), or sampling by considering spatial heterogeneity and others methods as reviewed in (Wang et al., 2012). In particular on could attempt to focus on sampling schemes in the Monte Carlo framework allowing variance reduction such as importance sampling (Glynn and Iglehart, 1989), stratified sampling (Imbens and Lancaster, 1996), or sampling using control variates (Baker et al., 2019) or on Markov chains methods (Evans and Swartz, 2000, Chapters 6 and 7). One could also try to identify an optimal sampling scheme by preliminary defining an optimality criterion and sampling constraints such as considering the simulated annealing method (Van Groenigen and Stein, 1998; Zhu and Stein, 2006) or by adapting entropy-based criteria (Bueso et al., 1998, 1999) for spatial sampling, and bayesian maximum entropy-based criteria (Hosseini and Kerachian, 2017) or by using genetic algorithm (Pourshahabi et al., 2018; Rana et al., 2008) for spatio-temporal sampling.

We restricted our study to three different sample schemes which are mainly used in the context of computing connections via trajectories. However, it can be extended to other cases such as the dynamic sampling schemes using Markov Chain techniques (Hastings, 1970; Van Ravenzwaaij et al., 2018) or hybrid design sampling (Hooten et al., 2009). This dynamicity could be beside the previous spatio-temporal sampled locations: one could adopt the sampling according to the spatial locations of the previous point to maintain maximum consistency by directing sampling spatial points in different parts of the area that could return different values of connections. These types of sampling can be summarized by the sampling via the elimination of spatial areas well known in ecological dynamic networks. (Saura et al., 2014; Thompson et al., 2017).

Another direction would be that each simulated trajectory would be used for all edges at the same time. This method increases the number of trajectories used for the estimation of each edge considerably. General type of temporal connectivity with a continuous range of values, (Holme and Saramäki, 2012), instead of $\{0,1\}$ discrete values used here, will probably give the same approximations.

4.7. Conclusion notes

Take home messages

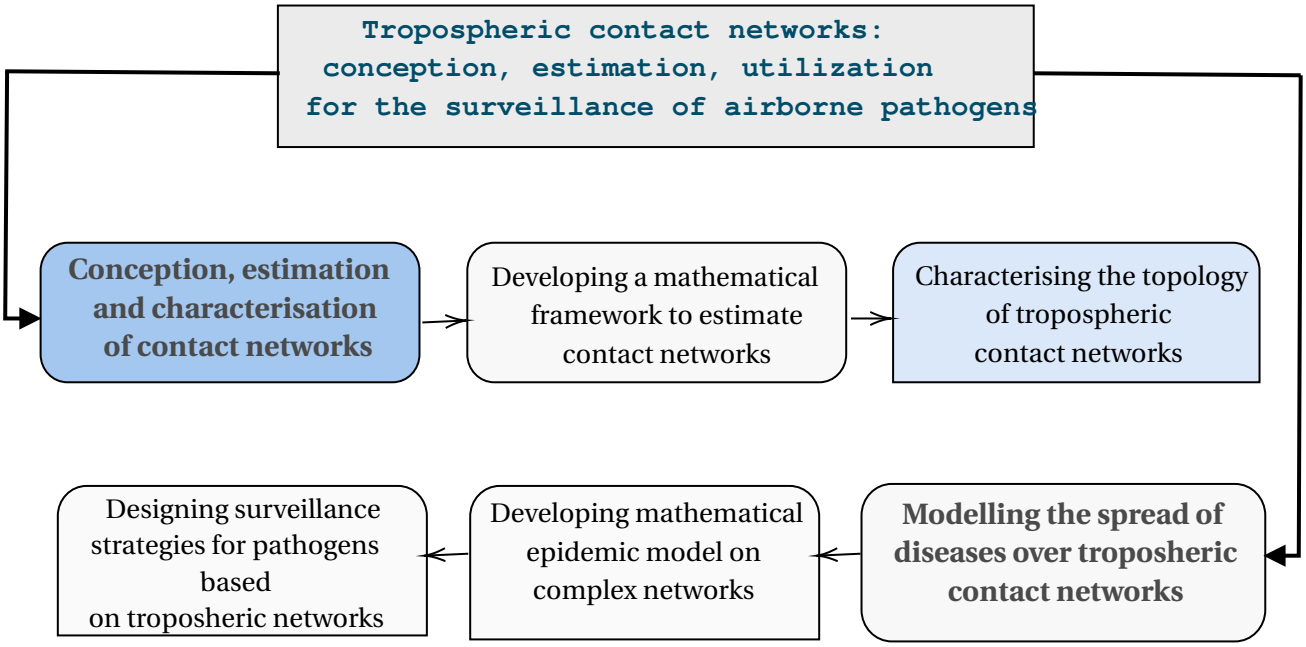
- The accuracy of the estimators of spatio-temporal quantities such as connectivity can be assessed by computing the variance or the quadratic error for different stochastic and deterministic spatio-temporal schemes.
- The stochastic schemes that we studied lead to unbiased and relatively accurate estimators in general, as shown by the computation of the variance. The deterministic scheme, even yielding constant estimators (with null variance), however leads to possibly highly biased (nevertheless bounded) estimators.
- Theoretically, we proved that the variance of the connectivity estimator with identically distributed but dependent (i.d.d.) random sampling locations in space-time is greater than the variance obtained with independent and identically distributed (i.i.d.) sampling points.
- Numerically, we showed that wider supports, flatter distributions, and higher variances are generally obtained with the i.d.d. scheme compared to the i.i.d. scheme (exceptions may be observed when the connectivity is small).

Perspectives

- The study is restricted to three different sampling schemes. It would be interesting to extend the study by considering others sampling schemes such as dynamic (or sequential) ones grounded on a Markov chain model.
- Mean squared errors of the connectivity estimators vary and may be occasionally relatively high, as illustrated by the numerical study. Therefore, the choice of the sampling probability μ should be balanced in a way that the estimators have both small biases and small variances for all the edges for which the connectivity is estimated. This may even require to adapt μ and the sampling size to each edge.
- The evaluation of accuracy is restrained to the pointwise based contact connectivity introduced in Chapter 3. It would be of interest to see whether the evaluation of the accuracy of other estimators for the duration and length-based pointwise connectivity gives similar results.

5. Long-distance connectivity shaped by air-mass movement: a complex network tool for experimental design in aerobiology

Table of Contents



5.1. Abstract

The collection and analysis of air samples for the study of microbial airborne communities or the detection of airborne pathogens is one of the few insights that we can grasp of a continuously moving flux of microorganisms from their sources to their sinks through the atmosphere. For large-scale studies, a comprehensive sampling of the atmosphere is beyond the scopes of any reasonable experimental setting, making the choice of the sampling locations and dates a key factor for the representativeness of the collected data.

In this chapter, we introduce a new method to reveal the main patterns of air-mass connectivity over a large geographical area using the formalism of trajectory-based spatio-temporal networks introduced in the previous chapters, that are particularly suitable for representing complex patterns of connection. To illustrate this approach, we use the coastline of the Mediterranean basin as an application introduced in Chapter 3. The study area, that extends for up to 1600 km vertically and 3860 km horizontally, is divided into 604 regions according to a regular grid, each of which will constitute one node of the network. The edges are then estimated using the formalism proposed in Chapter 3 and then the Mediterranean tropospheric connectivity network is inferred. Thereafter, we characterize the inferred network by computing different metrics revealing the spatial and temporal patterns of connectivity over the study area. The spatial pattern will be the key point helping to detect the regions that act as strong sources or strong receptors of aerial pathogens. The temporal pattern will be the key point for detecting the period during which the connectivity between the sources and the receptors is weak or strong. For the Mediterranean, we identified a seasonal

pattern represented by two main seasons: Summer and Winter. Also, we performed a comparison between the two tropospheric seasonal networks, which has also allowed us to suggest a new strategy to compare spatial weighted networks that are inspired by the small-world property of non-spatial networks.

5.1.1. Graphical outline

In this chapter, we answer the following problem by following the structure presented in the graphical outline 5.1:

- What are the topological characteristics of an estimated tropospheric network?

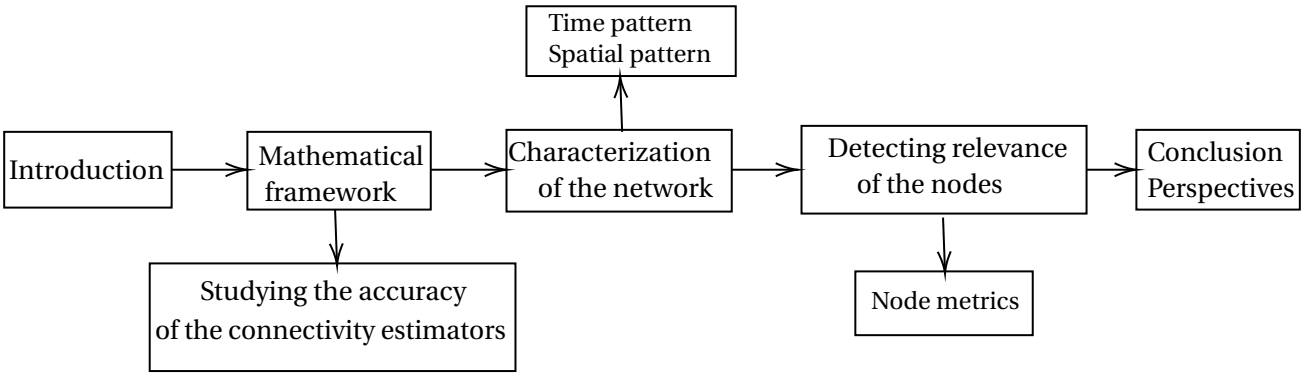


Figure 5.1. – Graphical outline of Chapter 5.

5.1.2. Status of the chapter

Research Article

- This article has been submitted.
- Authors of this article: Maria Choufany, Davide Martinetti, Samuel Soubeyrand and Cindy E. Morris.

5.2. Introduction

Organic particles are ubiquitous in the air (Barberán et al., 2015; Fröhlich-Nowoisky et al., 2012; Womack et al., 2010) and may originate from very different sources (Bowers et al., 2011a), such as plant canopies (Vacher et al., 2016), soils (Abrego et al., 2020), urban areas (Bowers et al., 2011b; Brodie et al., 2007) or surface waters (Powers et al., 2018). Despite their relative sparse density with respect to the volume of an air mass, their presence, and transportation across the planet has proven to have strong effects on many phenomena such as colonization and invasions by plants and

insects (Freeman, 1945; Hampe, 2011; Huestis et al., 2019; Kling and Ackerly, 2020; Kremer et al., 2012), human, animal or plant epidemics (Aylor, 1990; Bogawski et al., 2019; Hiraoka et al., 2017; Leyronas et al., 2018; Mahura et al., 2007; Mundt et al., 2009; Sadyś et al., 2014; Šauliene and Veriankaite, 2006; Wang et al., 2010) and atmospheric processes (Després et al., 2012; Morris et al., 2011, 2017). Some of these particles can be transported through the atmosphere over hundreds or even thousands of kilometers (Aylor, 2003; Barberán et al., 2015; Schmale and Ross, 2015), depending on their shape and mass (Yamamoto et al., 2014). Furthermore, seasonal trends of microbial compositions in the air have been repeatedly observed in several studies, especially in temperate climates (Bowers et al., 2013a; Cáliz et al., 2018; Fahlgren et al., 2010; Fröhlich-Nowoisky et al., 2009; Innocente et al., 2017; Nicolaisen et al., 2017b; Pace et al., 2019; Uetake et al., 2019; Vokou et al., 2012). All these factors make it particularly difficult to disentangle the complexity of the biological composition of air samples, which may include hundreds of species from both local and distant sources, and that can vary drastically over the course of a year.

During the last decades, an increasing number of studies have addressed this problem by collecting and analyzing air samples with a plethora of experimental settings that varied enormously from one study to the other in the number and choice of sampling locations, duration, and frequency of sampling, etc (Šantl-Temkiv et al., 2020). Regardless of these differences, most of these large-scale studies have in common a certain lack of forethought in the choice of sampling locations and dates, that are often dictated by logistical convenience, instead of relying on previous knowledge of air-mass sources. Nonetheless, there are a few exceptions in which the authors reconstructed the geographical origins of the air mass by looking at backward trajectories associated with the air masses from which they have collected their samples (Cáliz et al., 2018; Fahlgren et al., 2010; Innocente et al., 2017; Sarda-Esteve et al., 2019).

Here we present a framework to identify stable and recurrent connections between distant portions of territory via air-mass movements on large spatiotemporal scales. A priori knowledge of the location and seasons of occurrence of such aerial connections would provide a very useful rationale for designing the layout of air sampling schemes for early detection of airborne propagules of invasive plant and insect species, microbial pathogens, and their vectors. We considered the air-mass connectivity across the coast of the Mediterranean basin. The similar climates and vegetation along the northern and southern coasts of the Mediterranean basin would heighten the likelihood of survival of propagules that successfully migrated across the sea. The rather large body of water between the coasts will facilitate demonstrating events of long-distance aerial dispersal by limiting the number of potential intermediate sources. Importantly, the Mediterranean basin is also a hot spot for the spread of human and plant diseases where efforts to survey for invasions and emergences are intensifying. For example, there is a concerted effort to survey insect vectors of zoonotic diseases (Jourdain et al., 2019). This surveillance, which focuses mostly on traditional ground-based observations, can be strengthened in light of recent findings that some insects such as

mosquitos are indeed transported 100's of kilometers by air mass movements at 10's to 100's of meters above ground (Huestis et al., 2019). This raises the question of the most opportune sites and times for observing this long-distance dissemination. We have considered that any pair of points in space can have a certain degree of connection, regardless of their geographical distance, provided that there are recurrent air-mass trajectories that connect the two points. The direction and strength of these connections were estimated by looking at the trajectories linking every pair of points and weighting them. In this perspective, it seemed natural to resort to graph and network theory, since the formalism of nodes and edges provides an adequate environment for describing complex connections and can further be used to deepen the topology of the constructed networks to infer interesting properties of the graphs. We identified two seasonal patterns, one in summer and the other in winter, with relatively distinct behaviors. To characterize the two seasonal networks, we also proposed a new method of comparing spatial weighted and directed networks that accounts simultaneously for the geographical distance and the edge weight between nodes. Finally, we used different indices to assess the relevance of nodes within the network and their likelihood of infecting and/or being infected during a simulated Susceptible-Infected (SI) epidemic on the network. This allowed identifying sets of influential spreaders and strong receptors, which can be used for the design of efficient epidemic surveillance strategies.

5.3. Results

5.3.1. Description of the case study and data collection

Daily 48-hours backward air-mass trajectories from January 1, 2011 to December 31, 2017 were extracted using the HYSPLIT software from a set of 604 arrival points chosen across the study area. The arrival points of the air-mass trajectories, that represent the location of the nodes of the constructed spatial networks, correspond to the centroids of a grid with a mesh size of 74 km covering the coastline of the Mediterranean Sea from 5 km up to 250 km inland from the coast and including the four largest islands (namely Sicily, Sardinia, Cyprus, and Corsica) and the Balearic archipelago. The total number of computed trajectories was 1,543,220.

5.3.2. Network construction and properties

We constructed spatiotemporal networks with discrete (daily) timestamps whose nodes represent the cells of the mesh described earlier. An edge between two nodes i and j at timestamp t is equal to one of the air-mass trajectory arriving at cell j at day t has passed over the cell i during the previous 48 hours. This simple method allowed us to construct 2555 (365× 7) spatial directed networks, but we needed a way to downscale this complex information to fewer, longer periods of time. The first

approach was to concentrate the 2555 daily networks into a single projected static network by averaging the number of connections between each pair of nodes (referred hereafter as 2011-2017 network). Furthermore, we projected the 2555 daily networks at a yearly and monthly cadence, i.e. averaging all the networks of the year 2011, then year 2012, etc. and averaging all the networks of the 7 months of January, then all the networks of the 7 months of February, etc. We then computed different network metrics (diameter, density, transitivity and in- and out-degree correlation) as reported in Table .1 of the Appendix B. The comparison of the metrics for the 7 yearly networks showed no significant difference, meaning that the average yearly behavior has not changed during the 2011-2017 period. On the other hand, the comparison between the metrics of the 12 monthly-averaged networks highlighted a clear seasonal pattern. We hence used a hierarchical clustering method based on the Cut distance between the 12 monthly networks (see (Liu et al., 2018b) and Methods Section for the details) in order to confirm the observed patterns. The clustering algorithm identified two main seasons: summer (from May to September) and winter (from October to April) that have hence been gathered into two projected static networks whose network metrics are shown in Table 5.1 (referred hereafter simply as summer and winter networks). We can observe that the summer network has a significantly larger diameter (intended as the length of the longest of all the calculated shortest path) than the winter network, while it has a lower density (ratio of the number of edges and the number of possible edges). Also, the summer network has the longest average shortest path between any couple of nodes, higher degree correlation (i.e. coefficient of linear correlation between in-degree and out-degree of a node, a.k.a. assortativity) and lower transitivity (e.g. the average probability that the adjacent nodes of a node are also connected).

	Diameter	Density	Transitivity	Average Shortest Path	Degree Correlation
2011-2017	0.003	0.28	0.74	0.0008	0.306
winter	0.006	0.24	0.72	0.0015	0.171
summer	0.974	0.19	0.68	0.0021	0.281

Table 5.1. – Network metrics for the three networks representing the average connectivity during the entire period 2011-2017, and the summer and winter seasons.

One drawback of the metrics used so far is that they only account for the topological properties of the weighted networks, while they overlook the spatial signature that is inherently associated with the graph. The most natural consequence of considering the spatial structure of the network is that we expect geographically close nodes to be more strongly connected than nodes that are farther apart, a phenomenon that is also known as Tobler's first law of geography (Tobler, 1970). Indeed, we found that for the three considered networks, the correlation coefficients between edge weights and

distances are always negative (−0.38 for 2011-2017, −0.35 for winter, and −0.31 for summer), meaning that stronger weights are associated with shorter distances and, vice versa, weaker weights are associated to longer distances. Furthermore, consider the case of three nodes (i, j, k) such that the weight of the edge between i and the other two nodes is the same ($E(i, j) = E(i, k)$), while their geographical distance is significantly different, e.g. $d(i, j) \gg d(i, k)$. The fact that the weight between i and j is the same as between i and k, even if i and j are much further apart in space than i and k may be of great relevance in certain applications. These two considerations imply that in a spatial weighted network, certain edges are more prominent than others, in particular those that maximize, at the same time, weight and geographical distance. Here we introduce a new way of analyzing networks that accounts for these aspects and allows, first, to compare multiple spatial weighted networks, and second, to improve the visualization by pruning the number of edges hence avoiding the so-called 'hairball' effect (Dianati, 2016) (overly dense representation of edges that makes the network undecipherable). In the left panel of Figure 5.2 we depicted the weights and distances of all non-null edges of 2011-2017, summer, and winter networks, also highlighting the Pareto fronts that maximize both quantities. We can observe that the summer network is capable of longer and stronger connections than the winter network (particularly for nodes that are more than 700 km apart), while the winter network has generally stronger connections in the range between 300 and 500 km. In Figure 5.2b we mapped the directed edges that correspond to the 1% of points that lie closest to the Pareto front of Figure 5.2a. We can observe that the edges of the winter network with higher weights and small to medium distances tend to align with the Italian peninsula from North to South plus some edges from France and Spain towards Algeria, while longer edges cross the Mediterranean from South-East to North. Interestingly, we notice that nodes in Southern Italy receive edges both from Northern Italy (medium distance) and from Northeastern Africa (very long distance). For the summer network, we observe that most of the short and medium distance edges start from the Western European coast (France and Spain) and travel eastward towards Northwestern Africa (short distance and strong weight) and Northern Italy, Corsica, and Slovenia (medium distance and weight), while all long-distance edges connect Greece, Lybia and Egypt with Italy and Turkey. As expected, the 2011-2017 network integrates both summer and winter networks characteristics, with seasonal stability that can be observed in the Western Mediterranean with recurrent connections from South-East to North, West, and North-West.

5.3.3. Indices of the relevance of nodes on network structure

We computed local node properties of 2011-2017, summer, and winter networks. Local node properties measure the relevance of single nodes within a network with respect to certain dynamics, such as the flow of information or the spread of disease and their centrality with respect to the topology of the rest of the network. We hence

computed different node relevance indices in order to identify those nodes that play a prominent role in the network, namely betweenness centrality, closeness centrality, coreness centrality, eigenvector centrality, out-degree, in-degree, strength (see Methods section for the details). Furthermore, we simulated an SI epidemic spread starting from each node and recorded two more indices, namely SI persistence and SI frequency, that indicate, for every node i, the percentage of nodes that will be infected when an epidemic starts from i, and the percentage of times that node i has been infected across all simulated epidemics (see Methods for the details on the simulated SI epidemic model). Since the SI persistence index measures the likelihood of a node to spread the disease in case it is the outbreak node, it is expected to be related to other indices representing its outreach capacity and, indeed, we can observe in Figure 5.3 that the SI persistence index is positively correlated with the out-degree (0.78, 0.76 and 0.78 for 2011-2017, winter and summer networks, respectively), the strength index (0.62, 0.54 and 0.57) and the coreness centrality index (0.10, 0.30 and 0.26), while the correlation with the other indices is less straightforward. On the other hand, the SI frequency index measures the likelihood of a node of been infected by an epidemic that started somewhere else in the network, hence we expect it to be correlated with centrality measures, as we can observe in Figure 5.3, where we found that it is always positively correlated with the in-degree (0.59, 0.55 and 0.63 for 2011-2017, winter and summer networks, respectively), the strength index (0.17, 0.16 and 0.22) and the eigenvector centrality index (0.30, 0.56 and 0.69) and negatively correlated with the out-degree (-0.15, -0.17 and -0.17).

Finally, Figure 5.4 depicts the spatial distribution of the SI persistence and the SI frequency indices for the three networks considered here. We can observe that, on average, both indices show higher values in the 2011-2017 network than in the two seasonal networks, possibly due to the higher density of the 2011-2017 network (see Table 5.1). Nonetheless, the overall highest values for SI persistence and frequency are found in the summer season. In terms of SI persistence, the 2011-2017 network has higher values on the coast of France and Spain, Northern Italy, and the Sinai peninsula in Egypt. In winter, the nodes with higher values are located in Northeastern Italy, while non-negligible values are also found in the Balkans, Eastern coast of Egypt, and the coast of France and Spain. In summer, the highest values are found on the Eastern coast of Egypt, Southern Greece, and the Northern coast of Spain. In terms of SI frequency, the 2011-2017 network has relatively low values of the index that are mainly located around the Aegean and Adriatic Seas, plus Tunisia and Libya. In winter, moderate values of SI frequency are found in Greece, Libya, and Southern Italy. In the summer season, we observe the overall highest values concentrated in Central and Northeastern Italy, Slovenia, and Croatia, with moderate values in Libya.

5.4. Discussion

Understanding air-mass movements is a fundamental step for predicting how airborne microorganisms circulate across the planet, from their sources to their sinks via the atmosphere. Previous studies assessing microbial populations in air samples have focused on a reduced number of sampling sites and/or dates, while the choice of the sampling sites and dates is crucial in the experimental design. In this study, we considered the air-mass movements over a vast geographical region (the coastline of the Mediterranean basin) for an extended period of time (7 years with a daily resolution) in order to assess spatio-temporal patterns of connectivity between different locations using the formalism of spatio-temporal networks. We identified two distinct seasonal patterns in terms of connectivity over the study period, one from May to September (here referred as summer) and the second from October to April (winter), while the yearly regime of connectivity seems to be rather constant across the years. This observation resonates with analogous conclusions drawn from studies of microbial composition in air samples, both in the Mediterranean basin (Cáliz et al., 2018; Innocente et al., 2017; Pace et al., 2019; Vokou et al., 2012) and elsewhere (Bowers et al., 2013a; Fahlgren et al., 2010; Fröhlich-Nowoisky et al., 2009; Nicolaisen et al., 2017b; Uetake et al., 2019). Whether this pattern is due to a seasonal variation in the microbial sources, to the seasonal patterns of air-mass connectivity, or to a combination of both, is by itself an interesting research question. From the methodological point of view, we introduced a new way of analyzing spatial weighted networks that accounts for both edge weight and geographical distance, and that bears a certain resemblance to the small-world property of non-spatial networks, a property that is known to favor epidemic spread (Xu and Sui, 2009). Let us recall it briefly: a network has the small-world property (Watts and Strogatz, 1998) if most nodes are not connected to each another (low density), if the neighbors of any given node are likely to be neighbors of each other (high clustering), and if any given node is likely to be reached from every other node with a small number of steps (low average shortest path). In the context of spatial networks, the definition given above is no longer appropriate, since there are two ways of defining the neighborhood of a node and the length of the shortest path between nodes: the classical geographical one and the one given by the network topology. Nonetheless, if we consider the examples presented in the manuscript, we observe that the summer network has a joint distribution of edge weights and distances that dominates the distribution of the winter network when weights are low and distances are high (shortest paths can cross long distances) and the other way around, when weights are high and distances are low (high spatial clustering), while the winter network's joint distribution dominates the summer one for the middle range of weights and distances. We could then conclude that the summer network has more of the small-world property than the winter network, and this is confirmed by the fact that both SI persistence and SI frequency indices are overall higher in the summer than in the winter network. Finding a way of quantifying this difference is an

expected extension for the present work. Finally, the pruned network representation of Figure 5.2b allowed to identify strong spreaders (e.g. Northern Italy in winter vs. France, Spain, and the Northeastern coast of Africa in summer) and strong receptors (e.g. Italy in both winter and summer). Identifying strong spreaders and receptors is a key step for designing sampling campaigns. For example, in the context of epidemic surveillance of airborne diseases, it is important to optimize the allocation of sampling sites and dates in order to increase the detection rate and to reduce the delay between arrival and first detection. In this study, we identified a high-risk zone in Northeastern Italy, Croatia, and Slovenia that is particularly receptive during the summer months (i.e. high SI frequency index), that correspond, at these latitudes, to the growing season of most crops. On the other hand, epidemic outbreaks starting on the coast of France and Spain, Northern Italy and the Sinai peninsula are the ones showing the highest risk of rapid diffusion across the Mediterranean basin, based on the SI persistence maps.

5.5. Methods

5.5.1. Case study region and data collection

The study region corresponds to the coast of the Mediterranean Sea, ranging approximately 1,600 km from North to South and 3,860 km from East to West. The temperate climate of the chosen region is strongly influenced by the presence of the Mediterranean Sea, with mild winters, hot summers, and relatively scarce precipitations. The landscape is characterized by coastal vegetation, typically shrubs and pines, and densely populated areas with intensive crop production of wheat, barley, vegetables, and fruits, especially olive, grapes, and citrus. In this paper, we characterize recurrent movements of air masses through the Mediterranean region by defining a grid with mesh size 74 km covering the coastline from 5 km up to 250 km inland from the coast, including the four largest islands (namely Sicily, Sardinia, Cyprus, and Corsica) and the Balearic archipelago. Thus, we divided the region into $N = 604$ cells, where the centroids of the cells will be used as arrival locations of air-mass trajectories and will correspond to the nodes of the constructed network. The air-mass trajectories arriving at the prescribed locations in the period 2011-2017 with daily basis (hence $T = 365 \times 7 = 2555$) have been computed using the Hybrid Single-Particle Lagrangian Integrated Trajectory model (`HYSPLIT` (Stein et al., 2015)). The `HYSPLIT` model has been fed with meteorological data from the Global Data Assimilation System files with a 0.5-degree spatial resolution (GDAS: https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-data-assimilation-system-gdas) and was parametrized to return 48-hours backward air-mass trajectories arriving at the prescribed locations at 12:00 a.m. at an altitude of 500 m above mean sea level. A single trajectory consists of a vector containing the hourly positions (longitude, latitude, and altitude) visited by the air mass before arriving at the specified location and time.

5.5.2. From air mass trajectories to daily contact networks

To infer spatio-temporal networks according to air-mass trajectories, we used the formalism of graph theory, where nodes are defined as spatial points (the centroids of 604 cells formed by the mesh polygons of size 74 km covering the coastline) and edges are estimated using the following rule: a node i is connected to node j at time step t (i.e. the element $E_t(j,i)$ of the adjacency matrix E_t is equal to 1) if the trajectory arriving at i at time t has crossed the polygon whose centroids is j over the interval $[t-48h, t]$. Since the trajectories have been computed for every i and every t, we have T adjacency matrices E_t, $t \in [0, T]$ of size $N \times N$, each representing a directed spatial network. (See (Choufany et al., 2019a) for alternative connectivity measures)

In this paper, we characterize recurrent movements of air masses through the Mediterranean region by defining a grid with mesh size 74 km covering the coastline from 5 km up to 250 km inland from the coast, including the four largest islands (namely Sicily, Sardinia, Cyprus, and Corsica). Thus, we divide the region into 604 cells, where the centroids of the cells will be used as arrival locations of air-mass trajectories and will correspond to the nodes of the constructed network.

5.5.3. From daily contact networks to aggregated spatio-temporal networks

The information contained in the T spatial networks has been summarized by averaging all daily networks within a certain subset $S \subseteq \{0, T\}$ as follows (Holme and Saramäki, 2019):

$$E_S(i,j) = \sum_{t \in S} E_t(i,j)/|S|,$$

where $|S|$ denotes the cardinality of S. The chosen subsets are the entire interval $[0, T]$, the 7 years from 2011 to 2017, and the subsets of all days of the month of January, all the days of the month of February, etc.

5.5.4. General network metrics

The constructed networks are inherently complex by the sheer amount of spatial and temporal information that they encompass. Hence, there is no easy way of representing the results either graphically or numerically, without compromising the original complexity of the networks. In this aim, we explore the topology of the networks by looking at some generic properties through the following metrics (Csardi and Nepusz, 2006):

- Diameter: the longest of all possible shortest paths between any two pair of nodes computed using Dijkstra algorithm (Dijkstra et al., 1959) (this metric does not account for the geographical distance between nodes, but only the weight of the edge between them),

- Density: the ratio between the sum of all edge weights and the number of all possible edges (Liu et al., 2009),
- Transitivity (also known as clustering): the equivalent definition of density, but applied to triplets of nodes instead of pairs of node (Opsahl and Panzarasa, 2009),
- Shortest Path Length: the average of the shortest path between any possible pair of different nodes, computed using Dijkstra algorithm (Dijkstra et al., 1959) (this metric does not account for the geographical distance between nodes, but only the weight of the edge between them),
- Degree Correlation: in directed networks, it accounts for the correlation between the incoming and the outgoing degree of a node. Networks with positive (resp. negative) degree correlation foster (resp. hamper) epidemic spread (Pautasso et al., 2010).

5.5.5. Cut Distance

The Cut distance (Liu et al., 2018b) is a particularly suitable and elegant way for comparing weighted and directed networks having the same number of nodes (Tantardini et al., 2019) and it is based on the mathematical formalism of the cut distance or rectangle distance presented in (Frieze and Kannan, 1999). We do not present here the mathematical formulation of the distance and we invite the interested reader to refer to (Liu et al., 2018b) for the details. Since the algorithm for computing the Cut distance between two networks is based on the maximisation of a certain function over all possible pairs of disjoint and complementary subsets of nodes of a network, the problem becomes quickly unfeasible as the number of nodes increases (NP-hard problem (Liu et al., 2018b)). In order to solve this problem, we used a genetic algorithm from the `R` library `GA` (Scrucca, 2017).

5.5.6. Indices of network nodes relevance

Single node relevance can be computed in a multitude of ways and many indices have been proposed in the literature (Lü et al., 2016). Here we consider seven among the most widely used indices for weighted and directed networks: betweenness and closeness centralities (Freeman, 1978), coreness (Seidman, 1983), eigenvector centrality (Bonacich, 1987), in-degree (i.e. the sum of the weights of the edges pointing toward a node), out-degree (i.e. the sum of the weights of the edges outgoing from a node) and strength (i.e. the sum of in- and out-degree).

5.5.7. Susceptible-Infected epidemic model

In order to model the spread of an epidemic over the constructed spatiotemporal networks, we simulated a classical SI (Susceptible – Infected) compartmental

model (Keeling and Eames, 2005; Moreno et al., 2002; Newman, 2002; Pastor-Satorras et al., 2015). In this model, each individual can be assigned to two distinct states: susceptible or infected. Simulations start with all nodes being susceptible, except for a single inoculated initial node. At each time step, susceptible nodes become infected if the maximum weight of their infected contacts is over a certain threshold (here arbitrary set to 0.1). All simulations run for 6 time steps. We simulated $N = 604$ SI epidemics for each of the three considered periods (2011-2017, summer, and winter) by changing the initially infected node across all possible nodes of the networks, the rest of the parameters being constant. For every node $i \in \{1, \ldots, N\}$ and every period $G \in \{2011-2017, \text{summer}, \text{winter}\}$, we recorded two values: the SI persistence $P_G(i)$ that represents the maximum percentage of infected nodes at the end of the simulated epidemic that started from node i and spanned the period G and the SI frequency $F_G(i)$ that represents the percentage of times that node i has been infected across all the epidemics ran in the G period.

5.5.8. Software

Air-mass trajectories have been computed using the `HYSPLIT` (Stein et al., 2015) software installed on local cluster `https://informatique-mia.inrae.fr/biosp-cluster/cluster`. All the rest of the computations and graphics have been performed using the statistical software `R`, in particular using the packages `sf` (Pebesma, 2018), `ggplot2` (Wickham, 2016), `igraph` (Csardi and Nepusz, 2006), `GA` (Scrucca, 2017) and `corrplot` (Wei and Simko, 2017).

5.6. Acknowledgements

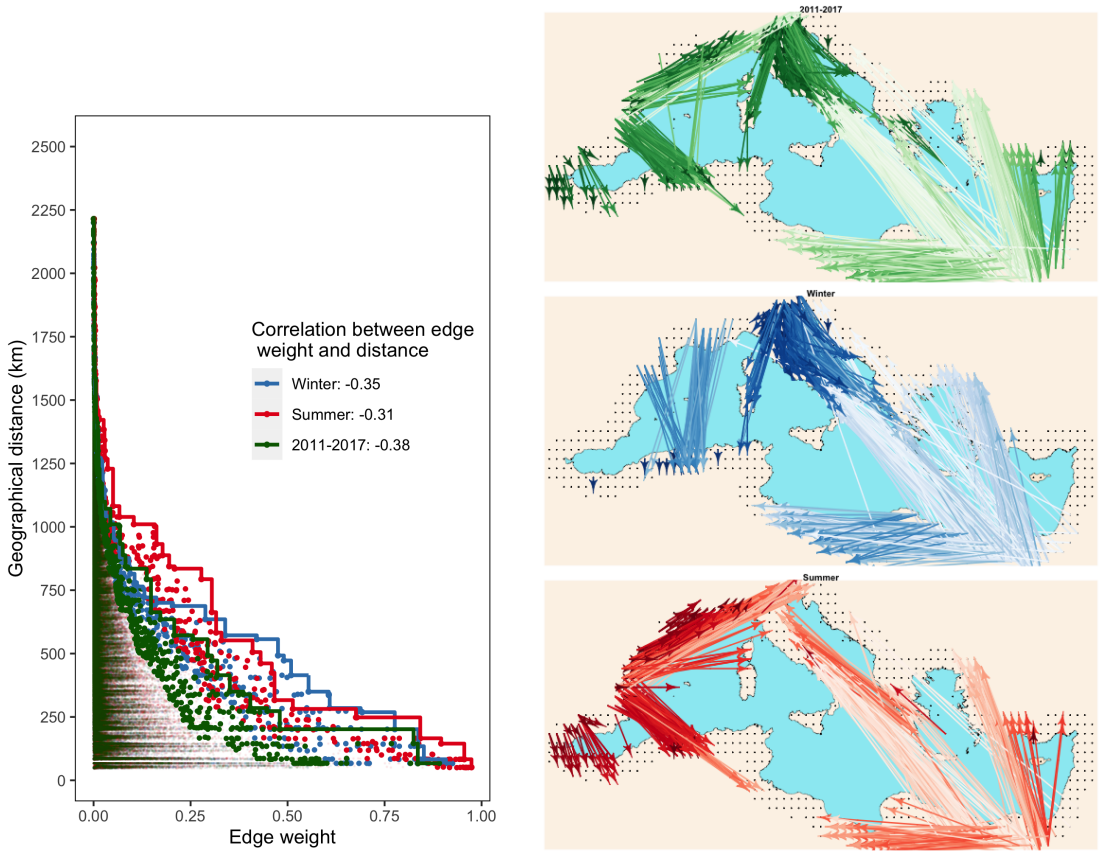
This research was funded by the SPREE project from the French National Research Agency (grant n. ANR-17-CE32-0004-01) and the PHYTOSENTINEL project (grant n. IB-2019-SPE). The authors thank Loïc Houde for his technical assistance in the calculation of trajectories with `HYSPLIT`.

5.7. Author contributions statement

M.C. and S.S. developed the theory; M.C., D.M., C.E.M., and S.S. designed the experiment. M.C. and D.M. performed the data extraction and analysis, and prepared all tables and figures. All authors contributed to writing the manuscript and approved the final version.

5.8. Competing interests

The authors declare no competing interests.



(a) Pareto Fronts

(b) Maps

Figure 5.2. – Panel a) shows the distribution of weights and distances for all pairs of non-null edges of 2011-2017, summer, and winter networks. The Pareto fronts and the 1% of the edges that lie closest to the Pareto fronts are depicted with a bigger dot. Panel b) represents the 1% of the edges that lie closest to the Pareto fronts, where the intensity of the color correspond to the strength of the weight.

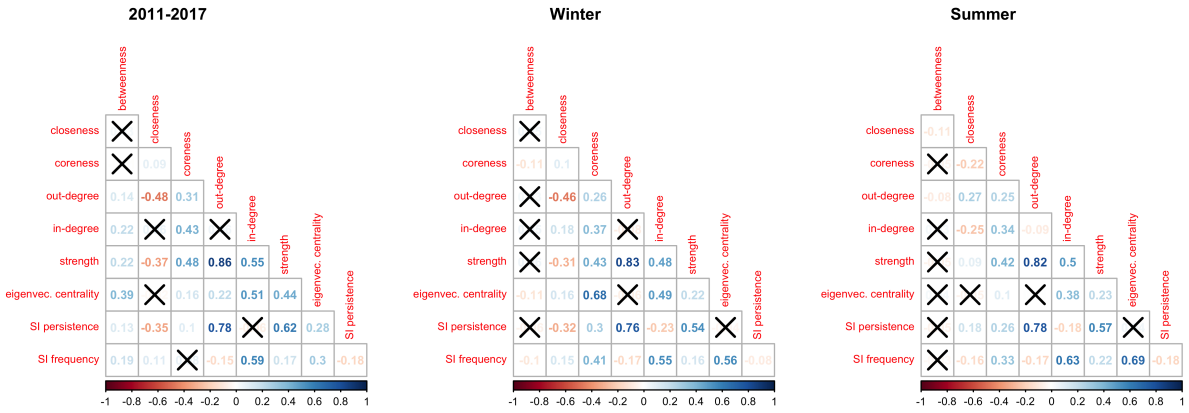


Figure 5.3. – Correlation plots of node indices. All reported correlation coefficients are statistically significant at 95% confidence level.

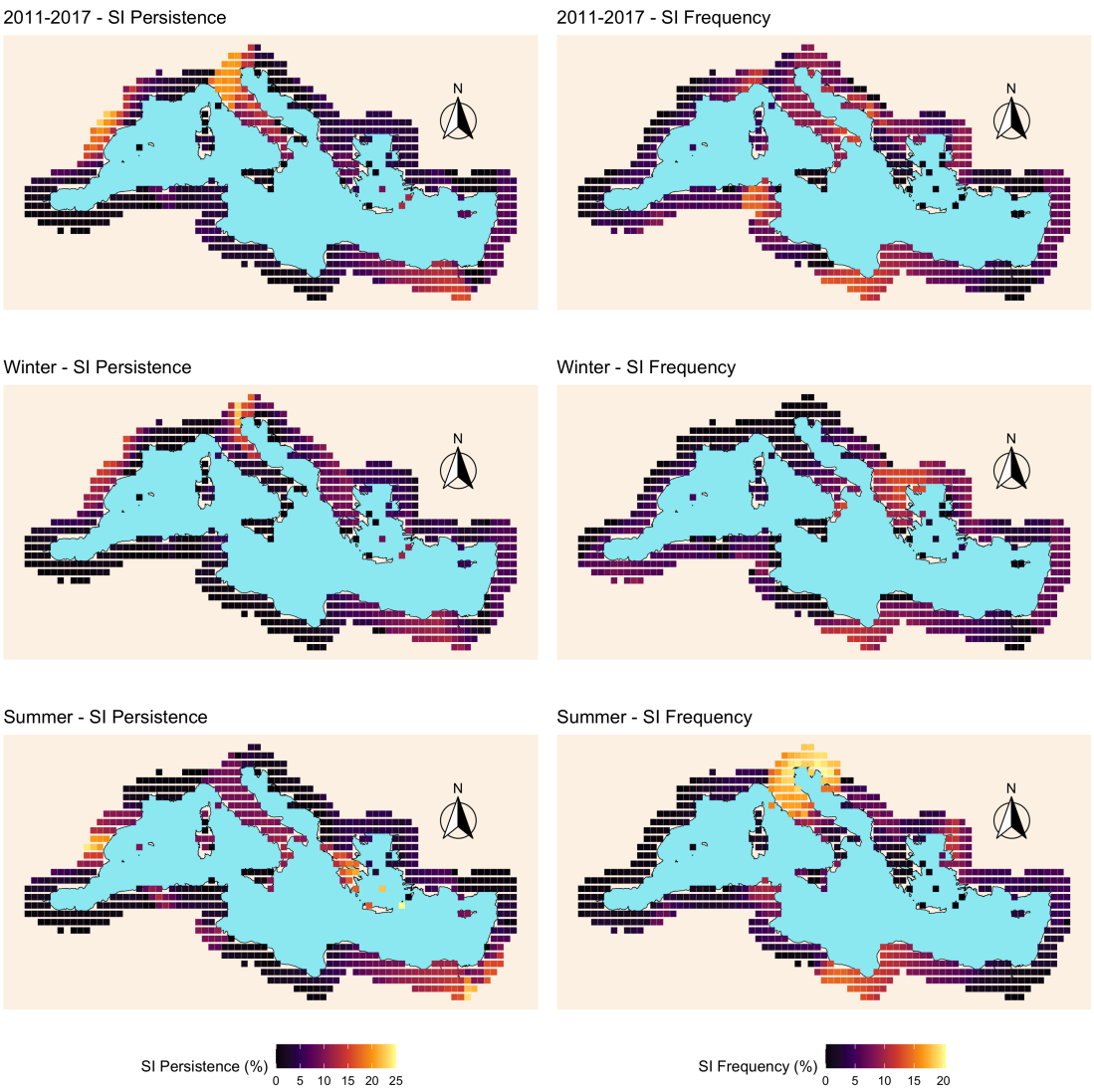


Figure 5.4. – SI Persistence and SI Frequency.

5.9. Conclusion notes

Take home messages

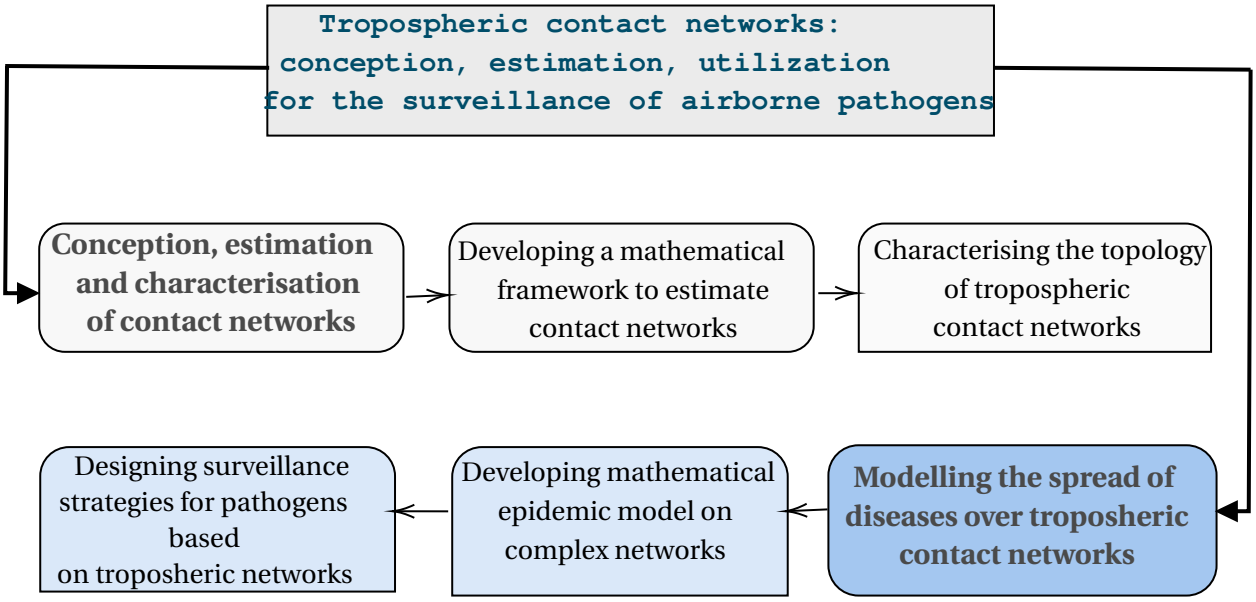
- Based on the inferred Mediterranean tropospheric network, we confirmed the seasonal pattern of air mass trajectories detected in Chapter 3, by using a more advanced method based on *cut distance*. The seasons roughly correspond to Summer (from May to September) and Winter (from October to April).
- We introduce a new measure for characterizing spatial complex networks that are based on the maximization of the geographical distances and edge weights between all pairs of nodes. This new measure allows us to detect the distant points with strong connectivity and it is inspired by the small-world property of non-spatial networks. Our results show that the summer network has more of the small-world property than the winter network.
- By simulating an epidemic model of type SI conditional to the tropospheric connectivity network, we identified those geographic areas that could act as strong spreaders by computing the SI-persistence and strong receptors by computing SI-frequency throughout the Mediterranean basin.

Perspectives

- The seasonal pattern detected in this work could be compared to the seasonal variation of the microbial populations sampled from the air within the study region.
- We need new ways of analyzing and comparing spatial networks that have to account for the possibility of spatial auto-correlation between close nodes.
- After identifying the critical areas, one could conceive and optimize different spatio-temporal sampling strategies that can be effective for the early detection of the emergence of new airborne pathogens. **This perspective is the key point of Chapter 6**.

6. Spatio-Temporal surveillance on complex networks: An epidemic prevention perspective

Table of Contents



6.1. Abstract

Infectious diseases spread between susceptible individuals can be tracked down by looking at the network of contacts between hosts. For certain diseases, reconstructing the network of contacts is a challenge by itself, but recent technical developments made possible the estimation of the contact structure generated by the flow of air masses over a large geographical territory. Airborne dispersion is a recognized and challenging pathway of dissemination of certain plant pathogens, whose arrival in new territories can trigger outbreaks and new epidemics.

In this chapter, we use the framework presented earlier in the manuscript to infer tropospheric connectivity networks in order to simulate the spread of an airborne disease over the French region PACA by means of a susceptible-Infectious-Recovered-Susceptible epidemic spread model. Then, we design different spatio-temporal surveillance strategies that should indicate where and when to conduct surveillance. Finally, we evaluate the efficacy of the proposed strategies in terms of frequency of detection over thousands of simulated epidemics.

Our results highlight the importance of using network-based strategies in order to anticipate the detection of network-dwelling epidemics and we also obtained reassuring indications about the surveillance effort that is needed in order to guarantee an effective early-detection, since we show that few surveillance sites and events could be very effective if chosen wisely.

6.1.1. Graphical outline

In this chapter, we answer the following questions by following the structure presented in the graphical outline 6.1:

- How can infectious diseases spread through a tropospheric network?
- What are the most effective spatio-temporal surveillance strategies for boosting the early detection of an epidemic?

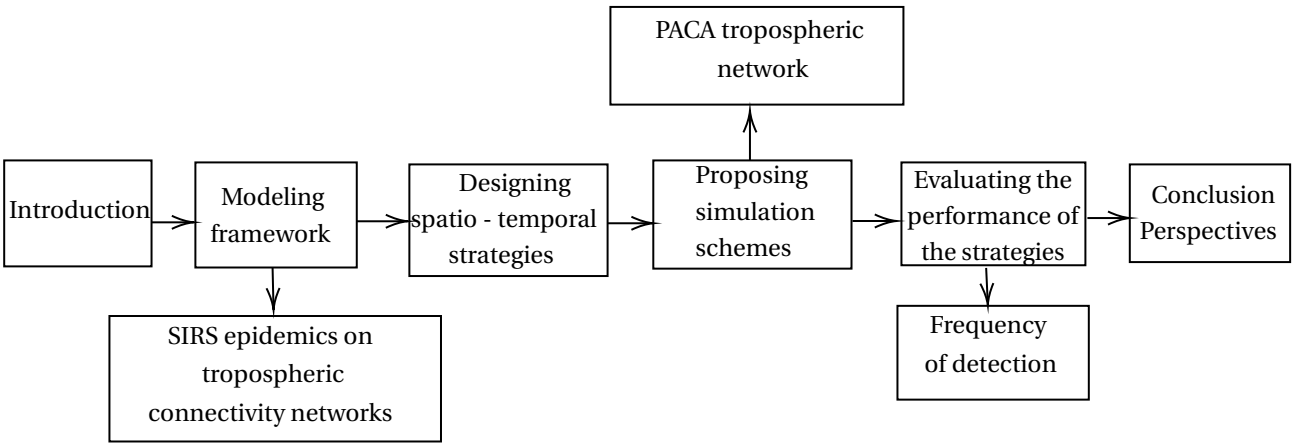


Figure 6.1. – Graphical outline of Chapter 6.

6.1.2. Status of the chapter

Research work

- This work is in progress.
- Authors of this chapter: Maria Choufany, Samuel Soubeyrand, Cindy E. Morris and Davide Martinetti.

6.2. Introduction

The impact of infectious disease outbreaks affecting plant populations can be major if the epidemics are not detected and controlled early enough (Parnell et al., 2015; Severns et al., 2019). Until today, risk managers, plant health inspectors, and researchers are mobilized to detect *where* and *when* new epidemic will strike by setting up appropriate surveillance systems to anticipate disease outbreaks in its early stages (Bai et al., 2017; Colman et al., 2019; Holme, 2018; Mastin et al., 2017). The sooner we take action to halt the spread of the diseases, the more we reduce human and plant suffering and economic loss. Moreover, by surveying, we prevent large epidemics or pandemics

due to the dispersion of the diseases (Colman et al., 2019), and several studies already proved that the cost of prophylaxis and surveillance is orders of magnitude lower than the cost of controlling an endemic disease (Schneider et al., 2020). Given the huge cost of surveying all the susceptible hosts, it is more suitable to identify the individuals with a high likelihood to get contaminated early during an infectious disease outbreak. This reasoning is known as the sentinel surveillance problem (Arita et al., 2004; Holme, 2018; Polgreen et al., 2009).

Plant diseases can spread over short and long geographical distance, and one of the long-range pathways of dissemination of certain pathogens is via air masses, which are responsible for the long-distance spread of emergent infectious diseases that can extend from hundreds of meters to thousands of kilometers (Brown and Hovmøller, 2002; Schmale III and Ross, 2015). Hence, modeling air mass trajectories to assess the epidemic spread is important because these trajectories can transport a significant volume of propagules for thousands of kilometers. Nonetheless, modeling air mass movements is not trivial, since the direction of the flow is constantly changing and cannot be observed directly, as for example in the case of river flows. Hence we resort to historical meteorological data in order to reconstruct past movements of air-masses over a large geographical area and over an extended period of time. Then, we summarize these movements by means of contact networks, that is a convenient tool for representing the connections between the individuals of a population, in our case, geographical areas hosting susceptible plants. Finally, our method allows us to represent dynamic contact by means of spatio-temporal networks, where the presence and strength of the links can vary over time.

Over these networks, it is possible to define mathematical epidemic spread models that take into account a realistic pattern of connections (Bansal et al., 2010; Moslonka-Lefebvre et al., 2011; Newman, 2002; Pastor-Satorras et al., 2015; Pautasso and Jeger, 2014). One of the basic mathematical models for representing the dynamics of infectious diseases is the SIR (Susceptible, Infected, Recovered) model introduced by Kermack and McKendrick (1932). The population in this model is divided into three classes where the individuals begin initially in the susceptible class and evolve to the infected class if they have sufficient contacts with infectious individuals. Infectious individuals can then recover and eventually become susceptible again after a certain time.

In this article, we use the approach presented in Choufany et al. (2019b) and compartmental models to simulate the spread of an airborne disease over inferred tropospheric networks through different scenarios, and the work of Martinetti and Soubeyrand (2019), Bai et al. (2017) and Holme (2018) to implement surveillance strategies over the networks (by targeting early detection). The objective of the simulations is to provide a conspicuous set of cases with different initial conditions where we can test the efficacy of diverse surveillance strategies. Our surveillance strategies are designed according to the features of the networks, i.e. they are both spatial and temporal, indicating where and when to conduct surveillance. The choice and design

of such strategies is based on centrality measures of the constructed networks and they are compared to random and classical spatial strategies. The comparison is performed with respect to a performance measure, the frequency of detection, that measures the likelihood of a surveillance strategy of detecting an outbreak over all the generated simulations. Finally, we test different budgetary scenarios, from very few and isolated surveillance events/sites to more comprehensive strategies with many sites and frequent surveillance events. Our results highlight the importance of using network-based strategies in order to anticipate the detection of network-dwelling epidemics and we also obtained reassuring indications about the surveillance effort that is needed in order to guarantee an effective early-detection, since we show that few surveillance sites and events could be very effective, if chosen wisely.

6.3. Simulated SIRS epidemics on tropospheric connectivity networks

In this manuscript, we aim at modeling epidemics of airborne diseases that can be transmitted via the movement of air masses. This movement can be estimated from past observations of air-mass trajectories over a given region and for a certain period of time T (Choufany et al., 2019b; Leyronas et al., 2018). The patterns of connectivity are represented through spatio-temporal networks where nodes represent a certain partition of the study region into N sub-regions, while edges represent average connectivity between pairs of nodes and they can change over time. In particular, this construction allows us to represent the intensity of tropospheric connectivity between distant nodes during the period of time T by using weighted directed adjacency matrices. In this section, we present a framework for simulating SIRS epidemics over such spatio-temporal networks.

Traditional compartmental SIR epidemic models (Kermack and McKendrick, 1932) describe the dynamics of an infectious disease within a population of N individuals that are divided into three distinct and non-overlapping compartments: susceptible (S), infectious (I), and recovered (R). The order of the S-I-R compartments usually indicates the phases through which hosts progress, from susceptible to recovered. Different variations of the standard SIR model have been proposed in the literature to account for non-lethal non-immunizing diseases without the R compartment (SI or SIS; Hethcote, 1994, 2000), diseases with a non-infectious exposition period (SEIR, where E stands for Exposed; Lekone and Finkenstädt, 2006) or diseases that do not confer lifelong immunity (SIRS; Lahrouz et al., 2011). In general, these types of models describe how individuals pass from one compartment to another. In this work, we consider the hypothesis of homogeneous mixing of individuals, i.e. all hosts have identical rates of disease-causing contacts. Whereas this assumption makes the analysis tractable using ordinary differential equations, once epidemiological parameters such as the infection and recovery rates have been defined, it may not

adequately reflect the complex reality of heterogeneous contact patterns between individuals (Bansal et al., 2007). A more realistic, but less tractable approach is to assume a known underlying contact structure, that can be conveniently represented through a network of contacts between individuals (Newman, 2002).

Formally, a network $G = (V, E)$ is defined by set of nodes $V = \{1, 2, \ldots, N\}$ connected by a set of edges E. A convenient way of representing a network is by means of an adjacency squared matrix $E(i, j)$, such that $E(i, j) > 0$ if and only of nodes i and j are connected in the i–to–j orientation ($i, j \in \{1, 2, \ldots, N\}$), $E(i, j) = 0$ otherwise. We say that the network G is directed if the corresponding adjacency is not symmetric, while we say that it is weighted if the adjacency matrix can take any value in the set of real numbers. In this particular exercise, we are going to consider spatio-temporal networks, in which the nodes represent some fixed geographical locations with known geographical coordinates. Furthermore, since the connectivity between sites may change over time, the set of edges E may vary during the time period $[0, T]$, that is hence divided into l intervals $\{[0, t_1), [t_1, t_2), \ldots, [t_{l-1}, t_l)\}$, each of which being associated to a specific network of connections $G_s = (V, E_s), s \in \{1, \ldots, l\}$. Notice that the set of nodes V remains fixed, meaning that the nodes of the networks remain unchanged during the entire period (no node is removed or added). On the other hand, the set of edges E_s of each G_s may vary, meaning that some connections can appear or disappear, and change their orientations or weights.

We chose to simulate a SIRS epidemic model, in which, at any given time, every node can only be in one of the three possible states: susceptible, infectious or recovered. Let us introduce a set of notations that will be useful in the description of the SIRS mechanisms. We will denote with $S_i(t) \in \{0, 1\}$, $I_i(t) \in \{0, 1\}$ and $R_i(t) \in \{0, 1\}$ the state of a node i at time t. Recall that, for every $t \in [0, T]$, $S_i(t) + I_i(t) + R_i(t) = 1 \ \forall i \in \{1, \ldots, N\}$ and $\sum_{i=1}^{N} (S_i(t) + I_i(t) + R_i(t)) = N = |V|$, N being the total number of nodes of the network, that is kept constant all over the study. Let $S(t)$, $I(t)$ and $R(t)$ be the three disjoint sets containing the susceptible, infectious and recovered nodes at time t respectively and such that $S(t) \cup I(t) \cup R(t) = V, \forall t \in [0, T]$. Let $V_t(i) = \{j \in \{1, \ldots, N\} : j \neq i, E(j, i) > 0\}$ be the set of nodes pointing towards i at time step t. We define $V_t^I(i) = V_t(i) \cap I(t-1)$ the set of nodes pointing towards i at time t and that were infectious at time $t-1$.

Before detailing the transitions between compartments S, I, and R, we introduce a few notations. Let $C : [0, 1]^2 \to [0, 1]$ be a binary operator satisfying the following properties, for all $a, b, c, d \in [0, 1]$:

— commutativity: $C(a, b) = C(b, a)$,

— monotonicity: $C(a, b) \leq C(c, d)$, if $a \leq c$ and $b \leq d$,

— associativity: $C(a, C(b, c)) = C(C(a, b), c)$,

— identity element: $C(a, 0) = a$.

Triangular conorms (or T-conorms), to which belong the maximum T-conorm $C_M(a, b) = \max\{a, b\}$, the probabilistic sum $C_P(a, b) = a + b - a \cdot b$ and the bounded sum (a.k.a.

Łukasiewicz T-conorm) $C_L(a,b) = \min\{a+b,1\}$, are examples of operators satisfying the previous set of conditions (Klement et al., 2004). Taking advantage of the associativity property of C, we will denote with $\underset{a\in X}{C}(a)$ the sequential application of C to all the elements a of a set X, as for the standard $\sum$ operator. Let f_β be a function that selects a random subset from a given set of nodes X ($f_\beta(X) \subseteq X$) such that each element of X has probability $\beta \in [0,1]$ of being chosen and the selection of the elements are independent. Finally, we define over $\mathbb{R}^+$ the exponentially decaying function $\mathscr{E}_\delta : t \mapsto 1 - e^{-\delta t}$ with $\delta \geq 0$.

In order to describe the progress of the SIRS epidemic model, whose state can only evolve on a discrete-time basis at integer times 1, 2, ..., we need to define the state of each node at each time $t \in \mathbb{N}$, knowing that the state of a node at time t depends on its state at time $t-1$ and on the state of its neighboring nodes according to the following set of rules:

— At the time $t = 0$, all nodes are considered susceptible except for a single randomly-chosen node that is inoculated.

— At every iteration, each susceptible node i can become infected if the following conditions are satisfied:

$$S_i(t-1) = 1 \quad \text{and} \quad \underset{j\in f_\beta(V_t^I(i))}{C}(E_t(j,i)) > u\,.$$

where u is a threshold randomly sampled (for each (t,i)) from a uniform distribution in $[0,1]$ and C is the Łukasiewicz T-conorm. In words, a node i that is susceptible at time $t-1$ becomes infected at time t if and only if the conorm C of the edge weights of a randomly chosen subset of its infected neighbors that point toward it, is larger than u.

— At every iteration, each infectious node can become recovered if the following conditions are satisfied:

$$I_i(t-1) = 1 \quad \text{and} \quad \mathscr{E}_{\delta_1}(T_I(i)) > v\,,$$

where $T_I(i)$ denotes the number of successive iterations up to time $t-1$ in which node i remained in the infectious state, $\delta_1 \in \mathbb{R}^+$ is a fixed parameter and $v \in [0,1]$ is sampled randomly from a uniform distribution in the unit interval. This means that the chance for an infectious node at time $t-1$ to recover at time t increases with the time spent in the infectious state.

— At every iteration, each recovered node can return to the susceptible state if the following conditions are satisfied:

$$R_i(t-1) = 1 \quad \text{and} \quad \mathscr{E}_{\delta_2}(T_R(i)) > w\,,$$

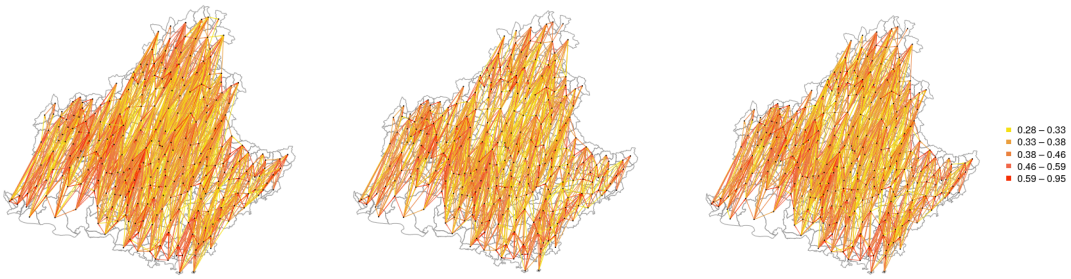
where $T_R(i)$ denotes the number of successive iterations up to time $t-1$ in which node i remained in the recovered state, $\delta_2 \in \mathbb{R}^+$ is a fixed parameter and

$w \in [0,1]$ is sampled (for each (t,i)) from a uniform distribution in the unit interval. This means that the chance for a recovered node at time $t-1$ to return to the susceptible state at time t increases with the time that it has spent in the recovered state.

6.4. Simulation settings

6.4.1. Geographical context and tropospheric connectivity networks

In this work, we consider the tropospheric connectivity over the South-Eastern French region of Provence-Alpes-Côte d'Azur (PACA hereafter). It is one of the metropolitan French regions with the highest plant diversity, distributed over a heterogeneous and complex Mediterranean agricultural landscape. Its alpine areas are mostly covered by natural vegetation, pastures, and grasslands, while most of its agricultural lands are cultivated with high yielding crops such as olive, vineyards, lavender, vegetables, and orchards. The tropospheric connectivity network considered in this study has been computed over the 294 watersheds of the region, which hence constitute the nodes of the spatio-temporal network, and the edges linking watersheds have been inferred by considering the trajectories of air masses movements from 2011 to 2017 as detailed in Choufany et al. (2019b). In that study we already found two strong seasonal patterns of connectivity across the PACA region that seemed rather stable through the years, one in Summer (Figure 6.2a), going from April to October and the other in Winter (Figure 6.2b), from November to March. Finally, we also consider the average yearly connectivity network (Figure 6.2c).



(a) Summer season (b) Winter Season (c) Whole period

Figure 6.2. – Networks weighted by the connectivities generated by air mass movements across PACA during the (a) Summer season, (b) Winter season, and (c) whole period.

6.4.2. Simulation schemes

We set up 5 simulation schemes by letting the tropospheric connectivity matrix vary in the course of the simulations. For the sake of comparison, all simulation schemes have the same duration (number of iterations) that has been set to 150, but we distinguish between the following cases:

Scheme 1. Winter - Summer: first 75 iterations using the winter connectivity network, then 75 iterations with the summer connectivity network;

Scheme 2. Summer - Winter: first 75 iterations using the Summer connectivity network, then 75 iterations with the Winter connectivity network;

Scheme 3. Year: 150 iterations with the average yearly connectivity network;

Scheme 4. Only Winter: 150 iterations with the winter connectivity network;

Scheme 5. Only Summer: 150 iterations with the summer connectivity network.

The simulations are not intended to reproduce any specific disease, but rather to explore diverse scenarios of diffusion of an airborne pathogen over the study area. Therefore, the values of the SIRS parameters are chosen in such a way that the simulated epidemics show a reasonable and realistic behavior, i.e. a steady progression without dying out or infecting the entire population too rapidly. The chosen parameters are as follows: $\beta = 0.5$, $\delta_1 = 0.013$ and $\delta_2 = 0.013$. Finally, each scheme is simulated 10000 times.

6.5. Spatio-temporal surveillance strategies

6.5.1. Strategy design

A spatio-temporal surveillance strategy consists of choosing a subset of k nodes of the network and s instants in the time period $[0, T]$ at which the k nodes are screened for the presence of the disease (under the simplifying assumption that tests have perfect detection accuracy). The objective of the strategy is to detect the disease as early as possible, and the efficiency of a given strategy will be assessed using appropriate measures that will be detailed in the following section. An interesting point in the design of such strategies is that both the surveillance nodes and the frequency of the surveillance may vary over time. Based on the spatial strategies proposed by Martinetti and Soubeyrand (2019), Herrera et al. (2016) and Colman et al. (2019), and the temporal strategies proposed by Bai et al. (2017) and Holme (2018), we couple two temporal surveillance strategies and six spatial strategies to obtain 12 spatio-temporal surveillance strategies. The six spatial surveillance strategies are based on the following rules:

Ran. Random: select k nodes at random.

Reg. Regular: select k nodes within a regular grid of k cells over the study area.

ND. Normalized degree: select the k nodes having the highest normalized degrees, this latter being defined by the ratio of the degree of the node and its spatial area.

In. In-degree: select the k nodes having the highest in-degree values, this latter being defined as the sum of the weights of the edges pointing at a given node.

Out. Out-degree: select the k nodes having the highest out-degree values, this latter being defined as the sum of the weights of the edges that originate from a given node.

Bet. Betweenness centrality: select the k nodes having the highest betweenness scores, this latter being defined as the fraction of shortest paths passing by a given node. This metric evaluates the degree of centrality of a node with respect to the path connecting other nodes.

While the 2 temporal surveillance strategies are defined as follows:

Ran. Random: select s random instants from the interval $[0, T]$.

Reg. Regular: select s instants at a regular frequency within the period $[0, T]$.

The 12 spatio-temporal surveillance strategies are hence the combination of all possible spatial and temporal strategies. Finally, in order to estimate the marginal gain of an added point of surveillance in space or in time, we explore, for each of the 12 spatio-temporal strategies, all possible combinations of spatial surveillance SS $\in$ {Ran,Reg,ND,In,Out,Bet} over $k \in \{1,3,5,10,15,20\}$ nodes and temporal surveillance TS $\in$ {Ran,Reg} at $s \in \{1,3,5,10,15,20\}$ instants (resulting in a total of 432 different sampling schemes).

6.5.2. Performance evaluation of surveillance strategies

To compare the performance in terms of the capability of detection of the proposed strategies, we compute the frequency of detection of a given surveillance strategy across all simulations of a given simulation scheme (Holme, 2018). Recall that a single spatio-temporal surveillance strategy is defined by four parameters: the spatial strategy SS with its number of surveillance nodes k, together with the temporal strategy TS and its number of surveillance instants s. We hence define the frequency of detection (FD) of a given spatio-temporal surveillance strategy $\mathscr{S} = (\text{SS}, k, \text{TS}, s)$ as the proportion of positive detections over the $M = 10000$ simulations of a given simulation scheme, i.e.:

$$\text{FD}(\mathscr{S}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left[\left(\sum_{(i,t) \in G^{(m)}(\mathscr{S})} I_i^{(m)}(t) \right) > 0 \right], \tag{6.1}$$

where $I^{(m)}$ is the m-th realization of the process I giving the infectious nodes across time (see Section 6.3), and $G^{(m)}(\mathscr{S})$ is the m-th realization of the sampling scheme $\mathscr{S}$ (which consists of a random irregular space-time grid).

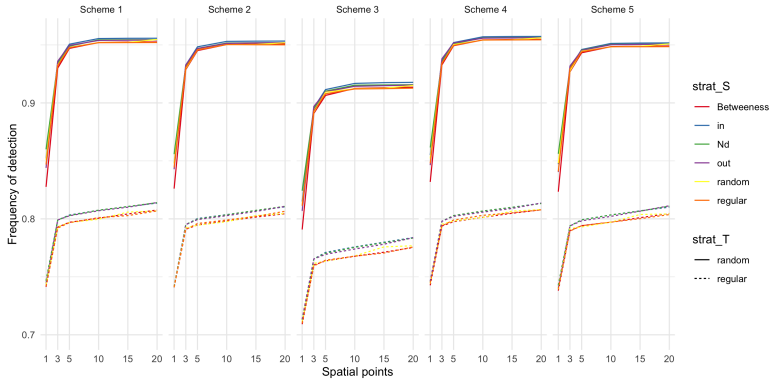
6.6. Results

The SIRS simulations have been used to test the performance of the proposed spatio-temporal surveillance strategies in terms of frequency of detection where, for each strategy, we recorded the number of positive detections over the 10000 simulations. An expected result is that by increasing the number of surveillance events, either in space or in time, we observe an increase in the frequency of detection, regardless of the simulation scheme used. This can be observed in the upper and lower panels of Figure 6.3: the upper panel shows how the frequency of detection increases with an increasing number of spatial surveillance points and similarly for increasing temporal points in the bottom panels. The colors of the lines represent the spatial strategy SS, while the line types the temporal strategy TS. Regardless of the simulation scheme, we can observe in the upper panel of Figure 6.3 that the random temporal strategy vastly outperforms the regular temporal strategy when the frequency of detection is averaged across all possible number of temporal points, while only smaller differences can be found between the spatial strategies, being the In and ND the ones achieving the higher frequency of detection. In the bottom panel of the same figure, where the frequency of detection is plotted against the number of temporal points and averaged across all possible number of spatial points, we can appreciate that the spatial strategies Out, In and ND outperforms the random, regular, and betweenness ones, regardless of the temporal strategy. By comparing the simulation schemes in both panels of Figure 6.3 we observe that the third scheme (the one corresponding to the tropospheric connectivity matrix averaged over the entire year) is the one achieving the lowest frequency of detection, while between the other scheme there is no significant difference.

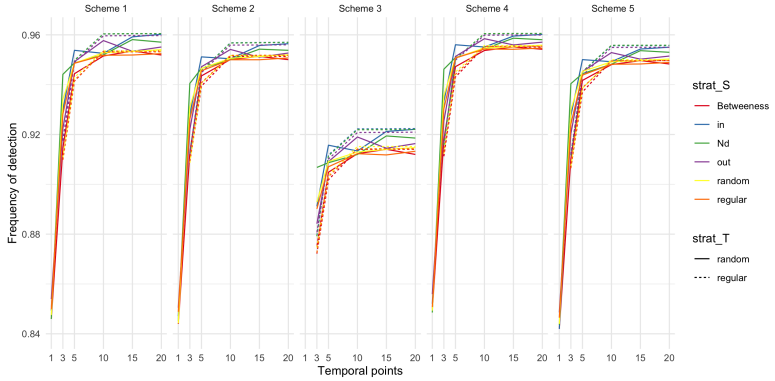
In order to determine the best-performing spatio-temporal strategies across all combinations of numbers of surveillance times and sites, we depicted in Figures 6.4a to 6.0e a 6×6 grid containing all $s \times k$ combinations. For each cell, there is a unique best spatio-temporal strategy in terms of frequency of detection that is represented using a coloring scheme for spatial strategies and a pattern scheme for the temporal strategy. Furthermore, we reported within each cell the frequency of detection associated with the best performing strategy. We can observe that, across all schemes, the random temporal strategy (cells with striped pattern) is preferred when the number of temporal points s and of spatial points k are equal to 1 or 3 (except a few exceptions). This means that the random temporal strategy is to be preferred when the available number of surveillance events and surveillance sites is very low. On the other hand, when $s > 3$ or $k > 3$, the regular time strategy is generally to be preferred. In terms of spatial strategy, when the number of surveillance sites and events is very low (between 1 and 3), the preferred spatial strategies are Out and ND, whereas if $s = 1$ and k is large, then the betweenness and regular spatial strategies should be preferred. When the number of surveillance events s is bigger than 3, then In and ND strategies become preferred (usually In for the lower values of k and ND for the higher values, but this depends on

the simulation scheme). Overall, there are no big differences across the simulation schemes, except for the third one, as before, in which the random spatial strategy is unusually well-performing for low values of s.

From the graphics in Figure 6.3 and the grids in Figures 6.4a to 6.0e, we can observe that the marginal gain in the frequency of detection per added unit of surveillance has a non-linear behavior, since small increases at the beginning imply a large gain in detection frequency, up to $k = 5$ and $s = 5$ (corresponding, respectively, to 1.6% of the total number of sites ($N = 297$) and to 3.3% of the 150 iterations of the simulations), whereas, after that threshold, further increases of the number of surveillance sites/events bring relatively smaller gains in detection.

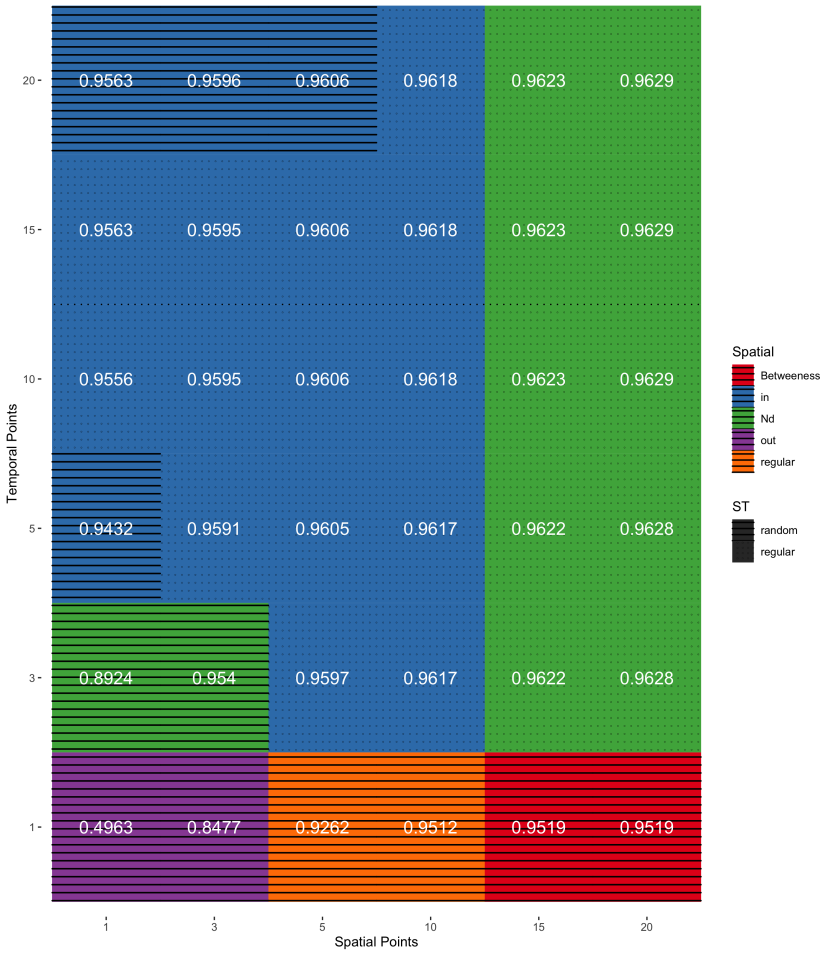


(a)

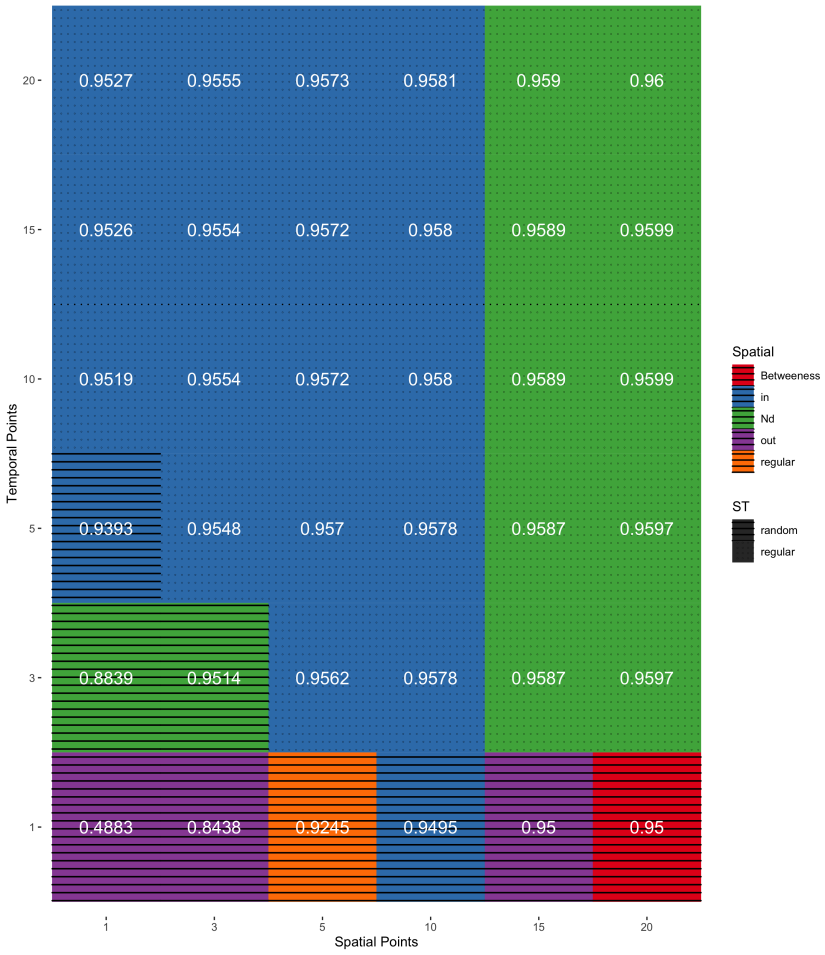


(b)

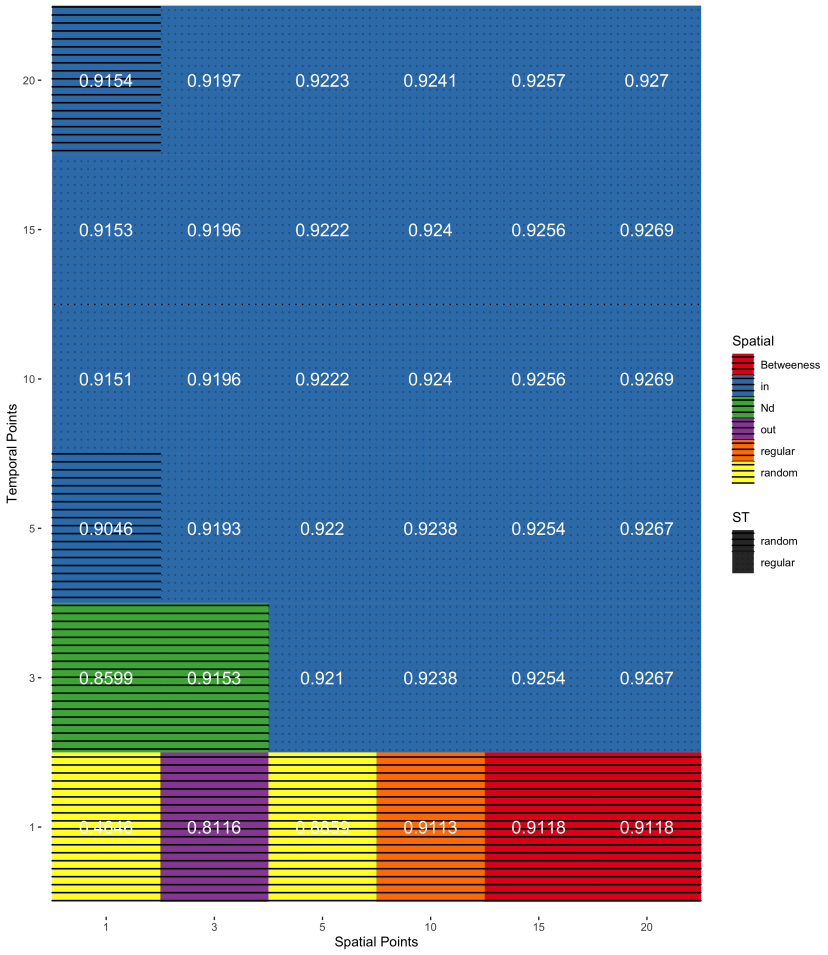
Figure 6.3. – The upper panel depicts the change in frequency of detection as a function of the number of surveillance sites, averaged over all possible numbers of surveillance events. The bottom panel is similar, but is a function of surveillance events and the frequency is averaged over all possible numbers of surveillance sites. The line color indicates the spatial surveillance strategy, while the line type indicates the temporal surveillance strategy.


0.9563
0.9596
0.9606
0.9618
0.9623
0.9629
0.9563
0.9595
0.9606
0.9618
0.9623
0.9629
0.9556
0.9595
0.9606
0.9618
0.9623
0.9629
0.9432
0.9591
0.9605
0.9617
0.9622
0.9628
0.8924
0.954
0.9597
0.9617
0.9622
0.9628
0.4963
0.8477
0.9262
0.9512
0.9519
0.9519
20
15
10
5
3
1
Temporal Points
1
3
5
10
15
20
Spatial Points
Spatial
Betweeness
in
Nd
out
regular
ST
random
regular

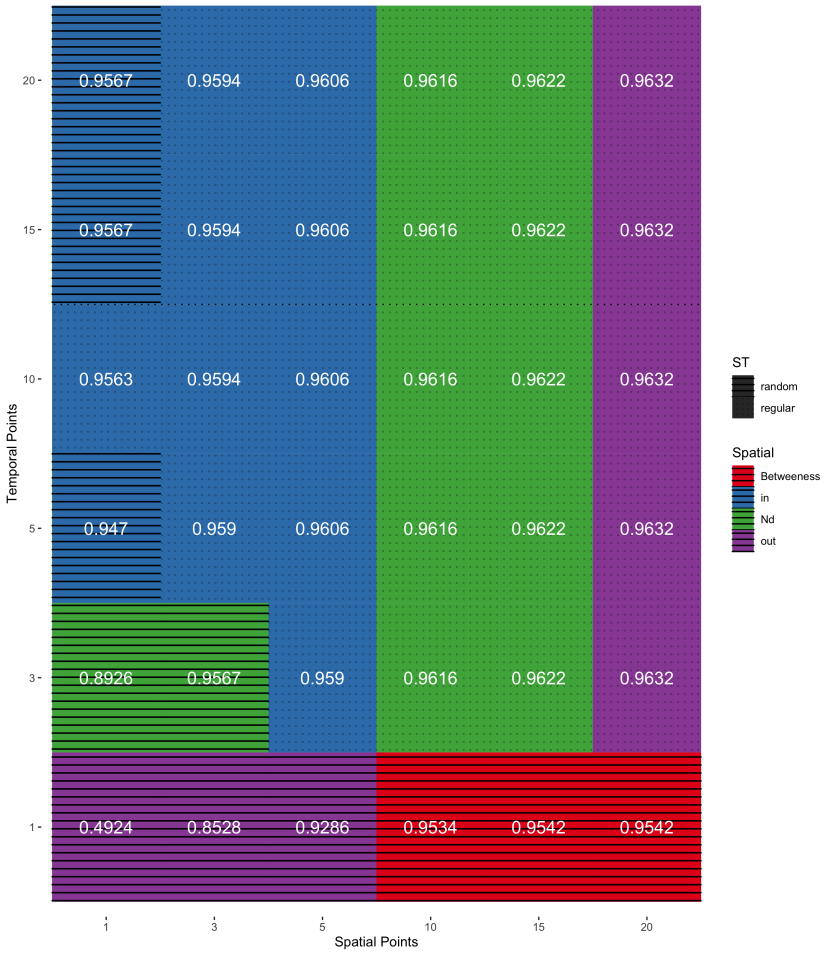
(a)


0.9527
0.9555
0.9573
0.9581
0.959
0.96
0.9526
0.9554
0.9572
0.958
0.9589
0.9599
0.9519
0.9554
0.9572
0.958
0.9589
0.9599
0.9393
0.9548
0.957
0.9578
0.9587
0.9597
0.8839
0.9514
0.9562
0.9578
0.9587
0.9597
0.4883
0.8438
0.9245
0.9495
0.95
0.95
20
15
10
5
3
1
Temporal Points
1
3
5
10
15
20
Spatial Points
Spatial
Betweeness
in
Nd
out
regular
ST
random
regular

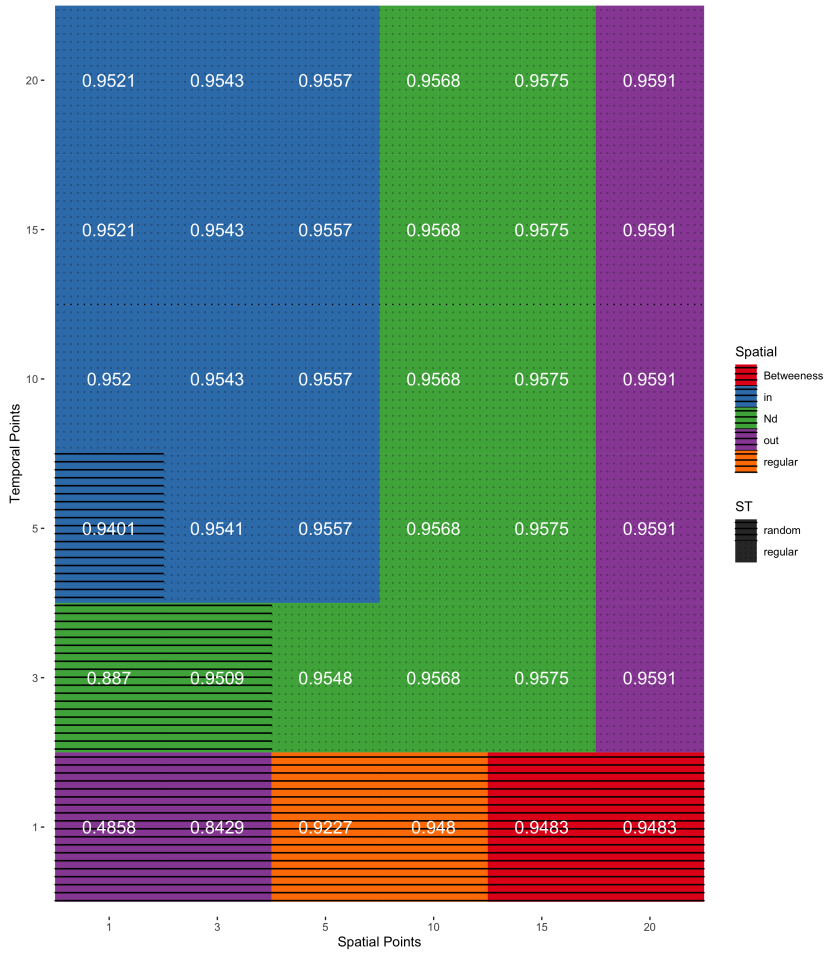
(b)


0.9154
0.9197
0.9223
0.9241
0.9257
0.927
0.9153
0.9196
0.9222
0.924
0.9256
0.9269
0.9151
0.9196
0.9222
0.924
0.9256
0.9269
0.9046
0.9193
0.922
0.9238
0.9254
0.9267
0.8599
0.9153
0.921
0.9238
0.9254
0.9267
0.8116
0.9113
0.9118
0.9118
20
15
10
5
3
1
Temporal Points
1
3
5
10
15
20
Spatial Points
Spatial
Betweeness
in
Nd
out
regular
random
ST
random
regular

(c)


0.9567
0.9594
0.9606
0.9616
0.9622
0.9632
0.9567
0.9594
0.9606
0.9616
0.9622
0.9632
0.9563
0.9594
0.9606
0.9616
0.9622
0.9632
0.947
0.959
0.9606
0.9616
0.9622
0.9632
0.8926
0.9567
0.959
0.9616
0.9622
0.9632
0.4924
0.8528
0.9286
0.9534
0.9542
0.9542
20
15
10
5
3
1
Temporal Points
1
3
5
10
15
20
Spatial Points
ST
random
regular
Spatial
Betweeness
in
Nd
out

(d)



(e)

Figure 6.0. – The 6×6 grid represents all possible combinations of s and k in $\{1,3,5,10,15,20\}$. Each cell of the grid gives the best performing spatio-temporal surveillance strategy in terms of frequency of detection (by coloring the cell according to the spatial strategy and using a different pattern for the temporal strategy) and the detection frequency attained by the best strategy (number in white between 0 and 1). From (a) to (e): Simulation schemes from 1 to 5, respectively, as specified in Section 6.4.2.

6.7. Discussion

In this paper, we define spatio-temporal surveillance strategies to predict the spread of an epidemic as early as possible by coupling spatial and temporal strategies. A spatio-temporal surveillance strategy is hence defined as a set of surveillance locations and dates that are deployed over a given territory in a given period of time. Spatial

strategies have been defined based on the constructed tropospheric contact networks and make use of centrality network metrics such as *normalised-degree*, *in-degree*, *out-degree*, *betweeness centrality* and two other generic strategies, namely the *random* and *regular* strategies. On the other hand, temporal strategies are based on generic strategies which consist on choosing *randomly* and *regularly* dates in a given period of time (Bai et al., 2017), whereas one could consider more specific temporal strategies informed by external data (for instance, by sampling nodes during favorable periods for infection —typically, after precipitation events for pathogens favored by wet deposition).

The proposed surveillance strategies are evaluated in terms of frequency of detection (Holme, 2018) based on simulated epidemics over the constructed contact networks. This method confirmed the usefulness of using strategies grounded on network centrality metrics, that generally show a higher frequency of detection than random or regular spatial strategies. Furthermore, we gained insight into the best strategy under the reasonable constraint of a limited number of surveillance sites and events per surveillance campaign. For those campaigns with a very limited number of events and sites, the spatial strategies based on Out degree and ND criteria seems to show better performance. For those campaigns with more surveillance dates and sites, the In-degree and ND strategies are preferable. In both scenarios, network-based strategies outperform purely spatial or random ones. Finally, the observed non-linear behavior of the frequency of detection for an increased unit of surveillance is an encouraging result, since it indicates that even with a reduced number of surveillance sites/events, if wisely placed in time and space, it is possible to achieve a high frequency of detection.

In this work, we presented a generic stochastic epidemic model of type SIRS conditional on air-mass-based contact networks. This work is developed bearing in mind the problem of detecting airborne plant diseases at an early stage of an epidemic by designing appropriate space and time surveillance strategies to know *where* and *when* it is more likely to intercept the epidemic. We applied the proposed approach to the PACA region, but the approach could be applied to study regions covering larger spatial scales, which may correspond to more structured networks and induce less redundant links. Such applications to larger scales (e.g. to the Mediterranean basin as in Chapter 5) could lead to more heterogeneous results between the competing strategies.

The generic approach that we proposed would deserve to be applied to real settings, e.g., applied to a specific pathogen by accounting for the pathogen life traits and by considering a relevant spatial scale and a relevant time period for its spread. This real application could be performed for a bacterial disease caused by *Pseudomonas syringae* (Morris et al., 2013). This bacteria is characterized by its capacity to lift in the air from the source location and to be transported by the air masses until sink locations (that can be far from the source) and then to spread locally. Since the life cycle of *P. syringae* is strongly correlated to the water cycle (Morris et al., 2008), it makes

sense to consider network nodes consisting of watersheds as we did for building the tropospheric network for PACA. Moreover, since *P. syringae* is preferentially carried by rain Monteil et al. (2014), the tropospheric network could be estimated by focusing on connectivity generated by trajectories arising during rainy periods.

The frequency of detection (given by Equation 6.1) that we used to measure the performance of the surveillance strategy is just one possible measure, while others may have been used (Holme, 2018; Martinetti and Soubeyrand, 2019) including measures better assessing the ability for early detection. For example, we could use the time of the first detection, the prevalence of the disease in the entire population at the time of the first detection, or the delay of detection on the surveillance sites/events with respect to a fixed threshold (e.g., the epidemics reaching more than 1% of the nodes). In certain situations, and depending on the context, the definition of strategy performance involves multiple criteria. For instance, finding the best surveillance strategy implies the optimization of a multi-objective function including budgetary constraints on the possible number of surveillance sites/events per sampling campaign. In this aim, we could, for instance, search a set of undominated sampling designs with the Pareto front method. Briefly, one has to find a compromise between the minimum number of surveillance points and the maximum frequency of detection (which are antagonistic objectives), and the Pareto front method consists of determining the sampling designs that are not dominated by any other sampling design

6.8. Conclusion notes

Take home messages

- Disease spread on spatio-temporal networks can be modeled using adapted SIRS-like epidemic models.
- For these spatio-temporal models, a surveillance strategy consists in defining when and where surveying the hosts of the pathogen.
- Spatial strategies based on centrality network metrics yield better performance.
- Sampling randomly in time seems to be more effective when the number of sampling sites/events is very small, while sampling regularly in time is more effective when the number of surveillance sites/events is relatively large.
- Increasing the number of surveillance sites or events increases the chances of disease detection. But the frequency increases non-linearly with the number of surveillance sites/events, reaching a plateau around 1%-3% of nodes/iterations (allowing a relatively large detection ability with few observation effort).

Perspectives

- Adapting the proposed methodology to a real epidemic caused by an airborne pathogen to produce a real proof-of-concept and to validate our preliminary results in a generic settings.
- Using other metrics of early-detection performance and possibly combining them for defining a multi-objective optimal strategy.
- Including budgetary constraints on the number and cost of surveillance units in order to design more relevant strategies from an epidemiological-economic perspective.

7. Conclusion and Perspectives

Table of Contents

7.1. Summary

The PhD thesis aims to conceive methods for estimating and characterizing spatio-temporal troposhperic networks allowing the improvement of airborne pathogen surveillance. First, in Chapter 3, we presented the approach allowing us to estimate the spatio-temporal and spatial networks that are generated by trajectory data (Choufany et al., 2019a). It can be resumed by the following: a spatial network is generated from a measure of connection generated by trajectories linking spatial units during a given time period, where the network nodes correspond to the spatial units and the network edges corresponds to the fluxes among the nodes. Network edges are weighted and directed according to the defined connectivities. If successive time period can be considered, then the successive networks form a spatio-temporal network. The measures of connection in this context can vary from an application to another. Therefore, we introduce different general definitions of measures relying on different bio-physical hypotheses such as the occurrence and the strength of the trajectory arriving at a defined node and previously going through another one, the length or the duration of the intersection between the trajectory and the node. The motivation of this approach has been expressed by applying it within the context of estimating tropospheric networks with epidemiological perspective. The so-called tropospheric network is defined as a spatio-temporal network where the trajectories are the air mass trajectories that are derived from `HYSPLIT`. To illustrate the proposed approach,

we applied it in two different contexts: the Mediterranean basin and the French region PACA. The two regions have different sizes, shapes and different topographic characteristics. After defining the spatial units and a temporal period ranging from 2011 to 2017, the spatio-temporal networks were estimated and their properties were investigated numerically. These networks are directed and weighted according to the contact-based estimator defined in Chapter 3. To portray the different networks, we computed several node and network metrics (Boccaletti et al., 2006; Lü et al., 2016) that allow the detection of seasonal patterns in terms of tropospheric connectivity. The developed framework remains generic and can be applied to other contexts where trajectory data can be used to characterize the connections between spatial units, such as human and animal movement data.

The overall developed approach is based on the estimation of the connectivity between couples of spatial units. However, when referring to an estimation, it is primordial to compute the efficiency / uncertainty of the estimators. Since the defined estimators depend on a spatio-temporal process (i.e., the movement of air masses) and its observation, their properties vary according to the choice of the spatio-temporal observation process, i.e., how the spatial and temporal dimensions are sampled. In this respect, we computed the accuracy of the connectivity estimator by theoretically comparing different sampling schemes in Chapter 4: a classical Monte Carlo sampling scheme, a random sampling scheme with dependencies (consisting in an irregular random grid) and a deterministic sampling scheme. The two stochastic schemes are unbiased, and the variance of the connectivity estimation with the classical Monte Carlo sampling scheme is lower than the one with the random grid. The deterministic scheme leads to a null variance, obviously, and we derived an upper bound of the estimation error.

The networks that we estimate with our approach may be quite complex and, therefore, their interpretation may be quite difficult. Thus, it is essential to searching new ways to characterize complex spatio-temporal networks. Thus, in Chapter 5, we consider the Mediterranean spatio-temporal connectivity network and we characterize it in a spatial and temporal way. We confirm the seasonal patterns detected in Chapter 3 by using an alternative analysis method: the Cut Distance introduced by (Liu et al., 2018b). According to this classification method, two networks are identified, which correspond to Summer and Winter season. Thereupon, we introduce a measure to characterize the connectivity by highlighting the distant points with strong connectivity by using Pareto Front measure. This measure helps to identify spatio-temporal pattern in the context of complex networks. In addition, we characterize the spatial pattern of each season and distinguish the differences among them. Furthermore, to characterise the nodes, we simulate a compartmental model of type SI (Susceptible – Infected) conditional on the network, network nodes viewed as strong spreaders and strong receptors were detected through the SI–persistence and the SI–frequency, respectively. Different spatial zones were detected via this method and they differed according to the seasons. Besides that, we looked for spatial pattern in a preliminary

way on the resulting seasonal networks. We identified that the Summer network is more likely to be a *small-world network* than the Winter one, i.e. the neighbors of any given node are likely to be neighbors of each other, and any given node is likely to be reached from every other node with a small number of steps.

Beyond the conception and the characterization of tropospheric networks, we were interested in the spread of epidemics over such types of networks and surveillance issues. In this aim, in Chapter 6, we developed a stochastic compartmental epidemic model of type *SIRS* conditional on the network (i.e., the network edges were used to spread the disease by taking into account the associated directed weights). Then, several spatio-temporal surveillance strategies were proposed and inspired by the centrality node metrics defined in the literature. This chapter led to some preliminary results that essentially concern the choice of an optimal number of spatio-temporal sampled points coupled with the spatio-temporal strategy. In space, we have deduced the importance of network centrality metrics as an effective tool to sample appropriate monitoring points. By contrast, in time, the proposed strategies resulted in approximately equal performances.

7.2. Preliminary applications

Numerically, the work presented in this PhD thesis necessitates to deal with big data bases and to conceive and apply a relatively complex computation workflow. To reduce the workload and the computational burden of this part, a web application `tropolink` is in the process of being deployed. This application is based on concepts and codes developed during the PhD thesis (codes were initially developed with `R` using the package `sf` (Pebesma, 2018) often used in spatial studies and the package `igraph` (Csardi and Nepusz, 2006) often used for graph representation).

`tropolink` will first take the form of a prototype web platform that allows its users to estimate the tropospheric connectivity between spatial units according to the developed framework in Chapter 3. Through `tropolink`, the user can make the extraction and computation of tropospheric connectivity in a practical and simplified way, without going into the computation complexity and the burden of managing massive data and using different softwares, including `HYSPLIT`.

The web application is presented as in Figure 7.1. To perform the calculations, the user should essentially provide 3 mandatory inputs: spatial arrival points/areas defined by geographical coordinates (longitude-latitude-altitude) and corresponding to the nodes of the network, a period of time during which the user wishes to compute the connectivity and the backward duration of the trajectory. Based on the inputs, `tropolink` will run `HYSPLIT` to extract the relevant trajectories and compute the tropospheric connectivities between every couple of spatial units through two stages:

1. Extraction of the tropospheric trajectories: an automatic procedure runs daily in the information system of the BioSP research unit to collect daily climatic global

data provided by the National Climatic Data Center of NOAA (National Oceanic and Atmospheric Administration). According to the inputs defined by the user through the web interface of `tropolink`, the relevant trajectories coordinates will be calculated based on the above-mentioned data using `HYSPLIT` also installed on the Biosp computation cluster (https://informatique-mia.inrae.fr/biosp-cluster/cluster).

2. Computation of the tropospheric connectivities: after extracting the coordinates of the tropospheric trajectories arriving at the specified spatial points during the specified period of time, the connectivities between spatial units are computed according to the mathematical framework presented in Chapter 3.

The output of this process is air connectivity networks where the nodes represent the input geographical points and the edges represent the probabilities of two sites being connected. The edges are represented by directed arrows that are colored according to their intensities. The networks are plotted over geographical maps and can be computed for different time patterns according to the choice of the user: daily, monthly, seasonal, and yearly patterns, etc. The user can make the choice of saving his/her work on the application, or extracting the adjacency matrices of the networks, or even extract the connectivity networks as a figure.

This project is under construction and will be deployed in the near future. It is led by Hervé Richard, Samuel , and Cindy E. Morris. Loic Houde implemented the automatic procedure for collecting climatic data from NOAA. Maria Choufany, Samuel Soubeyrand, Davide Martinetti, Rachid Senoussi, Christel Leyronas, and Cindy E. Morris developed the framework. Maria Choufany, Samuel Soubeyrand, Davide Martinetti developed initial R-codes for computing the connectivity. Makina Corpus https://makina-corpus.com/ and Hervé Richard make the development of the web application.

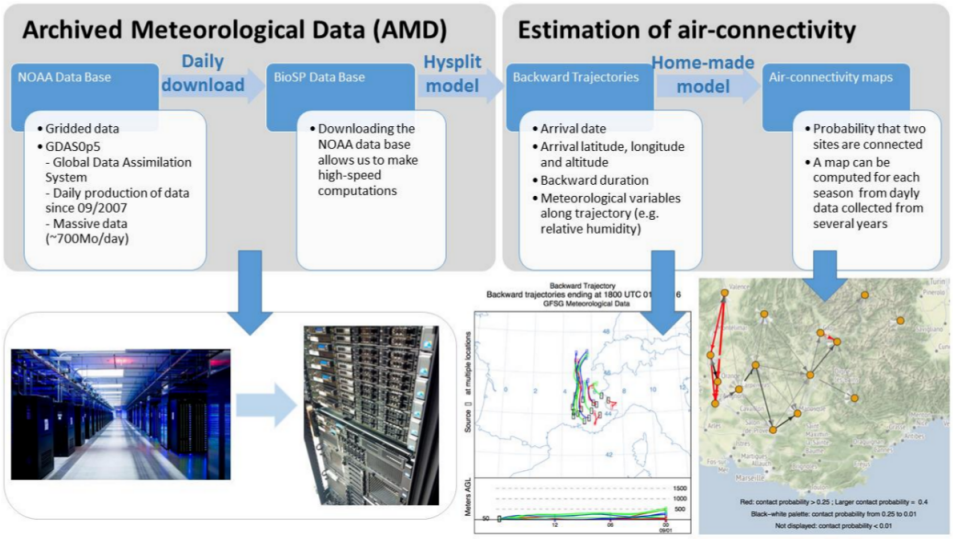


Figure 7.1. – Pipeline for computing air-connectivity maps from archived meteorological data. Figure prepared by Samuel Soubeyrand.

Figure 7.2. – Logo and short description of `tropolink`.

7.3. Perspectives

7.3.1. Linking tropospheric networks and genetic patterns

The work carried out in this thesis is, essentially, a statistical development for modeling networks generated by the movement of air masses, throughout which airborne pathogens spread and through which surveillance strategies could be improved (indeed, deepening the knowledge of connectivity and characterizing it at large spatial scales is a starting point to establish epidemic surveillance strategies with the objectives of monitoring the sanitary status of a host population or anticipating outbreaks of wind-dispersed pathogens). Some applications have been proposed to test and validate the approach but remain in a rather generic context, whereas different possibilities exist to embed the networks that we estimate in real contexts with real data.

For instance, from a demo-genetic perspective, transmission networks generally drive genetic patterns / networks (Banks et al., 2013; Garrett and Mundt, 1999). Thus,

developing an approach for evaluating the correlation between (i) contact networks obtained from air mass trajectory and (ii) large-scale genetic patterns of pathogen populations could be interesting. Such an approach is expected to give insights on the relatedness/discrepancies between the contact network and the genetic network. The contact network would be a good proxy of the transmission network if it is well *correlated* to the genetic network. In this case, the edges of the genetic network are weighted by the genetic similarity between pathogen populations observed at the nodes and the edges of the contact network are weighted by the probability of connections via the air-mass trajectories.

Within this context, I participated in a preliminary study given in Appendix C (published version: Leyronas et al., 2018), where we developed a method to potentially match the spatial genetic pattern of an airborne fungus, *S. sclerotiorum*, to a tropospheric connectivity network. *S. sclerotiorum*, a plant pathogenic fungus, can be dispersed via the air movement at short and long distances. Using a log-linear regression, Leyronas et al. (2018) determine the association between (i) the probability that the different components of the population sampled from the air in a given site are incoming from another specific site and (ii) the connectivity between sites via tropospheric movements:

$$\log q_{j\to i} = \alpha + \beta C_{j\to i} + \epsilon_{j\to i}$$

where $q_{j\to i}$ is, roughly speaking, the probability that the origin of isolates sampled in site i corresponds to site j ($q_{j\to i}$ is called the probability of the incoming component, PIC), $C_{j\to i}$ is the connectivity from site j to site i, α is the intercept coefficient, β is the slope coefficient and $\epsilon_{j\to i}$ is a random noise.

In four regions in France, air sampling was realized between 2014 and 2016. To infer airborne connectivity between the four sites, the framework presented in Chapter 3 was applied. Spatio-temporal tropospheric networks whose nodes are the four above-mentioned sites are estimated by extracting the air-mass trajectories arriving at the defined spatial areas between 2008-2017. The directed contact-based connectivity related to every couple of sites is estimated by considering the fraction of trajectories arriving at i and previously going throughout a buffer zone of 25 km around site j. In sum, monthly tropospheric networks are inferred for all the period of time. Each network has 4 nodes and 12 edges with varying intensities. The direction and intensities of the air mass connectivities vary according to the months associated with the seasonal variations. Figure 7.3 shows the resulting inferred tropospheric connectivity for March over the period 2008-2017. After conceiving the network, the least-squares approach is used to determine the parameters of the log-linear regression briefly introduced above. The significance of the link is assessed by a directional randomization test (testing the slope coefficient β). The regression lines between the probability of the incoming component $q_{j\to i}$ and the tropospheric connectivity $C_{j\to i}$ is stable across time as represented in Figure 7.4. It seems that there could be a positive link between

the probability and the tropospheric connectivity (but the reduced size of the network and the resulting size of the data set are small and the conclusions must be carefully drawn). It has to be noted also that the tropospheric connectivities better explain probabilities of incoming components than the simple geographic distances.

Beyond the application to *S. sclerotiorum* data, a simulation study is performed to evaluate the efficiency of the method for detecting an eventual link between a genetic pattern and a tropospheric network. In the simulation study, the number of sampling sites and the number of sampling isolates are varied, and the tropospheric connectivities are drawn from an exponential distribution. The simulation study shows a trade-off between the number of sites and the number of samples per site: Having more sites with few samples per site turns out to be generally better to determine the links between the sites than considering the reverse case; However, having too few samples per site can lead to inaccurate estimates of the PIC values.

This work remains a preliminary biological application of the methodology developed in this thesis. For carrying out such comparisons between tropospheric networks and genetic patterns, the genetic sampling should be extended as indicated by the simulation study (whereas the size of the tropospheric network is not really an issue, as illustrated with the networks estimated for PACA and the Mediterranean basin, where a large number of nodes were considered). Using relatively large observed genetic patterns and considering that the processes underlying the relationship between the long-distance dispersal of pathogens and the movement of air masses are not precisely known, it would be interesting to propose different competing tropospheric networks (e.g., using several options mentioned in Chapter 3) and to evaluate which one better matches the genetic pattern of an airborne infectious disease that is known for its capacity to spread in the region of interest. Indeed, dispersal of airborne pathogens may depend on diverse processes, and these processes vary according to the characteristics of the pathogens, e.g., (1) those transported at high altitude (Tanaka et al., 2019), (2) those whose concentration is affected by the nature of land cover (Smets et al., 2016), (3) those whose transmission is affected by the meteorological conditions such as temperature and humidity (Tang, 2009), (4) those linked to the water cycle (Morris et al., 2008). Hence, it would be interesting to build competing contact networks in accordance with these types of characteristics.

Concerning the comparison of large genetic patterns and tropospheric networks, one of the possible outlooks for the PACA region is the construction of the genetic network of *Pseudomonas syringae* (Morris et al., 2013). This airborne bacteria is present in the South-East region of France (Riffaud and Morris, 2002) (and in many regions in the world). It is characterized by its capacity of lifting up from its source location to the air and then being carried through the air masses to a sink location, following in some ways the water cycle. The fact that *P. syringae* can be advantageously deposited by rain (Monteil et al., 2014) incites to reconstruct its spatial genetic network by sampling rainwater. To this end, and in the framework of the ANR-funded project SPREE, a network for simultaneously collecting rainwater at different spatial points in PACA

is currently running. The collected samples will be used to test the hypothesis that microflora assemblages, deposited by rainfall along the same air trajectory, are similar, and that they are different from those deposited by rainfall along other trajectories. From these data, one will be able to reconstruction the genetic networks of *P. syringae* and have some direct indications about connectivity.

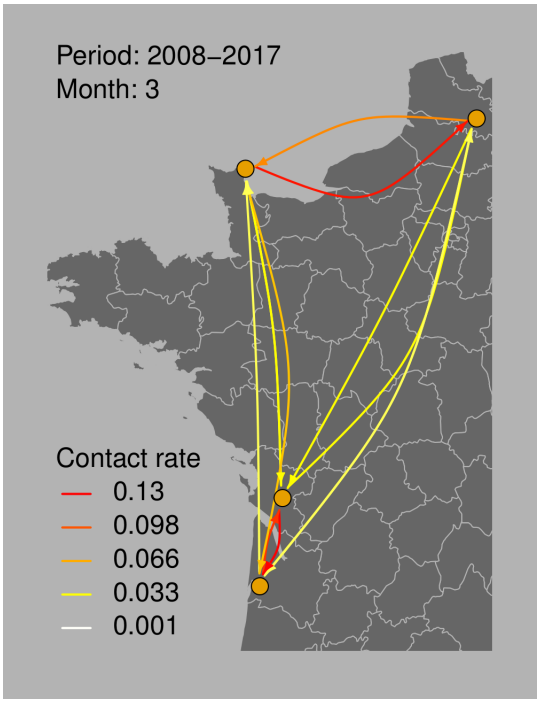


Figure 7.3. – Tropospheric connectivity inferred with `HYSPLIT` by Leyronas et al. (2018).

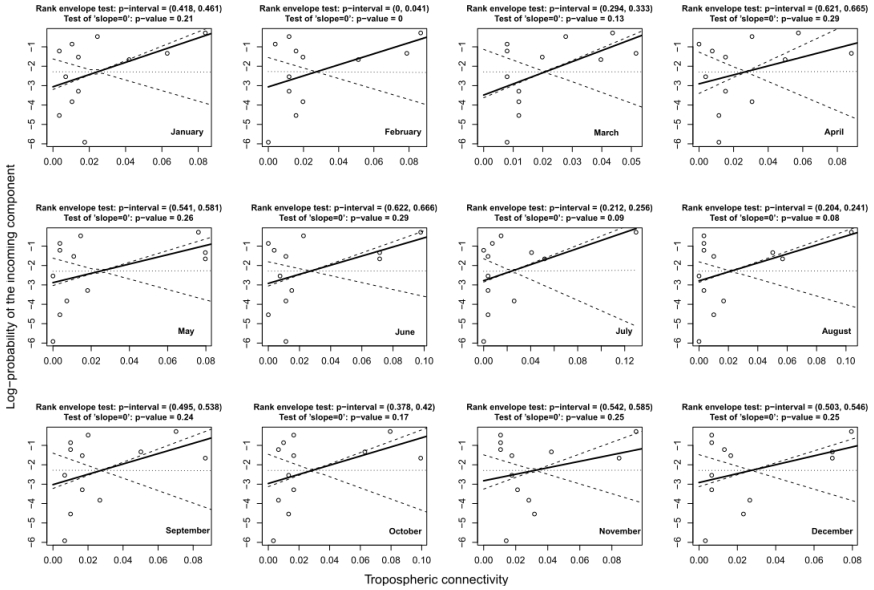


Figure 7.4. – Log-linear regression between the probability of incoming components and the tropospheric connectivity.

7.3.2. Beyond the contact-based connectivity

In Chapter 3, several types of connectivity were introduced: contact-based connectivity, length-based connectivity, duration-based connectivity, etc. We restricted our applications to the contact-based connectivity, from which we have inferred and characterized diverse tropospheric networks. As written in the previous section, in the epidemiological context, it is interesting to know, not only whether or not the air mass has passed through a given spatial unit, but also the time it has spent in the spatial unit, the length it has traveled in this spatial unit, whether it was raining or it was dry when it went through the spatial unit. The rationale behind these proposals are, for instance, that the length of the trajectory fragment (or the time spent by the trajectory) in an infectious zone may be viewed as a proxy of the number of pathogenic particles that can be transported towards other zones. In such a network, the edges are weighted by the average of the lengths (durations) of air mass trajectory fragments in distant spatial units.

7.3.3. Switching to other trajectory data

The approach developed in Chapter 3 is a general approach for estimating connectivity between spatial units over a defined period of time. These spatio-temporal connectivities depend on the application context. In the framework of the thesis, we

were interested in estimating connectivity via air mass trajectories between spatial units, so-called tropospheric connectivity. Our approach could be applied for inferring spatio-temporal connectivity in other scientific settings to explore other phenomena, such as:

- Estimating ocean-current contact network: where the nodes are spatial units over the ocean and the edges can be inferred by the flux of the ocean current between the spatial units. This kind of network can be informative when modeling the spatio-temporal dispersion of animals and particles transported by the current (Treml et al., 2008). The current trajectories can be inferred through the Lagrangian and Eulerian method as reviewed in Fossette et al. (2012). The Lagrangian measurement method consists of tracking objects such as buoys in order to record the water movements (Maximenko et al., 2009), while the Eulerian one consists of recording currents at a spatio-temporal point by using a current meter mooring (Klemas, 2012).
- Estimating animal-movement contact network: where the nodes are spatial units and the edges can be inferred according to the trajectories of animals. This kind of network could allow ones to characterize the connectivity between the landscape components (Bastille-Rousseau et al., 2018), and to estimate the transmission of epidemics affecting animals (Bigras-Poulin et al., 2007; Machado et al., 2019; VanderWaal et al., 2018). The trajectories of animals can be extracted via remote-sensing observation method (Handcock et al., 2009; Wark et al., 2007), the use of global positioning system (GPS; Calenge et al., 2009) or data on trade movements of farming animals (Qi et al., 2019).
- Estimating human-movement contact network: where the nodes are spatial units and the edges can be inferred according to the movements of humans. This kind of network can allow the tracking of infectious disease transmission over a large scale (Charu et al., 2017; Tizzoni et al., 2014) and the identification of super-spreading nodes such as the on-site eating and drinking locations in the case of COVID-19 transmission. The trajectories of the humans can be extracted via the mobile phone data (Hoteit et al., 2014), GPS data (Siła-Nowicka et al., 2016), or contact tracing enquiries (Eames et al., 2010).

7.3.4. Broadening the spatio-temporal sampling

To infer the connectivities, we used a sampling based on a grid (irregular in space but regular in time) in Chapter 3. To explore the relevancy of such a sampling choice, we characterized the error induced by the sampling design in Chapter 4, from an analytical and a numerical perspective, by considering, in particular, a classical Monte Carlo sampling scheme and a random sampling scheme with dependencies (corresponding to a random sampling grid). Beyond the obvious objective of selecting a sampling design that minimizes, for example, the quadratic error, one has to study

sampling designs that depend on practical constraints due to the application field and the phenomenon under study. In addition, it might be interesting to explore dynamic/sequential sampling approaches such as sampling schemes using Markov Chain techniques (Van Ravenzwaaij et al., 2018). These techniques take into account the previous spatio-temporal sampled locations but also the values obtained using the past samples. Thus, one can better explore the spatio-temporal domain with an expected lower computational cost compared to classical methods, for a specified target accuracy (Doucet et al., 2001).

7.3.5. Extension of network analysis

A broad (and partly open) question with network analysis is 'what metrics are appropriate to characterize a network?' This question can be addressed differently in each study and has no general answer. Many authors have adapted existing metrics such as centrality metrics (betweeness, degree, coreness, etc.), or proposed new metrics influenced by the specific application fields of the considered networks, e.g., social networks (Alahakoon et al., 2011; Chen et al., 2015; Galimberti et al., 2018; Kourtellis et al., 2013; Kumar and Panda, 2020), brain networks (Joyce et al., 2010; Zuo et al., 2012), landscape networks (Estrada and Bodin, 2008; Zamberletti et al., 2018), transportation networks (Wang et al., 2011; Zanin and Lillo, 2013) and biological networks (Koschützki and Schreiber, 2008; Ma and Zeng, 2003). In this thesis, we applied the centrality metrics over the inferred tropospheric networks in Chapter 3. Thereafter, in Chapter 5, we proposed a spatio-temporal metric, which consists in choosing the central spatial points in terms of strong spreaders by evaluating the Pareto Front over the edge weights and the geographical distance between the nodes. An interesting perspective for this method is to include the dynamic effect due to the temporal aspect. Instead of taking the average adjacency matrix over all days of the season, one can imagine running this method over all days selecting the points considered as strong spreaders for each day. Then, we could select a set of points by taking the most repeated ones across the season.

7.3.6. Estimation of epidemiological parameters

In Chapter 6, we developed a spatial SIRS model conditional on a tropospheric network, where the nodes are either susceptible, infectious or recovered, and transitions from one compartment to another are stochastic and depend on fixed parameters: transitions from the susceptible state to the infectious state are governed by parameter β, transitions from infectious to recovered are governed by δ_1, and transitions from recovered to susceptible are governed by δ_2.

In this model, parameter values were chosen without correspondence with a specific case study, but with the aim of generating a relatively large range of dynamics, from fast outbreaks throughout the whole network to rapidly vanishing outbreaks. To

handle a specific application and to improve model realism accordingly, parameters should ideally be estimated from epidemiological data. The estimation of parameters in epidemiological models can be performed with either a frequentist or a Bayesian approach, but is strongly dependent on available data as it is reviewed by Britton and Giardina (2014) and Andersson and Britton (2012). Classically, to inform the parameters of an epidemiological model, two types of data are considered: complete data and partially observed data.

Let us consider the ideal situation where the state $\{S(t), I(t), R(t)\}$ of any network node is observed along a given observation window. Then, the probability of any transition conditional on the past states of the whole network can be explicitly written (this probability takes into account the spatial and temporal dependencies). Subsequently, the likelihood can be expressed as a product of such probabilities. Hence, one can (numerically) maximize the likelihood or evaluate the posterior distribution of the parameters with an importance sampling algorithm or a Markov chain Monte Carlo algorithm (Albert et al., 2015).

For compartmental SIR-like models, one may encounter incomplete observations in very diverse settings. For instance, the state variables $\{S(t), I(t), R(t)\}$ are exhaustively observed in time, but only for a subset of individuals (or nodes in the context of networks); The state variables are observed for all the individuals, but only at some specific time points (hence the precise times of transitions are censored); Only one of the state variables is observed (e.g., if R is the removed / dead compartment, one observes the *first* time at which $R(t) = 1$, but we don't know if the individuals are susceptible or infectious when they are not dead. Many other situations (including a combination of the above-mentioned examples) can arise and be able to estimate the parameters in the wide diversity of possible situations is a challenge.

Before considering complex situations, let us highlight a particular case where the observations are incomplete but the inference is relatively straightforward. Roques et al. (2020) present a mechanistic-statistical approach where the outbreak is modeled with a SIRD (Susceptible - Infectious - Recovered - Dead) model based on a system of ordinary differential equations coupled with a model of the observation process, in which the observed variables are the daily increments in the number of deaths. Using the constraints embedded in the model and some simplifying assumptions, the authors are able to explicitly express the likelihood and perform maximum likelihood estimation.

Intrinsically, the reason why the estimation from partial observations presented above is straightforward is that the mechanistic component of the model is deterministic. When the mechanistic component of the model is stochastic and some state variables are unobserved (or latent, hidden, missing), the inference may be more complex. Indeed, in such a case, we are rarely able to write in a closed form the probabilities of transitions between observed states by taking into account all the temporal and/or spatial dependencies: in brief, analytical calculations are complicated (Britton and Giardina, 2014). A common practice is to explicitly introduce the hidden state

variables in estimation algorithms such as the EM (Becker, 1997), MCMC (McKinley et al., 2009; Roberts and Smith, 1994) and AMIS (Cornuet et al., 2012) algorithms.

For instance, the EM algorithm consists of maximizing the likelihood (of the observed variables) in repeating the following two steps until convergence: (1) Expectation: one computes the expected value of the likelihood of complete data (i.e., both the actually observed and the latent state variables), where the latent state variables are distributed with the current parameter values; (2) Maximization: the expected likelihood is maximized with respect to the parameters. Alternatively, one can choose to design an MCMC algorithm (in a Bayesian framework) for inferring parameters, in which one also explicitly handles the latent state variables and update, along the iterations, their values as well as the values of the parameters (see a simple toy example in Albert et al., 2015, Chap. 5). However, the MCMC approach may require a high computational memory or lead to chain mixing issues because of complex dependencies between the state variables and the parameters. If one encounters such difficulties, there are alternatives such as the recycling-based algorithms (e.g. AMIS), which can limit the number of times the likelihood is computed or which can more efficiently explore the space of unknown state variables and parameters (Abboud et al., 2019; Bugallo et al., 2015).

For more complex situations where the algorithms explicitly handling latent state variables do not converge or where the likelihood is theoretically or numerically intractable, approximate Bayesian computation (ABC; Marin et al., 2012) adapted to models of epidemiological dynamics or population dynamics (Britton and Giardina, 2014; Chong et al., 2018; McKinley et al., 2009; Soubeyrand et al., 2013) may be considered. ABC bypasses the evaluation of the likelihood function by instead (i) simulating a large number of times the model of interest for diverse parameter values typically drawn from their prior distributions, and (ii) evaluating the posterior distribution of the parameters using parameter values for which summary statistics computed from real data and summary statistics computed from simulation output match.

7.3.7. Refinement of the epidemio-surveillance methods

In Chapter 6, we proposed a combination of spatial and temporal strategies for testing the detection of an epidemic spreading on a spatio-temporal network of contacts. The spatial strategy is based on network centrality metrics, and the temporal strategy is based on strategies classically encountered in the literature. To improve this work, it would be relevant to define additional strategies: for instance, for spatial strategies on networks, it would be interesting to test the metrics Vote Rank and Random Generalized Accessibility as in Martinetti and Soubeyrand (2019). While for temporal strategies, it would be interesting to test the strategies according to more specific seasonal patterns, e.g. choosing the rainy or windy days as important surveillance events.

Once one has selected a sampling design for drawing spatio-temporal points form-

ing the sentinel surveillance scheme, one has to evaluate its performance in terms of disease detection. In Chapter 6, the performance criterion is restricted to the frequency of detection measure (Holme, 2018), but it could be extended by considering additional measures such as: (1) the time to detection or extinction (Holme, 2018), (2) early detection by 1% of persistence by computing the temporal lag between the surveillance set reaching 1% of infection and the entire network reaching the same threshold, (3) peak ratio that is computed through the difference between the epidemic persistence reached by the surveillance set and the entire network, (4) peak timing that can be calculated from the difference between the surveillance set reaching the epidemic peak and the entire network as defined in Martinetti and Soubeyrand (2019). These performance measures are more or less suitable according to the objectives of the surveillance and their relevancy can vary with respect to the severity of the epidemic, to the economic and sanitary risk posed by the disease, or to the degree of preparedness that society wants to achieve.

In this work, we evaluated the efficacy of the proposed spatio-temporal surveillance strategies with a single performance measure, i.e. the frequency of detection (FD), and by maximizing FD within a finite class of strategies. We could, however, consider an infinite class of strategies (e.g. strategies depending on continuous tuning parameters) and combine multiple surveillance efficacy metrics into a multi-objective optimization problem. In such a case, on has to rely to optimization methods to choose the best performing surveillance strategies, including stochastic optimization algorithms such as simulated annealing (Kirkpatrick et al., 1983) used for epidemiological surveillance by Mastin et al. (2019a), greedy algorithms (Church and ReVelle, 1974) used also for surveillance issues by Polgreen et al. (2009), and genetic algorithms (NSGA2; Deb et al., 2000; Ribaud et al., 2020).

Finally, even if adequate and efficient epidemiological surveillance can save millions in disease control and eradication costs by hampering the spread of epidemics thanks to early-stage interventions, it still involves a considerable effort for governmental agencies, farmers, private companies, other stakeholders, and the society in general. For further work in the area of surveillance throughout tropospheric networks, sampling strategies should be evaluated not only in terms of detection performance but also in terms of costs and possible savings by preventing the propagation of the disease. In this context, Picard (2018) used a Bayesian optimization method (Fang et al., 2005) for maximizing the mean of a net present value when the management strategy of the sharka disease varies. In the same vein, Schneider et al. (2020) developed an economic model to compute impact based on discounted foregone profits and losses in investment for the invasion of *Xylella fastidiosa* in European olive groves. Inspired by such work, we could extend our study by adding (i) a cost function for the sampling and (ii) a cost function for the non-detection of the disease (or the late detection of the disease), to find the optimal strategy making the balance between detection ability and economic cost.

Bibliography

C. Abboud, R. Senoussi, and S. Soubeyrand. Piecewise-deterministic markov processes for spatio-temporal population dynamics. Statistical Inference for Piecewise-deterministic Markov Processes, pages 209–255, 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, 79(2):765–789, 2019.

N. Abrego, B. Crosier, P. Somervuo, N. Ivanova, A. Abrahamyan, A. Abdi, K. Hämäläinen, K. Junninen, M. Maunula, J. Purhonen, et al. Fungal communities decline with urbanization—more in air than in soil. The ISME Journal, pages 1–10, 2020.

S. Aciego, C. Riebe, S. Hart, M. Blakowski, C. Carey, S. Aarons, N. Dove, J. Botthoff, K. Sims, and E. Aronson. Dust outpaces bedrock in nutrient supply to montane forest ecosystems. Nature Communications, 8:14800, 2017.

K. Aho, C. Weber, B. Christner, B. Vinatzer, C. Morris, R. Joyce, K. Failor, J. Werth, A. Bayless-Edwards, and D. Schmale III. Spatiotemporal patterns of microbial composition and diversity in precipitation. Ecological Monographs, 2019.

T. Alahakoon, R. Tripathi, N. Kourtellis, R. Simha, and A. Iamnitchi. K-path centrality: A new centrality measure in social networks. In Proceedings of the 4th workshop on social network systems, pages 1–6, 2011.

I. Albert, S. Ancelet, O. David, J.-B. Denis, D. Makowski, É. Parent, A. Rau, and S. Soubeyrand. Initiation à la statistique bayésienne-Bases théoriques et applications en alimentation, environnement, épidémiologie et génétique. 2015.

G. Alexanderson. About the cover: Euler and königsberg's bridges: A historical view. Bulletin of the american mathematical society, 43(4):567–573, 2006.

S. Ali, P. Gladieux, M. Leconte, A. Gautier, A. F. Justesen, M. S. Hovmøller, J. Enjalbert, and C. de Vallavieille-Pope. Origin, migration routes and worldwide population genetic structure of the wheat yellow rust pathogen puccinia striiformis f. sp. tritici. PLoS Pathog, 10(1):e1003903, 2014.

L. J. Allen and A. M. Burgin. Comparison of deterministic and stochastic sis and sir models in discrete time. Mathematical biosciences, 163(1):1–33, 2000.

L. J. Allen, F. Brauer, P. Van den Driessche, and J. Wu. Mathematical epidemiology, volume 1945. Springer, 2008.

A. Anastassopoulos, S. Nguyen, and X. Xu. On the use of the hysplit model to study air quality in windsor, ontario, canada. Environmental Informatics Archives, 2: 517–525, 2004.

P. K. Anderson, A. A. Cunningham, N. G. Patel, F. J. Morales, P. R. Epstein, and P. Daszak. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in ecology & evolution, 19(10):535–544, 2004.

R. M. Anderson. The population dynamics of infectious diseases: theory and applications. Springer, 2013.

H. Andersson and T. Britton. Stochastic epidemic models and their statistical analysis, volume 151. Springer Science & Business Media, 2012.

I. Arita, M. Nakane, K. Kojima, N. Yoshihara, T. Nakano, and A. El-Gohary. Role of a sentinel surveillance system in the context of global surveillance of infectious diseases. The Lancet Infectious Diseases, 4(3):171–177, 2004.

M. Armon, E. Dente, J. A. Smith, Y. Enzel, and E. Morin. Synoptic-scale control over modern rainfall and flood patterns in the levant drylands with implications for past climates. Journal of Hydrometeorology, 19(6):1077–1096, 2018.

K. Ashrafi, M. Shafiepour-Motlagh, A. Aslemand, and S. Ghader. Dust storm simulation over iran using hysplit. Journal of environmental health science and engineering, 12(1):9, 2014.

D. E. Aylor. The role of intermittent wind in the dispersal of fungal pathogens. Annual Review of Phytopathology, 28(1):73–92, 1990.

D. E. Aylor. Spread of plant disease on a continental scale: role of aerial dispersal of pathogens. Ecology, 84(8):1989–1997, 2003.

D. E. Aylor, G. S. Taylor, and G. S. Raynor. Long-range transport of tobacco blue mold spores. Agricultural Meteorology, 27(3-4):217–232, 1982.

J. Bae and S. Kim. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. Physica A: Statistical Mechanics and its Applications, 395:549–559, 2014.

Y. Bai, B. Yang, L. Lin, J. L. Herrera, Z. Du, and P. Holme. Optimizing sentinel surveillance in temporal network epidemiology. Scientific reports, 7(1):1–10, 2017.

N. T. Bailey et al. The mathematical theory of infectious diseases and its applications. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient mcmc. Statistics and Computing, 29(3):599–615, 2019.

D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences, 106(51):21484–21489, 2009.

S. C. Banks, G. J. Cary, A. L. Smith, I. D. Davies, D. A. Driscoll, A. M. Gill, D. B. Lindenmayer, and R. Peakall. How does ecological disturbance influence genetic diversity? Trends in Ecology & Evolution, 28(11):670–679, 2013.

S. Bansal, B. T. Grenfell, and L. A. Meyers. When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface, 4(16):879–891, 2007.

S. Bansal, J. Read, B. Pourbohloul, and L. A. Meyers. The dynamic nature of contact networks in infectious disease epidemiology. Journal of biological dynamics, 4(5): 478–489, 2010.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. Science, 286 (5439):509–512, 1999.

A.-L. Barabási and E. Bonabeau. Scale-free networks. Scientific American, 288(5): 60–69, 2003.

A.-L. Barabási et al. Network science. Cambridge University Press, 2016.

A. Barberán, J. Ladau, J. W. Leff, K. S. Pollard, H. L. Menninger, R. R. Dunn, and N. Fierer. Continental-scale distributions of dust-associated bacteria and fungi. Proceedings of the National Academy of Sciences, 112(18):5756–5761, 2015.

M. Barthelemy. Betweenness centrality in large complex networks. The European physical journal B, 38(2):163–168, 2004.

M. Barthélemy. Spatial networks. Physics Reports, 499(1-3):1–101, 2011.

M. Barthélemy. Spatial Networks. Springer, 2014.

P. Basaras, D. Katsaros, and L. Tassiulas. Detecting influential spreaders in complex, dynamic networks. Computer, 2013.

G. Bastille-Rousseau, I. Douglas-Hamilton, S. Blake, J. M. Northrup, and G. Wittemyer. Applying network theory to animal movements to identify properties of landscape space use. Ecological Applications, 28(3):854–864, 2018.

G. Beaunée, E. Vergu, and P. Ezanno. Modelling of paratuberculosis spread between dairy cattle farms at a regional scale. Veterinary research, 46(1):111, 2015.

N. Becker. The uses of epidemic models. Biometrics, pages 295–305, 1979.

N. G. Becker. Uses of the em algorithm in the analysis of data on hiv/aids and other infectious diseases. Statistical Methods in Medical Research, 6(1):24–37, 1997.

D. C. Bell, J. S. Atkinson, and J. W. Carlson. Centrality measures for disease transmission networks. Social networks, 21(1):1–21, 1999.

M. Bigras-Poulin, K. Barfod, S. Mortensen, and M. Greiner. Relationship of trade patterns of the danish swine industry animal movements network to potential disease spread. Preventive veterinary medicine, 80(2-3):143–165, 2007.

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. Physics Reports, 424(4-5):175–308, 2006.

P. Bogawski, K. Borycka, Ł. Grewling, and I. Kasprzyk. Detecting distant sources of airborne pollen for Poland: Integrating back-trajectory and dispersion modelling with a satellite-based phenology. Science of The Total Environment, 689:109–125, 2019.

P. Bonacich. Power and centrality: A family of measures. American journal of sociology, 92(5):1170–1182, 1987.

P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. Social networks, 23(3):191–201, 2001.

J. Bouttier, P. Di Francesco, and E. Guitter. Geodesic distance in planar graphs. Nuclear physics B, 663(3):535–567, 2003.

R. M. Bowers, S. McLetchie, R. Knight, and N. Fierer. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. The ISME Journal, 5(4):601–612, 2011a.

R. M. Bowers, A. P. Sullivan, E. K. Costello, J. L. Collett, R. Knight, and N. Fierer. Sources of bacteria in outdoor air across cities in the Midwestern United States. Applied and Environmental Microbiology, 77(18):6350–6356, 2011b. ISSN 0099-2240.

R. M. Bowers, N. Clements, J. B. Emerson, C. Wiedinmyer, M. P. Hannigan, and N. Fierer. Seasonal variability in bacterial and fungal diversity of the near-surface atmosphere. Environmental Science & Technology, 47(21):12097–12106, 2013a.

R. M. Bowers, N. Clements, J. B. Emerson, C. Wiedinmyer, M. P. Hannigan, and N. Fierer. Seasonal variability in bacterial and fungal diversity of the near-surface atmosphere. Environmental Science & Technology, 47(21):12097–12106, 2013b.

T. W. Bowyer, R. Kephart, P. W. Eslinger, J. I. Friese, H. S. Miley, and P. R. Saey. Maximum reasonable radioxenon releases from medical isotope production facilities and their effect on monitoring nuclear explosions. Journal of environmental radioactivity, 115:192–200, 2013.

U. Brandes. A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2):163–177, 2001.

F. Brauer. Compartmental models in epidemiology. In Mathematical epidemiology, pages 19–79. Springer, 2008.

T. S. Brett, J. M. Drake, and P. Rohani. Anticipating the emergence of infectious diseases. Journal of The Royal Society Interface, 14(132):20170115, 2017.

T. Britton. Stochastic epidemic models: a survey. Mathematical biosciences, 225(1): 24–35, 2010.

T. Britton and F. Giardina. Introduction to statistical inference for infectious diseases. arXiv preprint arXiv:1411.3138, 2014.

D. Brockmann and D. Helbing. The hidden geometry of complex, network-driven contagion phenomena. science, 342(6164):1337–1342, 2013.

E. L. Brodie, T. Z. DeSantis, J. P. M. Parker, I. X. Zubietta, Y. M. Piceno, and G. L. Andersen. Urban aerosols harbor diverse and dynamic bacterial populations. Proceedings of the National Academy of Sciences, 104(1):299–304, 2007. ISSN 0027-8424.

J. K. Brown and M. S. Hovmøller. Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. Science, 297(5581):537–541, 2002.

M. Bueso, J. Angulo, and F. Alonso. A state-space model approach to optimum spatial sampling design based on entropy. Environmental and Ecological Statistics, 5(1): 29–44, 1998.

M. Bueso, J. Angulo, J. Cruz-Sanjulian, and J. García-Aróstegui. Optimal spatial sampling design in a multivariate framework. Mathematical geology, 31(5):507–525, 1999.

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

A. G. Bunn, D. L. Urban, and T. H. Keitt. Landscape connectivity: a conservation application of graph theory. Journal of environmental management, 59(4):265–278, 2000.

R. E. Caflisch. Monte carlo and quasi-monte carlo methods. Acta Numerica, 7:1–49, 1998.

C. Calenge, S. Dray, and M. Royer-Carenzi. The concept of animals' trajectories from a data analysis perspective. Ecological informatics, 4(1):34–41, 2009.

J. Cáliz, X. Triadó-Margarit, L. Camarero, and E. O. Casamayor. A long-term survey unveils strong seasonal patterns in the airborne microbiome coupled to general and regional atmospheric circulations. Proceedings of the National Academy of Sciences, 115(48):12229–12234, 2018.

L. Candeloro, L. Savini, and A. Conte. A new weighted degree centrality measure: The application in an animal disease epidemic. PloS one, 11(11):e0165781, 2016.

V. Capasso and G. Serio. A generalization of the kermack-mckendrick deterministic epidemic model. Mathematical Biosciences, 42(1-2):43–61, 1978.

P. J. Carrington, J. Scott, and S. Wasserman. Models and methods in social network analysis, volume 28. Cambridge university press, 2005.

T. Chai, A. Crawford, B. Stunder, M. J. Pavolonis, R. Draxler, and A. Stein. Improving volcanic ash predictions with the hysplit dispersion model by assimilating modis satellite retrievals. Atmospheric Chemistry and Physics, 17(4):2865–2879, 2017.

V. Charu, S. Zeger, J. Gog, O. N. Bjørnstad, S. Kissler, L. Simonsen, B. T. Grenfell, and C. Viboud. Human mobility and the spatial transmission of influenza in the united states. PLoS computational biology, 13(2):e1005382, 2017.

D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou. Identifying influential nodes in complex networks. Physica a: Statistical mechanics and its applications, 391(4): 1777–1787, 2012.

M. Chen, T. Nguyen, and B. K. Szymanski. A new metric for quality of network community structure. arXiv preprint arXiv:1507.04308, 2015.

Y. Chen and Y. Luo. Analysis of paths and sources of moisture for the south China rainfall during the presummer rainy season of 1979–2014. Journal of Meteorological Research, 32(5):744–757, 2018.

K. C. Chong, B. C. Y. Zee, and M. H. Wang. Approximate bayesian algorithm to estimate the basic reproduction number in an influenza pandemic using arrival times of imported cases. Travel medicine and infectious disease, 23:80–86, 2018.

M. Choufany, D. Martinetti, R. Senoussi, C. E. Morris, and S. Soubeyrand. Spatiotemporal large-scale networks shaped by air mass movements. arXiv preprint arXiv:1911.07007, 2019a.

M. Choufany, D. Martinetti, R. Senoussi, C. E. Morris, and S. Soubeyrand. Spatiotemporal large-scale networks shaped by air mass movements. arXiv preprint arXiv:1911.07007, 2019b.

R. Church and C. ReVelle. The maximal covering location problem. In Papers of the Regional Science Association, volume 32, pages 101–118. Springer-Verlag, 1974.

V. Colizza and A. Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. Journal of theoretical biology, 251(3):450–467, 2008.

E. Colman, P. Holme, H. Sayama, and C. Gershenson. Efficient sentinel surveillance strategies for preventing epidemics on networks. PLoS computational biology, 15 (11), 2019.

L. M. Colon-Perez, M. Couret, W. Triplett, C. C. Price, and T. H. Mareci. Small worldness in dense and weighted connectomes. Frontiers in Physics, 4:14, 2016.

J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. Scandinavian Journal of Statistics, 39(4):798–812, 2012.

T. A. Crowl, T. O. Crist, R. R. Parmenter, G. Belovsky, and A. E. Lugo. The spread of invasive species and infectious disease as drivers of ecosystem change. Frontiers in Ecology and the Environment, 6(5):238–246, 2008.

G. Csardi and T. Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems:1695, 2006. URL http://igraph.org.

N. J. Cunniffe, F. F. Laranjeira, F. M. Neri, R. E. DeSimone, and C. A. Gilligan. Cost-effective control of plant disease when epidemiological knowledge is incomplete: modelling bahia bark scaling of citrus. PLoS Comput Biol, 10(8):e1003753, 2014.

N. J. Cunniffe, R. C. Cobb, R. K. Meentemeyer, D. M. Rizzo, and C. A. Gilligan. Modeling when, where, and how to manage a forest epidemic, motivated by sudden oak death in california. Proceedings of the National Academy of Sciences, 113(20):5640–5645, 2016.

D. J. Daley and J. Gani. Epidemic modelling: an introduction, volume 15. Cambridge University Press, 2001.

B. D. Dalziel, J. M. Morales, and J. M. Fryxell. Fitting probability distributions to animal movement trajectories: using artificial neural networks to link distance, resources, and memory. The American Naturalist, 172(2):248–258, 2008.

L. Danon, A. P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon. Networks and the epidemiology of infectious disease. Interdisciplinary perspectives on infectious diseases, 2011, 2011.

T. Das, M. H. D. Majumdar, R. T. Devi, and T. Rajesh. Climate change impacts on plant diseases. SAARC Journal of Agriculture, 14(2):200–209, 2017.

P. J. Davis and P. Rabinowitz. Methods of numerical integration. Courier Corporation, 2007.

G. F. De Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. da Fontoura Costa. Role of centrality for the identification of influential spreaders in complex networks. Physical Review E, 90(3):032812, 2014.

K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In International conference on parallel problem solving from nature, pages 849–858. Springer, 2000.

N. Deo. Graph theory with applications to engineering and computer science. Courier Dover Publications, 2017.

V. Després, J. A. Huffman, S. M. Burrows, C. Hoose, A. Safatov, G. Buryak, J. Fröhlich-Nowoisky, W. Elbert, M. Andreae, U. Pöschl, et al. Primary biological aerosol particles in the atmosphere: a review. Tellus B: Chemical and Physical Meteorology, 64(1): 15598, 2012.

N. Dianati. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. Phys. Rev. E, 93:012304, Jan 2016.

O. Diekmann and J. Heesterbeek. Mathematical epidemiology of infectious diseases. Model Building, Analysis, 1989.

E. W. Dijkstra et al. A note on two problems in connexion with graphs. Numerische mathematik, 1(1):269–271, 1959.

A. Doucet, N. De Freitas, and N. Gordon. An introduction to sequential monte carlo methods. In Sequential Monte Carlo methods in practice, pages 3–14. Springer, 2001.

R. Draxler, G. Hess, and R. Draxler. Description of the hysplit_4 modelling system, national oceanic & atmospheric administration technical memorandum erl arl. NOAA Tech. Mem. ERL ARL-224, 1997.

R. R. Draxler and G. Hess. An overview of the hysplit_4 modelling system for trajectories. Australian meteorological magazine, 47(4):295–308, 1998.

R. R. Draxler and G. D. Rolph. Evaluation of the transfer coefficient matrix (tcm) approach to model the atmospheric radionuclide air concentrations from fukushima. Journal of Geophysical Research: Atmospheres, 117(D5), 2012.

G. Dwyer and J. S. Elkinton. Host dispersal and the spatial spread of insect pathogens. Ecology, 76(4):1262–1275, 1995.

N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences, 106 (36):15274–15278, 2009.

K. Eames, S. Bansal, S. Frost, and S. Riley. Six challenges in measuring contact networks for use in modelling. Epidemics, 10:72–77, 2015.

K. T. Eames, C. Webb, K. Thomas, J. Smith, R. Salmon, and J. M. F. Temple. Assessing the role of contact tracing in a suspected h7n2 influenza a outbreak in humans in wales. BMC infectious diseases, 10(1):141, 2010.

Y. Elad and I. Pertot. Climate change impacts on plant pathogens and plant diseases. Journal of Crop Improvement, 28(1):99–139, 2014.

E. Estrada and Ö. Bodin. Using network centrality measures to manage landscape connectivity. Ecological Applications, 18(7):1810–1825, 2008.

L. Euler. Leonhard euler and the königsberg bridges. Scientific American, 189(1): 66–72, 1953.

M. Evans and T. Swartz. Approximating integrals via Monte Carlo and deterministic methods, volume 20. OUP Oxford, 2000.

C. Fahlgren, Å. Hagström, D. Nilsson, and U. L. Zweifel. Annual variations in the diversity, viability, and origin of airborne bacteria. Applied and Environmental Microbiology, 76(9):3015–3025, 2010.

K.-T. Fang, R. Li, and A. Sudjianto. Design and modeling for computer experiments. CRC press, 2005.

B. Fay, H. Glaab, I. Jacobsen, and R. Schrodin. Evaluation of eulerian and lagrangian atmospheric transport models at the deutscher wetterdienst using anatex surface tracer data. Atmospheric environment, 29(18):2485–2497, 1995.

L. Ferreira and D. B. Hitchcock. A comparison of hierarchical methods for clustering functional data. Communications in Statistics-Simulation and Computation, 38 (9):1925–1949, 2009.

B. D. Fitt, P. Walklate, H. McCartney, A. Bainbridge, N. Creighton, J. Hirst, M. Lacey, and B. Legg. A rain tower and wind tunnel for studying the dispersal of plant pathogens by rain and wind. Annals of Applied Biology, 109(3):661–671, 1986.

J. Flood. The importance of plant health to food security. Food Security, 2(3):215–231, 2010.

H. H. Flor. Current status of the gene-for-gene concept. Annual review of phytopathology, 9(1):275–296, 1971.

S. Fossette, N. F. Putman, K. J. Lohmann, R. Marsh, and G. C. Hays. A biologist's guide to assessing ocean currents: a review. Marine Ecology Progress Series, 457:285–301, 2012.

J. Freeman. Studies in the distribution of insects by aerial currents. The Journal of Animal Ecology, pages 128–154, 1945.

L. C. Freeman. Centrality in social networks conceptual clarification. Social networks, 1(3):215–239, 1978.

A. Frieze and R. Kannan. Quick approximation to matrices and applications. Combinatorica, 19(2):175–220, 1999.

J. Fröhlich-Nowoisky, D. A. Pickersgill, V. R. Després, and U. Pöschl. High diversity of fungi in air particulate matter. Proceedings of the National Academy of Sciences, 106(31):12814–12819, 2009. ISSN 0027-8424.

J. Fröhlich-Nowoisky, S. Burrows, Z. Xie, G. Engling, P. Solomon, M. Fraser, O. Mayol-Bracero, P. Artaxo, D. Begerow, R. Conrad, M. Andreae, V. Després, and U. Pöschl. Biogeography in the air: Fungal diversity over land and oceans. Biogeosciences, 9 (3):1125–1136, 2012.

E. Galimberti, A. Barrat, F. Bonchi, C. Cattuto, and F. Gullo. Mining (maximal) span-cores from temporal networks. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 107–116, 2018.

R. Gallotti and M. Barthelemy. The multilayer temporal network of public transport in great britain. Scientific data, 2(1):1–8, 2015.

A. Ganesh, L. Massoulié, and D. Towsley. The effect of network topology on the spread of epidemics. In Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies., volume 2, pages 1455–1466. IEEE, 2005.

Y. Gang, Z. Tao, W. Jie, F. Zhong-Qian, and W. Bing-Hong. Epidemic spread in weighted scale-free networks. Chinese Physics Letters, 22(2):510, 2005.

D. Garant, S. E. Forde, and A. P. Hendry. The multifarious effects of dispersal and gene flow on contemporary adaptation. Functional Ecology, pages 434–443, 2007.

K. Garrett and C. Mundt. Epidemiology in mixed host populations. Phytopathology, 89(11):984–990, 1999.

B. George and S. Shekhar. Time-aggregated graphs for modeling spatio-temporal networks. In Journal on Data Semantics XI, pages 191–212. Springer, 2008.

J. Geweke. Monte carlo simulation and numerical integration. Handbook of Computational Economics, 1:731–800, 1996.

P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. Management science, 35(11):1367–1392, 1989.

A. Grabowski and R. Kosiński. Epidemic spreading in a hierarchical social network. Physical Review E, 70(3):031908, 2004.

N. C. Grassly and C. Fraser. Mathematical models of infectious disease transmission. Nature Reviews Microbiology, 6(6):477–487, 2008.

A. G. Hallar, G. Chirokova, I. McCubbin, T. H. Painter, C. Wiedinmyer, and C. Dodson. Atmospheric bioaerosols transported via dust storms in the western united states. Geophysical Research Letters, 38(17), 2011.

R. K. Hamede, J. Bashford, H. McCallum, and M. Jones. Contact networks in a wild tasmanian devil (sarcophilus harrisii) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. Ecology letters, 12(11):1147–1157, 2009.

A. Hampe. Plants on the move: The role of seed dispersal and initial population establishment for climate-driven range expansions. Acta Oecologica, 37(6):666–673, 2011.

R. N. Handcock, D. L. Swain, G. J. Bishop-Hurley, K. P. Patison, T. Wark, P. Valencia, P. Corke, and C. J. O'Neill. Monitoring animal behaviour and environmental interactions using wireless sensor networks, gps collars and satellite remote sensing. Sensors, 9(5):3586–3603, 2009.

F. Harary and R. Z. Norman. Graph theory as a mathematical model in social science. Number 2. University of Michigan, Institute for Social Research Ann Arbor, 1953.

S. Hashimoto, Y. Murakami, K. Taniguchi, and M. Nagai. Detection of epidemics in their early stage through infectious disease surveillance. International journal of epidemiology, 29(5):905–910, 2000.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

J. C. Helton and F. J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering & System Safety, 81 (1):23–69, 2003.

M. Hernandez-Ceballos, J. Soares, H. García-Mozo, M. Sofiev, J. Bolivar, and C. Galán. Analysis of atmospheric dispersion of olive pollen in southern spain using silam and hysplit models. Aerobiologia, 30(3):239–255, 2014.

J. L. Herrera, R. Srinivasan, J. S. Brownstein, A. P. Galvani, and L. A. Meyers. Disease surveillance on complex social networks. PLoS computational biology, 12(7), 2016.

H. W. Hethcote. A thousand and one epidemic models. In Frontiers in mathematical biology, pages 504–515. Springer, 1994.

H. W. Hethcote. The mathematics of infectious diseases. SIAM review, 42(4):599–653, 2000.

D. L. Heymann, G. R. Rodier, et al. Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. The Lancet infectious diseases, 1(5): 345–353, 2001.

S. Hiraoka, M. Miyahara, K. Fujii, A. Machiyama, and W. Iwasaki. Seasonal analysis of microbial communities in precipitation in the greater Tokyo area Japan. Frontiers in Microbiology, 8:1506, 2017.

P. Holme. Temporal network structures controlling disease spreading. Physical Review E, 94(2):022305, 2016.

P. Holme. Objective measures for sentinel surveillance in network epidemiology. Physical Review E, 98(2):022313, 2018.

P. Holme and J. Saramäki. Temporal networks. Physics Reports, 519(3):97–125, 2012.

P. Holme and J. Saramäki. Temporal Network Theory. Springer International Publishing, 1 edition, 2019.

D. M. Hondula, L. Sitka, R. E. Davis, D. B. Knight, S. D. Gawtry, M. L. Deaton, T. R. Lee, C. P. Normile, and P. J. Stenger. A back-trajectory and air mass climatology for the Northern Shenandoah Valley, USA. International Journal of Climatology, 30(4): 569–581, 2010.

M. B. Hooten, C. K. Wikle, S. L. Sheriff, and J. W. Rushin. Optimal spatio-temporal hybrid sampling designs for ecological monitoring. Journal of Vegetation Science, 20(4):639–649, 2009.

M. B. Hooten, H. R. Scharf, T. J. Hefley, A. T. Pearse, and M. D. Weegman. Animal movement models for migratory individuals and groups. Methods in Ecology and Evolution, 9(7):1692–1705, 2018.

M. Hosseini and R. Kerachian. A bayesian maximum entropy-based methodology for optimal spatiotemporal design of groundwater monitoring networks. Environmental monitoring and assessment, 189(9):433, 2017.

S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle. Estimating human trajectories and hotspots through mobile phone data. Computer Networks, 64:296–307, 2014.

B. Hou, Y. Yao, and D. Liao. Identifying all-around nodes for spreading dynamics in complex networks. Physica A: Statistical Mechanics and its Applications, 391(15): 4012–4017, 2012.

D. L. Huestis, A. Dao, M. Diallo, Z. L. Sanogo, D. Samake, A. S. Yaro, Y. Ousman, Y.-M. Linton, A. Krishna, L. Veru, et al. Windborne long-distance migration of malaria mosquitoes in the sahel. Nature, 574(7778):404–408, 2019.

P. E. Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. Journal of applied ecology, 46(1):10–18, 2009.

A. Huppert and G. Katriel. Mathematical modelling and prediction in infectious disease epidemiology. Clinical microbiology and infection, 19(11):999–1005, 2013.

T. Hurst and C. Davis. Forecasting volcanic ash deposition using hysplit. Journal of Applied Volcanology, 6(1):5, 2017.

G. W. Imbens and T. Lancaster. Efficient estimation and stratified sampling. Journal of Econometrics, 74(2):289–318, 1996.

E. Innocente, S. Squizzato, F. Visin, C. Facca, G. Rampazzo, V. Bertolini, I. Gandolfi, A. Franzetti, R. Ambrosini, and G. Bestetti. Influence of seasonality, air mass origin and particulate matter chemical composition on airborne bacterial community structure in the Po Valley, Italy. Science of The Total Environment, 593-594:677 – 687, 2017.

M. J. Jeger, M. Pautasso, O. Holdenrieder, and M. W. Shaw. Modelling disease spread and control in networks: implications for plant sciences. New Phytologist, 174(2): 279–297, 2007.

M. J. Jeger, L. V. Madden, and F. Van den Bosch. The effect of transmission route on plant virus epidemic development and disease control. Journal of theoretical biology, 258(2):198–207, 2009.

G. Johnson et al. Peronospora hyoscyami de bary: taxonomic history, strains, and host range. Peronospora hyoscyami de Bary: taxonomic history, strains, and host range., pages 1–18, 1989.

R. A. Jones. Using epidemiological information to develop effective integrated virus disease management strategies. Virus Research, 100(1):5–30, 2004.

P. Jordano. What is long-distance dispersal? and a taxonomy of dispersal events. Journal of Ecology, 105(1):75–84, 2017.

F. Jourdain, A. M. Samy, A. Hamidi, A. Bouattour, B. Alten, C. Faraj, D. Roiz, D. Petrić, E. Pérez-Ramírez, E. Velo, et al. Towards harmonisation of entomological surveillance in the mediterranean area. PLoS neglected tropical diseases, 13(6):e0007314, 2019.

K. E. Joyce, P. J. Laurienti, J. H. Burdette, and S. Hayasaka. A new measure of centrality for brain networks. PloS one, 5(8):e12200, 2010.

B. Karrer and M. E. Newman. Competing epidemics on complex networks. Physical Review E, 84(3):036106, 2011.

J. Kean, C. Phillips, and M. McNeill. Surveillance for early detection: lottery or investment. Surveillance for Biosecurity: pre-border to pest management, KJ Froud, AI Popay, and SM Zydenbos (eds), pages 11–17, 2008.

M. J. Keeling. The effects of local spatial structure on epidemiological invasions. Proceedings of the Royal Society of London. Series B: Biological Sciences, 266 (1421):859–867, 1999.

M. J. Keeling and K. T. Eames. Networks and epidemic models. Journal of the Royal Society Interface, 2(4):295–307, 2005.

D. Kempe, J. Kleinberg, and A. Kumar. Connectivity and inference problems for temporal networks. Journal of Computer and System Sciences, 64(4):820–842, 2002.

W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. ii.—the problem of endemicity. Proceedings of the Royal Society of London. Series A, containing papers of a mathematical and physical character, 138 (834):55–83, 1932.

W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115(772):700–721, 1927.

Y. O. Khaniabadi, S. M. Daryanoosh, A. Amrane, R. Polosa, P. K. Hopke, G. Goudarzi, M. J. Mohammadi, P. Sicard, and H. Armin. Impact of Middle Eastern Dust storms on human health. Atmospheric Pollution Research, 8(4):606–613, 2017.

N. Kinoshita, K. Sueki, K. Sasa, J.-i. Kitagawa, S. Ikarashi, T. Nishimura, Y.-S. Wong, Y. Satou, K. Handa, T. Takahashi, et al. Assessment of individual radionuclide distributions from the fukushima nuclear accident covering central-east japan. Proceedings of the National Academy of Sciences, 108(49):19526–19529, 2011.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. science, 220(4598):671–680, 1983.

I. Z. Kiss, J. C. Miller, P. L. Simon, et al. Mathematics of epidemics on networks. Cham: Springer, 598, 2017.

M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. Identification of influential spreaders in complex networks. Nature physics, 6(11): 888, 2010.

V. Klemas. Remote sensing of coastal and ocean currents: An overview. Journal of Coastal Research, 28(3):576–586, 2012.

E. P. Klement, R. Mesiar, and E. Pap. Triangular norms. position paper i: basic analytical and algebraic properties. Fuzzy sets and systems, 143(1):5–26, 2004.

M. M. Kling and D. D. Ackerly. Global wind patterns and the vulnerability of wind-dispersed species to climate change. Nature Climate Change, 10(9):868–875, 2020.

A. S. Klovdahl. Social networks and the spread of infectious diseases: the aids example. Social science & medicine, 21(11):1203–1216, 1985.

M. D. Koenig and S. Battiston. From graph theory to models of economic networks. a tutorial. Networks, topology and dynamics, 613:23–63, 2009.

E. D. Kolaczyk and G. Csárdi. Statistical analysis of network data with R, volume 65. Springer, 2014.

D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene regulation and systems biology, 2:GRSB–S702, 2008.

N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi. Identifying high betweenness centrality nodes in large social networks. Social Network Analysis and Mining, 3(4):899–914, 2013.

A. Kremer, O. Ronce, J. J. Robledo-Arnuncio, F. Guillaume, G. Bohrer, R. Nathan, J. R. Bridle, R. Gomulkiewicz, E. K. Klein, K. Ritland, et al. Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecology letters, 15(4):378–392, 2012.

S. Kumar and B. Panda. Identifying influential nodes in social networks: Neighborhood coreness based voting approach. Physica A: Statistical Mechanics and its Applications, page 124215, 2020.

A. Lahrouz, L. Omari, and D. Kiouach. Global analysis of a deterministic and stochastic nonlinear sirs epidemic model. Nonlinear Analysis: Modelling and Control, 16(1): 59–76, 2011.

J. C. Lang, H. De Sterck, J. L. Kaiser, and J. C. Miller. Analytic models for sir disease spread on random spatial networks. Journal of Complex Networks, 6(6):948–970, 2018.

P. E. Lekone and B. F. Finkenstädt. Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. Biometrics, 62(4):1170–1177, 2006.

C. Leyronas, C. E. Morris, M. Choufany, and S. Soubeyrand. Assessing the aerial interconnectivity of distant reservoirs of sclerotinia sclerotiorum. Frontiers in Microbiology, 9:2257, 2018.

C. Li, L. Wang, S. Sun, and C. Xia. Identification of influential spreaders based on classified neighbors in real-world complex networks. Applied Mathematics and Computation, 320:512–523, 2018.

W. Li, Y. Lin, and Y. Liu. The structure of weighted small-world networks. Physica A: Statistical Mechanics and its Applications, 376:708–718, 2007.

B. Liu, Y. Ma, W. Gong, M. Zhang, and J. Yang. Study of continuous air pollution in winter over Wuhan based on ground-based and satellite observations. Atmospheric Pollution Research, 9(1):156–165, 2018a.

G. Liu, L. Wong, and H. N. Chua. Complex discovery from weighted ppi networks. Bioinformatics, 25(15):1891–1897, 2009.

J.-G. Liu, Z.-M. Ren, Q. Guo, and B.-H. Wang. Node importance ranking of complex networks. Acta Physica Sinica, 62(17):178901, 2013.

Q. Liu, Z. Dong, and E. Wang. Cut based method for comparing complex networks. Scientific reports, 8(1):1–11, 2018b.

T. Liu, M. E. Marlier, R. S. DeFries, D. M. Westervelt, K. R. Xia, A. M. Fiore, L. J. Mickley, D. H. Cusworth, and G. Milly. Seasonal impact of regional outdoor biomass burning on air pollution in three Indian cities: Delhi, Bengaluru, and Pune. Atmospheric Environment, 172:83–92, 2018c.

Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. Controllability of complex networks. Nature, 473(7346):167, 2011.

L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou. Vital nodes identification in complex networks. Physics Reports, 650:1–63, 2016.

H.-W. Ma and A.-P. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics, 19(11):1423–1430, 2003.

X. Ma and L. Gao. Biological network analysis: insights into structure and functions. Briefings in functional genomics, 11(6):434–442, 2012.

G. Machado, C. Vilalta, M. Recamonde-Mendoza, C. Corzo, M. Torremorell, A. Perez, and K. VanderWaal. Identifying outbreaks of porcine epidemic diarrhea virus through animal movements and spatial neighborhoods. Scientific reports, 9(1): 1–12, 2019.

A. G. Mahura, U. S. Korsholm, A. A. Baklanov, and A. Rasmussen. Elevated birch pollen episodes in Denmark: contributions from remote sources. Aerobiologia, 23(3):171, 2007.

L. Makra, T. Sánta, I. Matyasovszky, A. Damialis, K. Karatzas, K.-C. Bergmann, and D. Vokou. Airborne pollen in three European cities: Detection of atmospheric circulation pathways by applying three-dimensional clustering of backward trajectories. Journal of Geophysical Research: Atmospheres, 115(D24), 2010.

M. L. Margosian, K. A. Garrett, J. S. Hutchinson, and K. A. With. Connectivity of the American agricultural landscape: assessing the national risk of crop pest and disease spread. BioScience, 59(2):141–151, 2009.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. Statistics and Computing, 22(6):1167–1180, 2012.

D. Martinetti and S. Soubeyrand. Identifying lookouts for epidemio-surveillance: Application to the emergence of xylella fastidiosa in france. Phytopathology, pages PHYTO–07, 2019.

A. J. Mastin, F. van den Bosch, T. R. Gottwald, V. A. Chavez, and S. R. Parnell. A method of determining where to target surveillance efforts in heterogeneous epidemiological systems. PLoS computational biology, 13(8):e1005712, 2017.

A. J. Mastin, T. R. Gottwald, F. van den Bosch, N. J. Cunniffe, and S. R. Parnell. Optimising risk-based surveillance for early detection of invasive plant pathogens. bioRxiv, page 834705, 2019a.

A. J. Mastin, F. van den Bosch, F. van den Berg, and S. R. Parnell. Quantifying the hidden costs of imperfect detection for early detection surveillance. Philosophical Transactions of the Royal Society B, 374(1776):20180261, 2019b.

N. Maximenko, P. Niiler, L. Centurioni, M.-H. Rio, O. Melnichenko, D. Chambers, V. Zlotnicki, and B. Galperin. Mean dynamic topography of the ocean derived from satellite and drifting buoy data using three different techniques. Journal of Atmospheric and Oceanic Technology, 26(9):1910–1919, 2009.

J. D. Mayer. Geography, ecology and emerging infectious diseases. Social science & medicine, 50(7-8):937–952, 2000.

T. McKinley, A. R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. The International Journal of Biostatistics, 5(1), 2009.

D. S. Meredith et al. Banana leaf spot disease (sigatoka) caused by mycosphaerella musicola leach. Phytopathological papers, (11), 1970.

L. A. Meyers, M. Newman, M. Martin, and S. Schrag. Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. Emerging infectious diseases, 9(2):204, 2003.

J. C. Miller and I. Z. Kiss. Epidemic spread in networks: Existing methods and current challenges. Mathematical modelling of natural phenomena, 9(2):4–42, 2014.

B. K. Mishra and N. Jha. Seiqrs model for the transmission of malicious objects in computer network. Applied Mathematical Modelling, 34(3):710–715, 2010.

C. L. Monteil, M. Bardin, and C. E. Morris. Features of air masses associated with the deposition of pseudomonas syringae and botrytis cinerea by rain and snowfall. The ISME journal, 8(11):2290–2304, 2014.

Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. The European Physical Journal B-Condensed Matter and Complex Systems, 26(4):521–529, 2002.

B. E. Moroz, H. L. Beck, A. Bouville, and S. L. Simon. Predictions of dispersion and deposition of fallout from nuclear testing using the NOAA-HYSPLIT meteorological model. Health Physics, 99(2), 2010.

C. E. Morris, D. C. Sands, B. A. Vinatzer, C. Glaux, C. Guilbaud, A. Buffiere, S. Yan, H. Dominguez, and B. M. Thompson. The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. The ISME Journal, 2(3):321, 2008.

C. E. Morris, D. Sands, M. Bardin, R. Jaenicke, B. Vogel, C. Leyronas, P. Ariya, and R. Psenner. Microbiology and atmospheric processes: research challenges concerning the impact of airborne micro-organisms on the atmosphere and climate. Biogeosciences, 8(1):17–25, 2011.

C. E. Morris, C. L. Monteil, and O. Berge. The life history of pseudomonas syringae: linking agriculture to earth system processes. Annual review of phytopathology, 51: 85–104, 2013.

C. E. Morris, S. Soubeyrand, E. K. Bigg, J. M. Creamean, and D. C. Sands. Mapping rainfall feedback to reveal the potential sensitivity of precipitation to biological aerosols. Bulletin of the American Meteorological Society, 98(6):1109–1118, 2017.

M. Moslonka-Lefebvre, A. Finley, I. Dorigatti, K. Dehnen-Schmutz, T. Harwood, M. J. Jeger, X. Xu, O. Holdenrieder, and M. Pautasso. Networks in plant epidemiology: from genes to landscapes, countries, and continents. Phytopathology, 101(4):392–403, 2011.

C. C. Mundt, K. E. Sackett, L. D. Wallace, C. Cowger, and J. P. Dudley. Aerial dispersal and multiple-scale spread of epidemic disease. EcoHealth, 6(4):546–552, 2009.

J. Murray. Mathematical biology ii. spatial models and biological applications. Springer-Verlag, New York, 2003.

M. E. Newman. Models of the small world. Journal of Statistical Physics, 101(3-4): 819–841, 2000.

M. E. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical Review E, 64(1):016132, 2001.

M. E. Newman. Spread of epidemic disease on networks. Physical Review E, 66(1): 016128, 2002.

M. E. Newman. The structure and function of complex networks. SIAM Review, 45(2): 167–256, 2003.

M. E. Newman. Analysis of weighted networks. Physical Review E, 70(5):056131, 2004.

M. Nicolaisen, J. S. West, R. Sapkota, G. G. Canning, C. Schoen, and A. F. Justesen. Fungal communities including plant pathogens in near surface air are similar across Northwestern Europe. Frontiers in Microbiology, 8:1729, 2017a.

M. Nicolaisen, J. S. West, R. Sapkota, G. G. M. Canning, C. Schoen, and A. F. Justesen. Fungal communities including plant pathogens in near surface air are similar across Northwestern Europe. Frontiers in Microbiology, 8:1729, 2017b.

H. Niederreiter. Quasi-monte carlo methods for multidimensional numerical integration. In Numerical Integration III, pages 157–171. Springer, 1988.

T. Opsahl and P. Panzarasa. Clustering in weighted networks. Social Networks, 31(2): 155–163, 2009.

L. Pace, L. Boccacci, M. Casilli, and S. Fattorini. Temporal variations in the diversity of airborne fungal spores in a Mediterranean high altitude site. Atmospheric Environment, 210:166 – 170, 2019. ISSN 1352-2310.

S. Pady and L. Kapica. Fungi in air over the atlantic ocean. Mycologia, 47(1):34–50, 1955.

S. Parnell, T. Gottwald, N. Cunniffe, V. Alonso Chavez, and F. van Den Bosch. Early detection surveillance for an emerging plant pathogen: a rule of thumb to predict prevalence at first discovery. Proceedings of the Royal Society B: Biological Sciences, 282(1814):20151478, 2015.

R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. Reviews of Modern Physics, 87(3):925, 2015.

M. Pautasso and M. J. Jeger. Network epidemiology and plant trade networks. AoB Plants, 6, 2014.

M. Pautasso, X. Xu, M. J. Jeger, T. D. Harwood, M. Moslonka-Lefebvre, and L. Pellis. Disease spread in small-size directed trade networks: the role of hierarchical categories. Journal of Applied Ecology, 47(6):1300–1309, 2010.

E. Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, 10(1):439–446, 2018.

I. A. Pérez, F. Artuso, M. Mahmud, U. Kulshrestha, M. L. Sánchez, and M. García. Applications of air mass trajectories. Advances in Meteorology, 2015, 2015.

C. Picard. Modélisation et optimisation de la gestion d'une épidémie: quel impact du paysage? PhD thesis, 2018.

P. M. Polgreen, Z. Chen, A. M. Segre, M. L. Harris, M. A. Pentella, and G. Rushton. Optimizing influenza sentinel surveillance at the state level. American journal of epidemiology, 170(10):1300–1306, 2009.

S. Pourshahabi, N. Talebbeydokhti, G. Rakhshandehroo, and M. R. Nikoo. Spatiotemporal multi-criteria optimization of reservoir water quality monitoring network using value of information and transinformation entropy. Water Resources Management, 32(10):3489–3504, 2018.

C. W. Powers, R. Hanlon, H. Grothe, A. J. Prussin, L. C. Marr, and D. G. Schmale. Coordinated sampling of microorganisms over freshwater and saltwater environments using an unmanned surface vehicle (USV) and a small unmanned aircraft system (sUAS). Frontiers in Microbiology, 9:1668, 2018.

A. Pozzebon, G. Loeb, and C. Duso. Grape powdery mildew as a food source for generalist predatory mites occurring in vineyards: effects on life-history traits. Annals of Applied Biology, 155(1):81–89, 2009.

L. Qi, G. Beaunée, S. Arnoux, B. L. Dutta, A. Joly, E. Vergu, and P. Ezanno. Neighbourhood contacts and trade movements drive the regional spread of bovine viral diarrhoea virus (bvdv). Veterinary research, 50(1):1–15, 2019.

J. L. Rabinowitz, A. R. Lupo, P. S. Market, and P. E. Guinan. An investigation of atmospheric rivers impacting heavy rainfall events in the North-Central Mississippi River Valley. International Journal of Climatology, 2019.

T. Rana, S. Khan, and M. Rahimi. Spatio-temporal optimisation of agricultural drainage using groundwater models and genetic algorithms: an example from the murray irrigation area, australia. Hydrogeology journal, 16(6):1145–1157, 2008.

S. Rautureau, B. Dufour, and B. Durand. Vulnerability of animal trade networks to the spread of infectious diseases: a methodological approach applied to evaluation and emergency control strategies in cattle, france, 2005. Transboundary and emerging diseases, 58(2):110–120, 2011.

M. Ribaud, C. Blanchet-Scalliet, C. Helbert, and F. Gillot. Robust optimization: a kriging-based multi-objective optimization approach. Reliability Engineering & System Safety, page 106913, 2020.

C.-H. Riffaud and C. E. Morris. Detection of pseudomonas syringae pv. aptata in irrigation water retention basins by immunofluorescence colony-staining. European Journal of Plant Pathology, 108(6):539–545, 2002.

D. Rigling and S. Prospero. Cryphonectria parasitica, the causal agent of chestnut blight: invasion history, population biology and disease control. Molecular Plant Pathology, 19(1):7–20, 2018.

G. O. Roberts and A. F. Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. Stochastic processes and their applications, 49(2):207–216, 1994.

M. Roberts, V. Andreasen, A. Lloyd, and L. Pellis. Nine challenges for deterministic epidemic models. Epidemics, 10:49–53, 2015.

L. E. Rocha, A. Decuyper, and V. D. Blondel. Epidemics on a stochastic model of temporal network. In Dynamics On and Of Complex Networks, Volume 2, pages 301–314. Springer, 2013.

A. W. Roddam. Mathematical epidemiology of infectious diseases: Model building, analysis and interpretation: O diekmann and jap heesterbeek, 2000, chichester: John wiley pp. 303,£ 39.95. isbn 0-471-49241-8, 2001.

G. Rolph, F. Ngan, and R. Draxler. Modeling the fallout from stabilized nuclear clouds using the HYSPLIT atmospheric dispersion model. Journal of Environmental Radioactivity, 136:41–55, 2014.

G. D. Rolph, R. R. Draxler, A. F. Stein, A. Taylor, M. G. Ruminski, S. Kondragunta, J. Zeng, H.-C. Huang, G. Manikin, J. T. McQueen, et al. Description and verification of the noaa smoke forecasting system: the 2007 fire season. Weather and Forecasting, 24 (2):361–378, 2009.

L. Roques, E. K. Klein, J. Papaix, A. Sar, and S. Soubeyrand. Using early data to estimate the actual infection fatality ratio from covid-19 in france. Biology, 9(5):97, 2020.

R. Ross. The mathematics of malaria. British Medical Journal, 1(2626):1023, 1911.

H. Sachs, M. Stiebitz, and R. J. Wilson. An historical note: Euler's königsberg letters. Journal of Graph Theory, 12(1):133–139, 1988.

M. Sadyś, C. A. Skjøth, and R. Kennedy. Back-trajectories show export of airborne fungal spores (Ganoderma sp.) from forests to agricultural and urban areas in England. Atmospheric Environment, 84:88–99, 2014.

T. Šantl-Temkiv, B. Sikoparija, T. Maki, F. Carotenuto, P. Amato, M. Yao, C. E. Morris, R. Schnell, R. Jaenicke, C. Pöhlker, et al. Bioaerosol field measurements: Challenges and perspectives in outdoor studies. Aerosol Science and Technology, 54(5):520–546, 2020.

R. Sarda-Esteve, D. Baisnee, B. Guinot, J. Sodeau, D. O'Connor, J. Belmonte, J.-P. Besancenot, J.-E. Petit, M. Thibaudon, G. Oliver, C. Sindt, and V. Gros. Variability and geographical origin of five years airborne fungal spore concentrations measured at Saclay, France from 2014 to 2018. Remote Sensing, 11(14), 2019.

I. Šauliene and L. Veriankaite. Application of backward air mass trajectory analysis in evaluating airborne pollen dispersion. Journal of Environmental Engineering and Landscape Management, 14(3):113–120, 2006.

S. Saura, Ö. Bodin, and M.-J. Fortin. Stepping stones are crucial for species' long-distance dispersal and range expansion through habitat networks. Journal of Applied Ecology, 51(1):171–182, 2014.

D. G. Schmale and S. D. Ross. Highways in the sky: Scales of atmospheric transport of plant pathogens. Annual Review of Phytopathology, 53(1):591–611, 2015.

D. G. Schmale III and S. D. Ross. Highways in the sky: Scales of atmospheric transport of plant pathogens. Annual review of phytopathology, 53, 2015.

K. Schneider, W. Van der Werf, M. Cendoya, M. Mourits, J. A. Navas-Cortés, A. Vicent, and A. O. Lansink. Impact of xylella fastidiosa subspecies pauca in european olives. Proceedings of the National Academy of Sciences, 117(17):9250–9259, 2020.

J. Scott. Social network analysis. Sociology, 22(1):109–127, 1988.

L. Scrucca. On some extensions to GA package: hybrid optimisation, parallelisation and islands evolution. The R Journal, 9(1):187–206, 2017.

H. Seherm and S. M. Coakley. Plant pathogens in a changing world. Australasian Plant Pathology, 32(2):157–165, 2003.

S. B. Seidman. Network structure and minimum degree. Social networks, 5(3):269–287, 1983.

M. Á. Serrano and M. Boguna. Clustering in complex networks. i. general formalism. Physical Review E, 74(5):056114, 2006.

P. M. Severns, K. E. Sackett, D. H. Farber, and C. C. Mundt. Consequences of long-distance dispersal for epidemic spread: Patterns, scaling, and mitigation. Plant disease, 103(2):177–191, 2019.

W. Shan, Y. Yin, H. Lu, and S. Liang. A meteorological analysis of ozone episodes using hysplit model and surface data. Atmospheric Research, 4(93):767–776, 2009.

M. Shaw and M. Pautasso. Networks and plant disease management: Concepts and applications. Annual Review of Phytopathology, 52:477–493, 2014.

M. W. Shaw and T. M. Osborne. Geographic distribution of plant pathogens in response to climate change. Plant Pathology, 60(1):31–43, 2011.

M. D. Shirley and S. P. Rushton. The impacts of network topology on disease spread. Ecological Complexity, 2(3):287–299, 2005.

K. Siła-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demšar, and A. S. Fotheringham. Analysis of human mobility patterns from gps trajectories and contextual information. International Journal of Geographical Information Science, 30(5):881–906, 2016.

W. Smets, S. Moretti, S. Denys, and S. Lebeer. Airborne bacteria in the atmosphere: presence, purpose, and potential. Atmospheric Environment, 139:214–221, 2016.

K. M. Smith, C. C. Machalaba, R. Seifman, Y. Feferholtz, and W. B. Karesh. Infectious disease and economics: The case for considering multi-sectoral impacts. One Health, 7:100080, 2019.

S. Soubeyrand. Construction of semi-markov genetic-space-time seir models and inference. 2016.

S. Soubeyrand, F. Carpentier, F. Guiton, and E. K. Klein. Approximate bayesian computation with functional statistics. Statistical Applications in Genetics and Molecular Biology, 12(1):17–37, 2013.

J. Staples, D. A. Nickerson, and J. E. Below. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. Genetic epidemiology, 37(2):136–141, 2013.

L. Steele, E. Orefuwa, and P. Dickmann. Drivers of earlier infectious disease outbreak detection: a systematic literature review. International Journal of Infectious Diseases, 53:15–20, 2016.

A. Stein, R. R. Draxler, G. D. Rolph, B. J. Stunder, M. Cohen, and F. Ngan. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. Bulletin of the American Meteorological Society, 96(12):2059–2077, 2015.

S. H. Strogatz. Exploring complex networks. Nature, 410(6825):268, 2001.

G. Strona, C. J. Carstens, and P. S. Beck. Network analysis reveals why xylella fastidiosa will persist in europe. Scientific Reports, 7(1):71, 2017.

A. Talbi, Y. Kerchich, R. Kerbachi, and M. Boughedaoui. Assessment of annual air pollution levels with PM1, PM2.5, PM10 and associated heavy metals in Algiers, Algeria. Environmental Pollution, 232:252–263, 2018.

D. Tanaka, K. Sato, M. Goto, S. Fujiyoshi, F. Maruyama, S. Takato, T. Shimada, A. Sakatoku, K. Aoki, and S. Nakamura. Airborne microbial communities at high-altitude and suburban sites in toyama, japan suggest a new perspective for bioprospecting. Frontiers in bioengineering and biotechnology, 7:12, 2019.

J. W. Tang. The effect of environmental parameters on the survival of airborne infectious agents. Journal of the Royal Society Interface, 6(suppl_6):S737–S746, 2009.

S. Tang. A modified si epidemic model for combating virus spread in wireless sensor networks. International Journal of Wireless Information Networks, 18(4):319–326, 2011.

M. Tantardini, F. Ieva, L. Tajoli, and C. Piccardi. Comparing methods for comparing networks. Scientific Reports, 9(1):17557, 2019.

P. L. Thompson, B. Rayfield, and A. Gonzalez. Loss of habitat and connectivity erodes species diversity, ecosystem functioning, and stability in metacommunity networks. Ecography, 40(1):98–108, 2017.

M. Tizzoni, P. Bajardi, A. Decuyper, G. K. K. King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza. On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol, 10(7):e1003716, 2014.

W. R. Tobler. A computer movie simulating urban growth in the Detroit region. Economic geography, 46(sup1):234–240, 1970.

M. W. Traunmueller, N. Johnson, A. Malik, and C. E. Kontokosta. Digital footprints: Using wifi probe and locational data to analyze human mobility trajectories in cities. Computers, Environment and Urban Systems, 72:4–12, 2018.

E. A. Treml, P. N. Halpin, D. L. Urban, and L. F. Pratson. Modeling population connectivity by ocean currents, a graph-theoretic approach for marine conservation. Landscape Ecology, 23(1):19–36, 2008.

J. Uetake, Y. Tobo, Y. Uji, T. C. J. Hill, P. J. DeMott, S. M. Kreidenweis, and R. Misumi. Seasonal changes of airborne bacterial communities over Tokyo and influence of local meteorology. Frontiers in Microbiology, 10:1572, 2019.

D. Urban and T. Keitt. Landscape connectivity: a graph-theoretic perspective. Ecology, 82(5):1205–1218, 2001.

C. Vacher, A. Hampe, A. Port, U. Sauer, S. Compant, and C. Morris. The Phyllosphere: Microbial jungle at the plant-climate interface. Annual Review of Ecology Evolution and Systematics, 47, 12 2016.

F. Van den Bosch, J. Zadoks, and J. Metz. Continental expansion of plant disease: a survey of some recent results. In Predictability and Nonlinear Modelling in Natural Sciences and Economics, pages 274–281. Springer, 1994.

P. Van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Mathematical biosciences, 180(1-2):29–48, 2002.

J. Van Groenigen and A. Stein. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality, 27(5):1078–1086, 1998.

D. Van Ravenzwaaij, P. Cassey, and S. D. Brown. A simple introduction to markov chain monte–carlo sampling. Psychonomic bulletin & review, 25(1):143–154, 2018.

K. VanderWaal, A. Perez, M. Torremorrell, R. M. Morrison, and M. Craft. Role of animal movement and indirect contact among farms in transmission of porcine epidemic diarrhea virus. Epidemics, 24:67–75, 2018.

M. C. Vernon and M. J. Keeling. Representing the uk's cattle herd as static and dynamic networks. Proceedings of the Royal Society B: Biological Sciences, 276(1656):469–476, 2009.

D. Vokou, K. Vareli, E. Zarali, K. Karamanoli, H.-I. A. Constantinidou, N. Monokrousos, J. M. Halley, and I. Sainis. Exploring biodiversity in the bacterial community of the Mediterranean phyllosphere and its relationship with airborne bacteria. Microbial Ecology, 64(3):714–724, 2012.

E. Volz and L. A. Meyers. Susceptible–infected–recovered epidemics in dynamic contact networks. Proceedings of the Royal Society B: Biological Sciences, 274(1628): 2925–2934, 2007.

A. Wain, S. Lee, G. Mills, G. Hess, M. Cope, and N. Tindale. Meteorological overview and verification of hysplit and aaqfs dust forecasts for the duststorm of 22-24 october 2002. Australian Meteorological Magazine, 55(1):35–46, 2006.

H. Wallace. Dispersal in time and space: soil pathogens. Plant disease, pages 181–202, 2012.

J. Wallinga, W. J. Edmunds, and M. Kretzschmar. Perspective: human contact patterns and the spread of airborne infectious diseases. Trends in Microbiology, 7(9):372–377, 1999.

H. Wang, X. Yang, and Z. Ma. Long-distance spore transport of wheat stripe rust pathogen from Sichuan, Yunnan, and Guizhou in southwestern China. Plant Disease, 94(7):873–880, 2010.

J. Wang, H. Mo, F. Wang, and F. Jin. Exploring the network structure and nodal centrality of china's air transport network: A complex network approach. Journal of Transport Geography, 19(4):712–721, 2011.

J.-F. Wang, A. Stein, B.-B. Gao, and Y. Ge. A review of spatial sampling. Spatial Statistics, 2:1–14, 2012.

X. F. Wang and G. Chen. Complex networks: small-world, scale-free and beyond. IEEE circuits and systems magazine, 3(1):6–20, 2003.

H. M. Ward et al. Researches on the life history of hemileia vastatrix. Journal of the Linnaean Society., 19:299–335, 1882.

T. Wark, P. Corke, P. Sikka, L. Klingbeil, Y. Guo, C. Crossman, P. Valencia, D. Swain, and G. Bishop-Hurley. Transforming agriculture through pervasive wireless sensor networks. IEEE Pervasive Computing, 6(2):50–57, 2007.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. nature, 393(6684):440–442, 1998.

T. Wei and V. Simko. R package corrplot: Visualization of a Correlation Matrix, 2017. URL https://github.com/taiyun/corrplot.

D. B. West et al. Introduction to graph theory, volume 2. Prentice Hall Upper Saddle River, NJ, 1996.

J. S. West. Plant pathogen dispersal. eLS, 2014.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

K. Wickwire. Mathematical models for the control of pests and infectious diseases: a survey. Theoretical population biology, 11(2):182–238, 1977.

M. J. Williams and M. Musolesi. Spatio-temporal networks: reachability, centrality and robustness. Royal Society open science, 3(6):160196, 2016.

M. E. Wilson. Geography of infectious diseases. Infectious Diseases, page 1055, 2010.

A. M. Womack, B. J. M. Bohannan, and J. L. Green. Biodiversity and biogeography of the atmosphere. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1558):3645–3653, 2010.

H. Woźniakowski. Efficiency of quasi-monte carlo algorithms for high dimensional integrals. In Monte-Carlo and Quasi-Monte Carlo Methods 1998, pages 114–136. Springer, 2000.

Z. Xu and D. Z. Sui. Effect of small-world networks on epidemic propagation and intervention. Geographical Analysis, 41(3):263–282, 2009.

N. Yamamoto, W. W. Nazaroff, and J. Peccia. Assessing the aerodynamic diameters of taxon-specific fungal bioaerosols by quantitative PCR and next-generation DNA sequencing. Journal of Aerosol Science, 78:1 – 10, 2014. ISSN 0021-8502.

L.-X. Yang and X. Yang. A new epidemic model of computer viruses. Communications in Nonlinear Science and Numerical Simulation, 19(6):1935–1944, 2014.

M. Youssef and C. Scoglio. An individual-based approach to sir epidemics in contact networks. Journal of theoretical biology, 283(1):136–144, 2011.

D. Yu, M. Kim, G. Xiao, and T. H. Hwang. Review of biological network data and its applications. Genomics & informatics, 11(4):200, 2013.

H. Yuan and G. Chen. Network virus-epidemic model with the point-to-group information propagation. Applied Mathematics and Computation, 206(1):357–367, 2008.

P. Zamberletti, M. Zaffaroni, F. Accatino, I. F. Creed, and C. De Michele. Connectivity among wetlands matters for vulnerable amphibian populations in wetlandscapes. Ecological Modelling, 384:119–127, 2018.

M. Zanin and F. Lillo. Modelling the air transport with complex networks: A short review. The European Physical Journal Special Topics, 215(1):5–21, 2013.

J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao. Identifying a set of influential spreaders in complex networks. Scientific Reports, 6:27823, 2016.

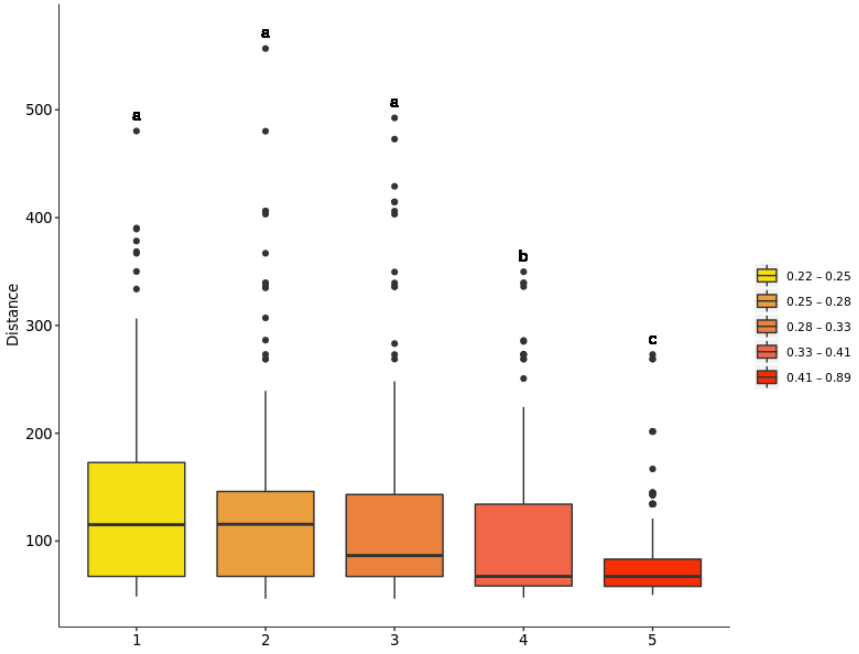
X. Zhou, Y. Hu, Y. Wu, and X. Xiong. Influence analysis of information erupted on social networks based on sir model. International Journal of Modern Physics C, 26 (02):1550018, 2015.

Z. Zhu and M. L. Stein. Spatial sampling design for prediction with estimated parameters. Journal of agricultural, biological, and environmental statistics, 11(1):24, 2006.

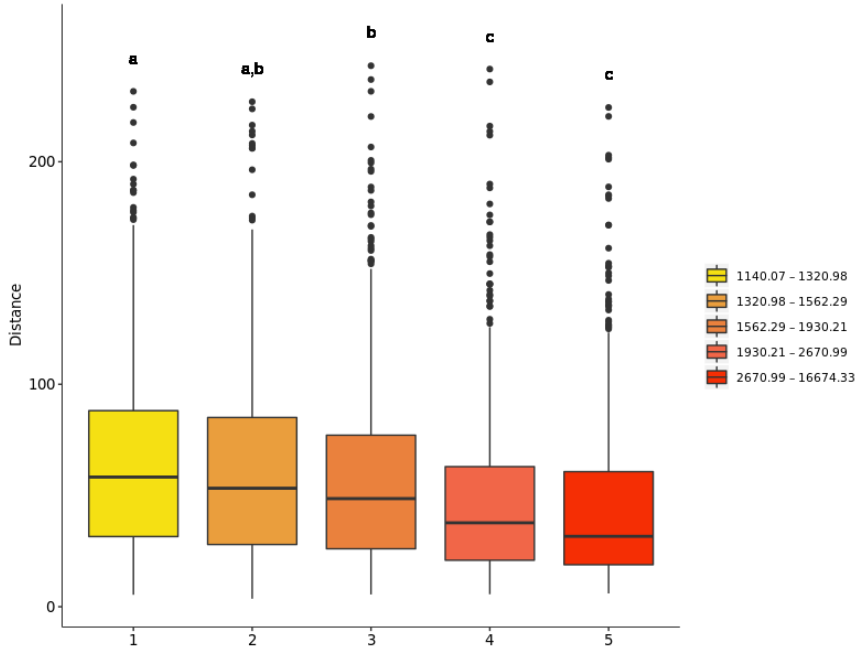
X.-N. Zuo, R. Ehmke, M. Mennes, D. Imperati, F. X. Castellanos, O. Sporns, and M. P. Milham. Network centrality in the human functional connectome. Cerebral cortex, 22(8):1862–1875, 2012.

ANNEXES

A. Appendix of Chapter 2

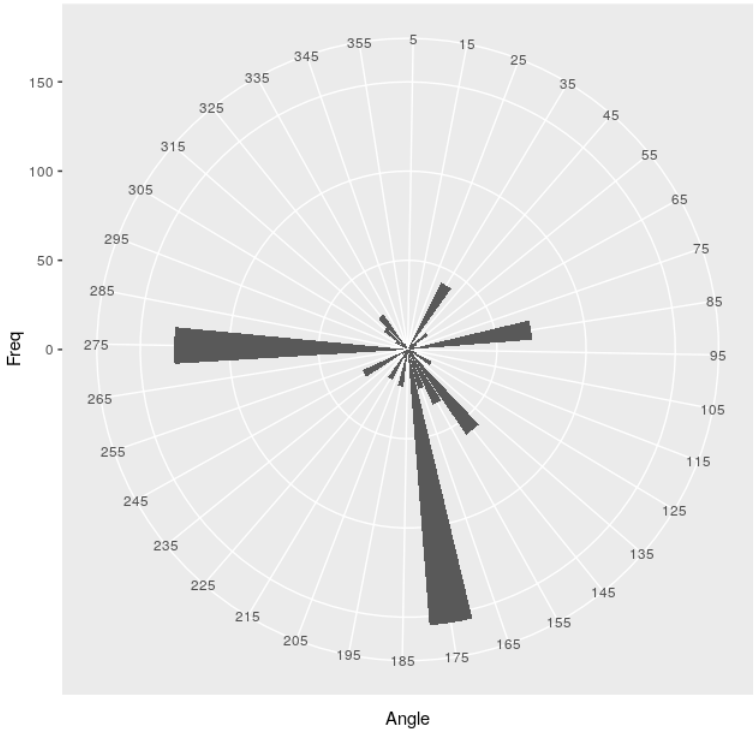


(a) Mediterranean region

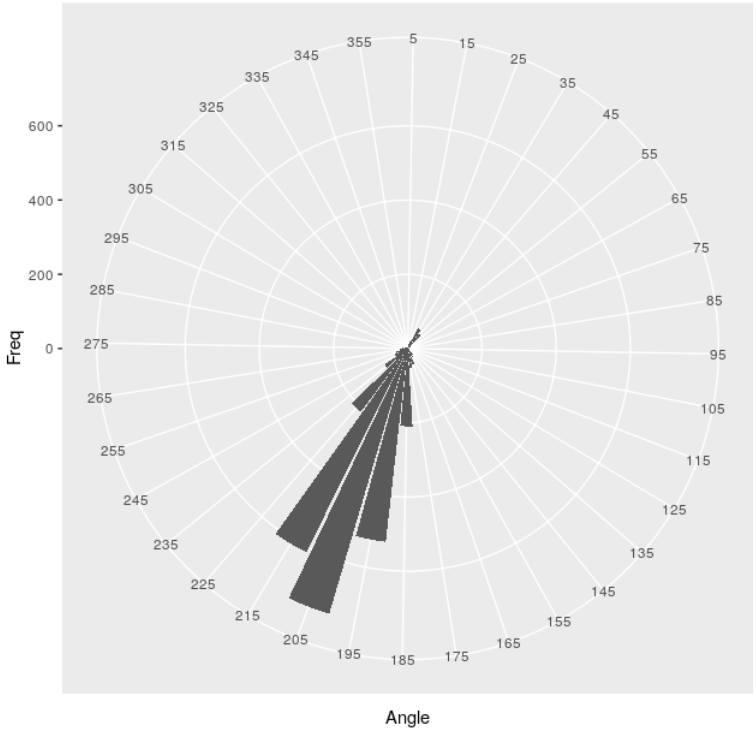


(b) PACA

Figure .5. – Boxplot for the distance between the points within the categories. The categories correspond to the intensity of connection between the nodes (a) of the Mediterranean region and (b) PACA. The letters are chosen according to the resulting p-values of Wilcoxon signed-rank test, based on the significance level of 0.05, to compare the distribution of the distances between every couple of the categories in Figure .5. The categories having the same letter doesn't have a significant difference between them.



(a) Mediterranean region



(b) PACA

Figure .6. – Circular histogram illustrating the direction of the connectivities between the nodes whithin the categories (a) of the Mediterranean region and (b) PACA.

B. Appendix of Chapter 4

Time period	Diameter	Density	Transitivity	Degree Correlation	Shortest Path
2011-2017	0.003	0.278	0.74	0.31	0.0008
2011	0.030	0.175	0.678	0.132	0.0062
2012	0.024	0.176	0.673	0.085	0.0061
2013	0.019	0.180	0.668	0.232	0.0061
2014	0.082	0.148	0.659	0.239	0.0065
2015	0.068	0.163	0.661	0.236	0.0062
2016	0.553	0.156	0.660	0.242	0.0063
2017	0.126	0.171	0.657	0.175	0.0061
January	0.23	0.131	0.636	0.168	0.0111
February	0.72	0.138	0.631	0.275	0.0123
March	0.29	0.136	0.629	0.266	0.0111
April	0.70	0.135	0.640	0.221	0.0113
May	0.99	0.126	0.635	0.181	0.0116
June	1.14	0.116	0.616	0.176	0.0119
July	1.36	0.102	0.600	0.061	0.0115
August	1.81	0.098	0.599	0.125	0.0118
September	1.52	0.100	0.606	0.168	0.0123
October	0.12	0.102	0.623	0.205	0.0121
November	0.10	0.113	0.629	0.378	0.0127
December	1.03	0.131	0.619	0.369	0.0112

Table .1. – Network indices (Diameter, density, transitivity, degree correlation, average shortest path) calculated from the networks covering the Mediterranean region and estimated in three temporal contexts: the entire period 2011-2017, yearly time periods from 2011 to 2017 and monthly time periods.

C. Appendix of chapter 6

frontiers in Microbiology

ORIGINAL RESEARCH
published: 25 September 2018
doi: 10.3389/fmicb.2018.02257

Assessing the Aerial Interconnectivity of Distant Reservoirs of *Sclerotinia sclerotiorum*

Christel Leyronas[1], Cindy E. Morris[1], Maria Choufany[2] and Samuel Soubeyrand[2]*

[1] *Pathologie Végétale, INRA, Montfavet, France,* [2] *BioSP, INRA, Avignon, France*

Many phytopathogenic fungi are disseminated as spores via the atmosphere from short to long distances. The distance of dissemination determines the extent to which plant diseases can spread and novel genotypes of pathogens can invade new territories. Predictive tools including models that forecast the arrival of spores in areas where susceptible crops are grown can help to more efficiently manage crop health. However, such models are difficult to establish for fungi with broad host ranges because sources of inoculum cannot be readily identified. *Sclerotinia sclerotiorum*, the pandemic agent of white mold disease, can attack >400 plant species including economically important crops. Monitoring airborne inoculum of *S. sclerotiorum* in several French cropping areas has shown that viable ascospores are present in the air almost all the time, even when no susceptible crops are nearby. This raises the hypothesis of a distant origin of airborne inoculum. The objective of the present study was to determine the interconnectivity of reservoirs of *S. sclerotiorum* from distant regions based on networks of air mass movement. Viable airborne inoculum of *S. sclerotiorum* was collected in four distinct regions of France and 498 strains were genotyped with 16 specific microsatellite markers and compared among the regions. Air mass movements were inferred using the HYSPLIT model and archived meteorological data from the global data assimilation system (GDAS). The results show that up to 700 km could separate collection sites that shared the same haplotypes. There was low or no genetic differentiation between strains collected from the four sites. The rate of aerial connectivity between two sites varied according to the direction considered. The results also show that the aerial connectivity between sites is a better indicator of the probability of the incoming component (PIC) of inoculum at a given site from another one than is geographic distance. We identified the links between specific sites in the trajectories of air masses and we quantified the frequencies at which the directional links occurred as a proof-of-concept for an operational method to assess the arrival of airborne inoculum in a given area from distant origins.

Keywords: airborne inoculum, air-mass movement, contact network, polyphagous fungi, risk forecasting

OPEN ACCESS

Edited by:
Pierre Amato,
UMR6296 Institut de Chimie de Clermont-Ferrand (ICCF), France

Reviewed by:
Dale Warren Griffin,
United States Geological Survey, United States
M. Elias Dueker,
Bard College, United States

****Correspondence:***
Christel Leyronas
christel.leyronas@inra.fr

Specialty section:
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology

Received: *28 May 2018*
Accepted: *05 September 2018*
Published: *25 September 2018*

Citation:
Leyronas C, Morris CE, Choufany M and Soubeyrand S (2018) Assessing the Aerial Interconnectivity of Distant Reservoirs of Sclerotinia sclerotiorum. Front. Microbiol. 9:2257. doi: 10.3389/fmicb.2018.02257

INTRODUCTION

Many phytopathogenic fungi are disseminated via the atmosphere from micro- to macro-geographical scales in the form of spores (Aylor et al., 1982; Brown and Hovmoller, 2002; Prospero et al., 2005). The most striking examples of long distance dispersal (LDD) can be illustrated by intercontinental movements of spores leading to rust emergence in previously unaffected areas. Evidence for atmospheric transoceanic jumps has been reported for coffee leaf rust, sugarcane rust and wheat stem rust, for example (Bowden et al., 1971; Purdy et al., 1985; Brown and Hovmoller, 2002). Long distance dispersal can also lead to the annual recolonization of some areas by periodic migrations of aerial spores. Indeed some pathogens make long distance jumps from one susceptible host to another throughout the growing season following prevailing winds. This is, for example, the case of Oomycetes responsible for tobacco blue mold and cucurbit downy mildew. It is also the case for cereal stem rust that migrates following the "*Puccinia* pathway" in the United States wheat belt, or for wheat yellow rust in China (Aylor et al., 1982; Hovmoller et al., 2002; Ojiambo and Holmes, 2011). Finally, LDD can lead to the invasion of areas where plant disease is already present by novel strains. This is the recent case for Ug99, a race of *P. graminis tritici,* which is virulent on cultivars of wheat carrying a particular, widely used resistance gene thereby making it an important threat to world wheat production. Modeled wind trajectories indicate that air movements are likely responsible for the spread of Ug99 in Africa and for its predicted trajectory into Asia (Singh et al., 2011; Meyer et al., 2017).

Forecasting the movements of fungal pathogens in the atmosphere is useful for anticipating epidemic risks. To forecast the arrival of spores in areas where susceptible crops are grown, local and distant sources of spores need to be identified and air-mass trajectories modeled (Isard et al., 2005; Tao et al., 2009). Predictive tools can help growers in rationalizing their practices, particularly chemical control, and thus can lead to a reduction in the number of fungicide applications. Studying atmospheric pathways frequently taken by plant pathogens may also be useful in organizing crops in regional landscapes, e.g., to reduce the cultivation of identical susceptible varieties in areas located on a dispersal route of a threatening pathogen (Singh et al., 2015; Meyer et al., 2017). Forecasting epidemic risks can be achieved for host-specific, obligate parasitic fungi such as rust or some oomycetes whose sources, and hosts along their trajectory can be readily identified (IPM-PIPE) (Isard et al., 2005; Tao et al., 2009). Indeed, each of these parasitic fungal species is able to grow on a very limited number of plant species. Moreover the fungi colonize only living host plants and cannot live saprophytically. Thus the identification of fields that are potential sources or sink of spores can be readily determined.

In contrast with host specific rusts and oomycetes, many phytopathogenic fungi are generalists with many environmental sources including host plants as well as substrates for saprophytic growth. Therefore, in such cases forecasting spore arrival is particularly difficult because sources of inoculum cannot be readily identified. Some fungal spores have been found up to several kilometers above ground level, above oceans or far away from any sources (Hirst et al., 1967; Maldonado-Ramirez et al., 2005; Prospero et al., 2005; Smith et al., 2012; Damialis et al., 2017) indicating long-distance dispersal events. Tracking the movements in the atmosphere of phytopathogenic fungi with large host ranges, in order to forecast epidemic risk, is challenging.

Sclerotinia sclerotiorum, the pandemic causal agent of white mold disease, can attack more than 400 plant species including some with economic importance such as field crops and vegetables (Purdy, 1979; Boland and Hall, 1994). Monitoring of airborne inoculum of *S. sclerotiorum* in several French cropping areas has shown that viable ascospores were present in the air almost all the time, even when there were no susceptible crops nearby suggesting that spores were from distal sources. This raises the hypothesis of a distant origin of airborne inoculum. Several approaches can be used to assess the pathway of spores over a large area (Golan and Pringle, 2017). Whereas monitoring fungal propagules with air samplers can give an idea about abundance of spores at a definite time in a definite location, it has limited application for assessing LDD. Molecular approaches allow comparative genetics of populations across a geographic range. However, they give little information about dispersal mechanisms and about the moment (recent or more ancient) when they occurred. Atmospheric transport models can be used for back-trajectory analysis to determine the origin of air masses and establish source–receptor relationships (Stein and Ngan, 2015a) while taking into account the errors of these estimates. Finally the most efficient methods to assess LDD of fungal spores are the ones that combine the three approaches (Golan and Pringle, 2017).

The objective of this study was to establish a method to determine the relationships of reservoirs of *S. sclerotiorum* and if certain strains are likely to be exchanged between regions located several 100 km apart. Based on air samplings, genetic characteristics of regional populations were compared. Aerial interconnectivity (networks of the air masses that would have transported spores aerially) was assessed by using archived meteorological data providing information about air mass movements.

MATERIALS AND METHODS

Sampling of Viable Airborne Inoculum

Air sampling was carried out at four distinct regions of France between 2014 and 2016 (**Table 1**). Airborne propagules were collected using a volumetric sampler called Portable Air Sampler for Agar Plates (Burkard Manufacturing, Rickmansworth, United Kingdom) with a flow rate of 20 L min^{-1}. Petri plates were filled with a semi-selective medium amended with bromophenol blue (Steadman et al., 1994). The samplers were set to run for 9 min twice during the day at each sampling date. One plate was used for each 9 min sampling. Samplers were placed in the North region of France (N) in a chicory witloof field (*Cichorium intybus* L.), in the Center-West region (CW) in a cantaloupe field (*Cucumis melo* L.) and in the South-West (SW) and North-West regions (NW) in carrot fields (*Daucus carota* L.).

TABLE 1 | Genetic characteristics of *S. sclerotiorum* isolates collected in the air at four different French regions.

	Sample size	Year of collection	Nb of distinct MLH[a]	Haplotypic diversity	Allelic richness[b]	Hnb[c]
Total	498	2014–15–16	241	0.48	5.36	0.62 (0.19)
North	105	2014–15–16	59	0.55	4.47	0.59 (0.18)
North–West	18	2014–15–16	17	0.94	4.13	0.60 (0.17)
Center–West	49	2014–16	37	0.75	5.68	0.64 (0.21)
South–West	326	2014–15	154	0.47	5.27	0.60 (0.20)

[a]*Multilocus haplotype.* [b]*Unbiased allelic richness corrected for a minimum sample size of 18 isolates.* [c]*Unbiased gene diversity (standard deviation between brackets).*

After exposure, the plates were incubated in the lab at room temperature (ca 22°C) and the presence of *Sclerotinia* sp. on the selective medium was assessed by the development of yellow halos caused by the production of oxalic acid (Steadman et al., 1994). Mycelial colonies associated with yellow halos were transferred to fresh PDA (Potato Dextrose Agar) to obtain pure cultures. Isolates showing the typical morphological features of *Sclerotinia* sp. were subjected to a step of single hypha isolation prior to their entry in the fungal collection of the laboratory after which they were considered as strains (i.e., pure lines). For this, single pieces of hyphal tip were excised from the growing margin of a colony after 2 days of incubation on PDA and transferred to fresh medium as described by Lehner et al. (2016). The single-hypha strains were then stored as sclerotia at −20°C.

Genotyping

Sclerotia, typical mycelial stroma formed by *S. sclerotiorum*, were spread on PDA to germinate and produce mycelium. Genomic DNA was extracted in 96-well plates from aliquots of 100 mg (fresh weight) of frozen mycelium, following the Dneasy Plant extraction Kit protocol (Qiagen). Sixteen microsatellite markers designed for *S. sclerotiorum* by Sirjusingh and Kohn (2001) were amplified with forward primers conjugated with fluorescent dyes following the protocol described by Leyronas et al. (2018). Reverse primers did not carry any fluorescent. To determine the size of the microsatellites, the PCR products were diluted and multiplexed prior to scanning with an ABI 3730 sequencer (Applied Biosystems). GeneMapper software version 4.1 (Applied Biosystems) was then used for the microsatellite size analysis. Complete microsatellite size profiles (referred to as "haplotypes" hereafter) were obtained for all strains.

Genetic Characterization of Airborne *S. sclerotiorum* Strains

Unbiased gene diversity (Hnb) and unbiased allelic richness were computed separately for the strains collected in distinct regions with the Genetix software (Belkhir et al., 1996–2004). The number of different multilocus haplotypes (MLH) was computed with GenClone 1.0 software (Arnaud-Haond and Belkhir, 2007). We used the index of haplotypic diversity (based on the number of individuals and the number of distinct MLH), which estimates the proportion of haplotypes present in a population and is equal to 1 when a population is composed exclusively of unique haplotypes (Arnaud-Haond et al., 2007). The software FSTAT version 2.9.3 (Goudet, 1995) was used to compute allelic richness per locus corrected for a minimum sample size of 18 isolates. Genetic differentiation was estimated with a hierarchical analysis of molecular variance (AMOVA) using Arlequin version 3.5 (Excoffier et al., 2005). AMOVA was used to determine the proportion of genetic variation partitioned among the four regions, among the years of collection within a region, or among isolates within a sampling year. The number of permutations to test for significance was set at 8,000. Arlequin was also used to assess genetic differentiation between strains of pairs of regions by computing R_{ST} values (Slatkin, 1995).

Two-Part Decomposition of an Observed Pathogen Composition

To decompose the composition of aerial samples into strains of local and distant origins we used the following logic:

We let $N_i = (N_{i1}, \ldots, N_{in})$ denote the counts of isolates from strains 1, … , n collected from site i. Then N_i was split into two parts: the part issued from site j (via long-distance dispersal) and a part not issued from site j (i.e., either generated locally or issued from other distant sites) (**Figure 1**). Here, we make the crude assumption that for any strain s that was collected in both sites i and j, the corresponding count N_{is} is issued from site j and, reversely, for any strain s that was collected in site i but not in site j, the corresponding count j is not issued from site j. Mathematically, this translates into: $N_i = f_j N_i - \left(1 - f_j\right) N_i$ where f_i is a vector of dimension n whose component s is equal to 1 if strain s was collected in site j; otherwise it equals 0. Thus, f_i plays the role of a filter and $f_i N_i$ stands for the part in N_i issued from site j whereas $(1 - f_j)N_i$ stands for the part in N_i not issued from site j. The specification used here for f_i implies that all strains in j have an equal propensity to be dispersed from j to i and an equal propensity to establish in i.

How likely $f_j N_i$ issues from site j can be measured by assuming that this vector of counts is drawn from a multinomial distribution with the vector of probability$p_j = N_j / \sum_{s=1}^{n} N_{js}$, i.e., the proportion of the strain in site j. We let $q_{j \to i}$ denote the multinomial probability computed at $f_j N_i$ with parameter p_j. Thereafter, $q_{j \to i}$ is called the probability of the incoming component (PIC) in site i from site$j$.

To avoid inconsistencies because of varying sample sizes at different sites, probability $q_{j \to i}$ was computed by subsampling with replacement k isolates in N_i and N_j, where k is the minimum sample size across all sites (not only i and j), computing f_j^* and

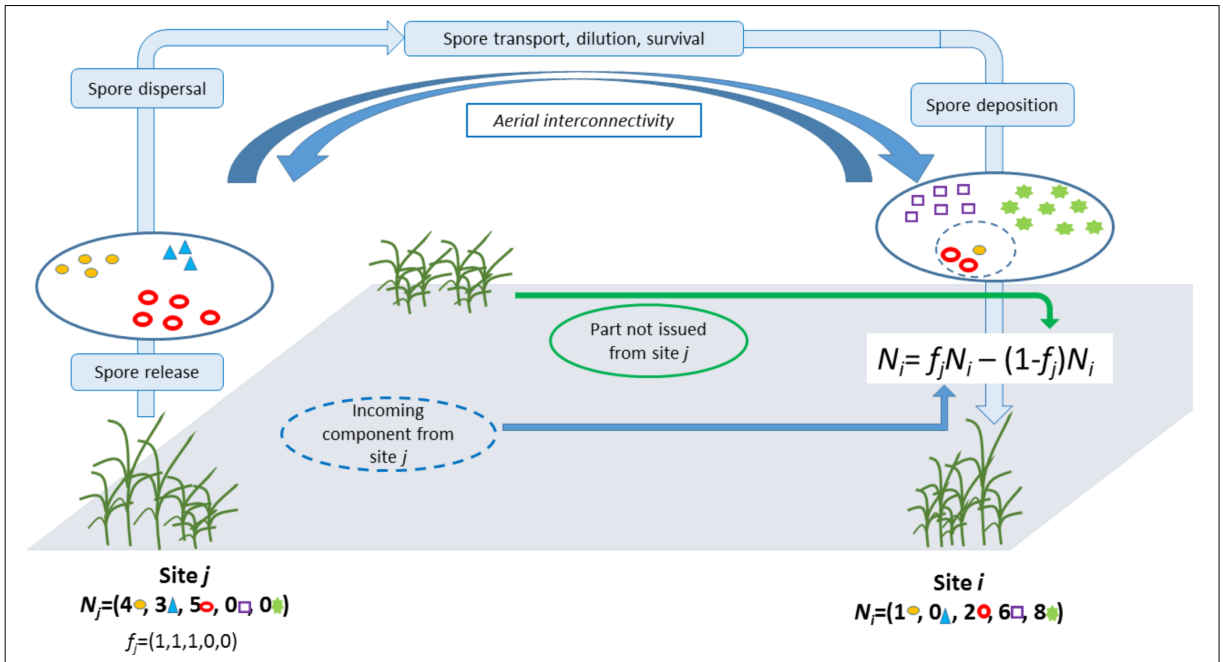


FIGURE 1 | Aerobiological phases of spore dispersal and composition of aerial samples.

p_j^* from the subsampled vectors of counts, say N_i^* and N_j^*, and computing the PIC $q_{j\to i}^*$ in site i from site j using N_j^* and a subsample of size 1 of N_i^*. This procedure was repeated 10^4 times and the average of the 10^4 $q_{j\to i}^*$ runs was used as the value of $q_{j\to i}$.

Relationship Between Incoming Components of the Pathogen Population and Tropospheric Connectivity

We aim to explore the association between (a) the probability that the different components of the population sampled from the air are incoming from elsewhere (PIC) and (b) the connectivity between sites via tropospheric movements. The rationale is that the greater the connectivity, the greater the PIC. This association was assessed with a log-linear regression:

$log q_{j\to i} = \alpha + \beta C_{j\to i} + \varepsilon_{j\to i}$ where $C_{j\to i}$ measures the directional connectivity from site j to site i, α is the intercept coefficient, β is the slope coefficient and $\varepsilon_{j\to i}$ is random noise.

Directional connectivity was inferred by exploiting archived meteorological data provided by the Global Data Assimilation System (GDAS) of NOAA and the software HYSPLIT (Stein et al., 2015b; Rolph et al., 2017) that allows the modeling of air mass trajectories from the aforementioned data. For each date between 2008-01-01 and 2017-12-31, i.e., a 10 years period over which GDAS data are available, HYSPLIT allowed us to reconstruct over 48 h the trajectory $H(t, x)$ of the air mass arriving at location × (1,000 m above the sea) and date t (12:00 GMT). We computed the connectivity from site j to site i as the fraction of trajectories arriving at site i and previously going throughout a buffer zone around site j. Here, the connectivity is considered in a multisite context, and can therefore be expressed in the graph theory framework. A graph $G = (V, E)$ is characterized by a set V of vertices and a set E of links between vertices, i.e., the edges of the graph. In our application, the vertices correspond to the sampling sites and the edges are weighted by the possibility of pathogen diffusion between two different sites via air mass movements, namely the directional connectivity terms. For our application, we calculated different connectivities corresponding to the 12 months of the year. Thus, each monthly connectivity value was based on approximately 300 (10 years × 30 days) binary variables indicating whether or not an air mass arriving at a given site went through the vicinity of another given site. More precisely, the monthly connectivity for the m-th month of the year from site j to site i was computed as follows:

$$C_{j\to i}^{(m)} = \frac{1}{K_m}\sum_{k=1}^{K_m} 1\left(H\left(x_i, t_{mk}\right) \cap B\left(x_j, r\right) \neq \emptyset\right)$$

where $\{t_{mk} : k = 1, \ldots, K_m\}$ are all dates of the m-th month of the year between 2008-01-01 and 2017-12-31, $B\left(x_j, r\right)$ is a circular buffer of radius r around site j, and $1\left(H\left(x_i, t_{mk}\right) \cap B\left(x_j, r\right)\right)$ is the indicator function equal to 1 if the intersection between the trajectory $H\left(x_i, t_{mk}\right)$ and the buffer $B\left(x_j, r\right)$ is not empty; otherwise it equals to 0. In our application, buffer zones were defined as disks of radius 25 km centered on sampling sites, and allowed us to take into account the uncertainty in the reconstruction of air mass trajectories. The use of buffer zones also implies that an

assumption of local stability in pathogen composition is made. Thereafter, for the sake of simplicity in the notation, we omit the index m.

The log-linear regression between $q_{j\to i}$ and $C_{j\to i}$ was fitted to data with a least-squares approach. To achieve a satisfactory statistical power, for each pair (i, j) we only used the point $(q_{j\to i}, C_{j\to i})$ or $(q_{j\to i}, C_{j\to i})$ corresponding to the maximum connectivity. The significance of the link was assessed via a randomization directional test (Manly, 1997) assessing whether the slope coefficient β is equal to 0, the alternative being that β is positive. Under randomization, the regression is expected to be flat (i.e., $\beta = 0$), with some variations. For each test, 10^4 randomizations were independently carried out by conserving the network structure of data: for each repetition, identifiers of sites were simply randomly re-sampled without replacement for the connectivities and the following log-linear regression was fitted to the randomized dataset:

$log q_{j\to i} = \alpha + \beta C^*_{j\to i} + \varepsilon^*_{j\to i}$ where $C^*_{j\to i}$ denotes the randomized directional connectivity from site j to site i. The p-value of the test was computed as the proportion of estimates β^* obtained for the 10^4 repetitions larger than the estimate β obtained with non-randomized data.

Simulation Study

A simulation study was designed to assess the efficiency of the method aimed at linking the tropospheric connectivity with the PIC and to compare different sampling schemes varying in the number of sampling sites, I, and the number of sampled isolates per site, M. For this goal, we generated datasets from an empirical simulation model that does not represent real processes, but produces multi-site pathogen compositions with various levels of similarities depending on arbitrarily generated connectivities.

In the model, for any target site i and any source site $j \neq i$, the tropospheric connectivity $C_{j\to i}$ was drawn in an exponential distribution with mean 0.02. Thus, 95% of connectivities are lower than 0.060, and 99% of connectivities are lower than 0.092.

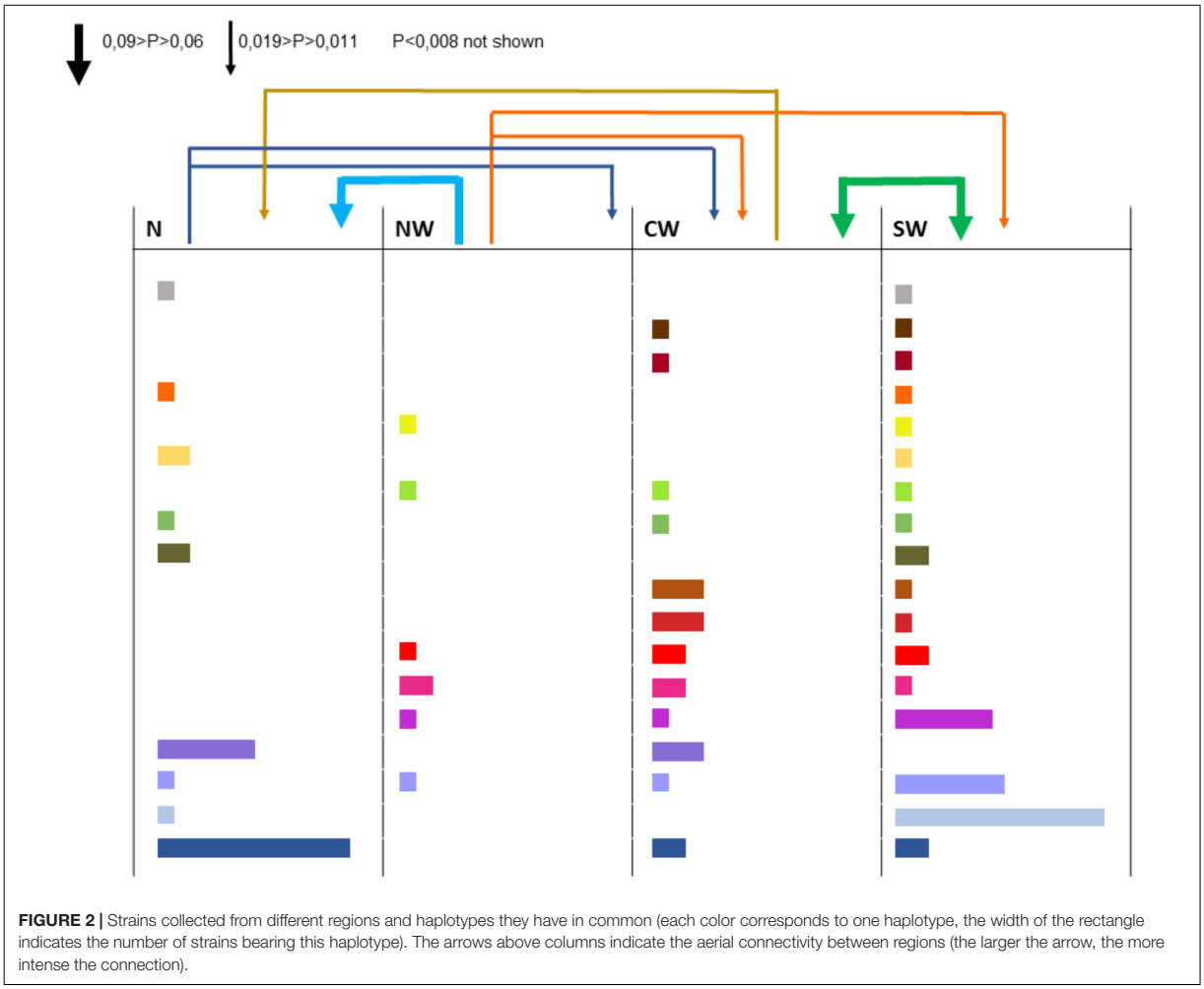


FIGURE 2 | Strains collected from different regions and haplotypes they have in common (each color corresponds to one haplotype, the width of the rectangle indicates the number of strains bearing this haplotype). The arrows above columns indicate the aerial connectivity between regions (the larger the arrow, the more intense the connection).

Each site i was supposed to be occupied by a proportion $1 - \tau_i$ of endemic strains of the pathogen and a proportion τ_i of exogenous strains, where τ_i depends on the connectivities toward site i and a parameter, τ_{max}, giving the maximum exogenous fraction of strains of the pathogen at the site. The set of endemic strains in i consists of K strains with random proportions. The set of exogenous strains in i is potentially made of all the strains that are endogenous in distant sites $j \in \{1, \ldots, I\} - \{i\}$ such that $C_{j \to i} > 0$ (here, for the sake of simplicity, we suppose that the set of I sampling sites form a closed system).

More precisely, pathogen compositions N_i in site $i \in \{1, \ldots, I\}$ were drawn with the following three-stage procedure.

(1) The proportion τ_i of exogenous strains satisfies.
(2) $\tau_i = \tau_{max} \frac{C_{\cdot \to i}}{\sum_{i'=1}^{I} C_{\cdot \to i'}}$,
(3) where $C_{\cdot \to i} = \sum_{j \neq i} C_{j \to i}$ denotes the overall connectivity of site i.
(4) Proportions $\gamma_{(i-1)K+1}, \ldots, \gamma_{(i-1)K+K}$ of the K endemic strains in site i, that are indexed by $(i-1)K+1, \ldots, (i-1)K+K$, are randomly generated by simulating K independent values U_{ik} $(k = 1, \ldots, K)$ in the uniform distribution over $[0, 1]$ and setting $\gamma_{(i-1)K+k} = U_{ik} / \sum_{k'=1}^{K} U_{ik'}$. Then, counts of sampled isolates of endemic strains in site i, denoted by $N_{i,(i-1)K+1}, \ldots, N_{i,(i-1)K+K}$, are obtained by a multinomial draw with size $[(1 - \tau_i) M]$ and probabilities $\gamma_{(i-1)K+1}, \ldots, \gamma_{(i-1)K+K}$, where the rounding operator $[\bullet]$ is applied to obtain integer values.
(5) Finally, counts of sampled isolates of exogenous strains in site i are obtained by a multinomial draw with size $[\tau_i M]$ and vector of probabilities proportional to Δ, whose element $(j-1)K + k$, $k = 1, \ldots, K$, is equal to zero if $j = i$, and equal to $\gamma_{(j-1)K+k} C_{j \to i}$ otherwise:
(6) $$\Delta_{(j-1)K+k} = \begin{cases} 0 & if \quad j = i \\ \gamma_{(j-1)K+k} C_{j \to i} & if \quad j \neq i \end{cases}$$

Hence, the probability that an endemic strain of j is sampled in i depends on the relative importance of this strain in j and the connectivity from j to i.

Using the empirical simulation model described above, we made several series of 100 simulations with varying values of the number of sites I, the number of isolates per site M and the maximum exogenous fraction of pathogen strains at the site level τ_{max}. Then, we carried out two comparisons:

(1) We compared the situation where $(I, M) = (4, 250)$ (i.e., a few sites with large samples) and the situation where $(I, M) = (10, 100)$ (i.e., a larger number of sites with smaller samples); 1,000 isolates were sampled in both situations. For this comparison we set $\tau_{max} = 0.2$ (i.e., the maximum proportion of exogenous strains at the site level is 20%) and $K = 10$ (i.e., each site has 10 endemic strains).
(2) Then, we studied the case where the maximum proportion of exogenous strains at the site level is low, namely 2% ($\tau_{max} = 0.02$) and compared two different sampling efforts: $(I, M) = (10, 100)$ and $(I, M) = (10, 1000)$. For this comparison, $K = 10$.

RESULTS

Genetic Characteristics of Airborne *S. sclerotiorum* Strains

All the 16 markers used to characterize the 498 strains were polymorphic with a number of alleles ranging from 2 (for locus 36-4) to 21 (for locus 106-4). Strains collected from the Center-West region showed the highest gene diversity and allelic richness (**Table 1**). A total of 241 different multilocus haplotypes (MLH) were identified, of which 172 were represented by only one strain. Our analyses of dissemination and connectivity were based on the 69 MLH that were represented by at least 2 strains. The most frequent MLH was represented by 20 strains, all coming from the South-West. Furthermore, 18 of the MLH represented by multiple strains were collected from more than one region (**Figure 2**). The geographic distance between the collection sites of strains sharing the same haplotype was as far as 700 km.

Hierarchical AMOVA revealed that 89.2% of the genetic variability was distributed within sampling years, compared to 1.6% within a region and 9.2% among regions. These three factors contributed significantly to total genetic variance ($P < 0.05$).

There was no evidence of genetic differentiation between *S. sclerotiorum* strains with the R_{ST} fixation index statistic not significant ($P > 0.05$) for pairwise combination SW-CW, SW-NW, CW-NW and NW-N (clonally-corrected data). There was evidence of low genetic differentiation between *S. sclerotiorum* strains with the R_{ST} highly significant ($P < 0.01$) for pairwise combination North/Center-West (R_{ST} = 0.094 and 0.064 with clonally corrected data) and North/South-West (R_{ST} = 0.152 and

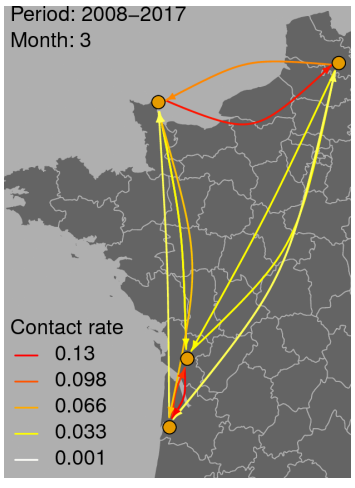


FIGURE 3 | Tropospheric connectivity in March that was inferred with HYSPLIT from archived meteorological data over the period 2008–2017.

0.099 with clonally corrected data). There was also a low genetic differentiation (value < 0.10) between *S. sclerotiorum* strains from the North and the North-West (R_{ST} = 0.061, P = 0.03) for the non-corrected dataset.

Tropospheric Connectivity

Figure 3 shows the graph of tropospheric connectivity in March that was inferred with HYSPLIT from archived meteorological data over the period 2008–2017. Our approach provides an assessment of the directional links between the sites where *S. sclerotiorum* has been sampled. Thus, we obtained, for each month of the year, a network with 4 nodes and 12 edges with varying intensities (**Supplementary Figures S1–S11** are graphs for months other than March). Directionality in the links can be clearly seen on these graphs. **Figure 4** gives a clearer assessment of the variation in the connectivity with time. Despite temporal variation in the directional connectivity between two given sites, we observe stability with respect to the most connected sites

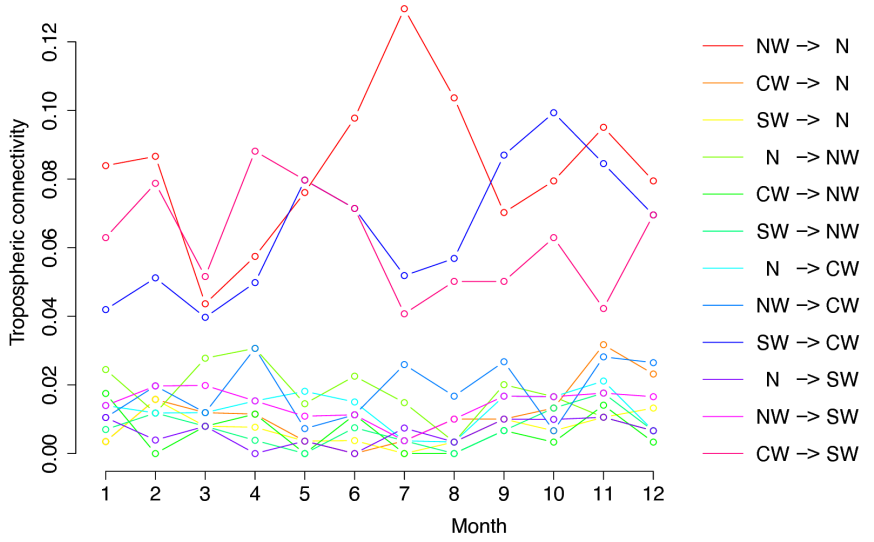


FIGURE 4 | Temporal variation in the tropospheric connectivity between sites.

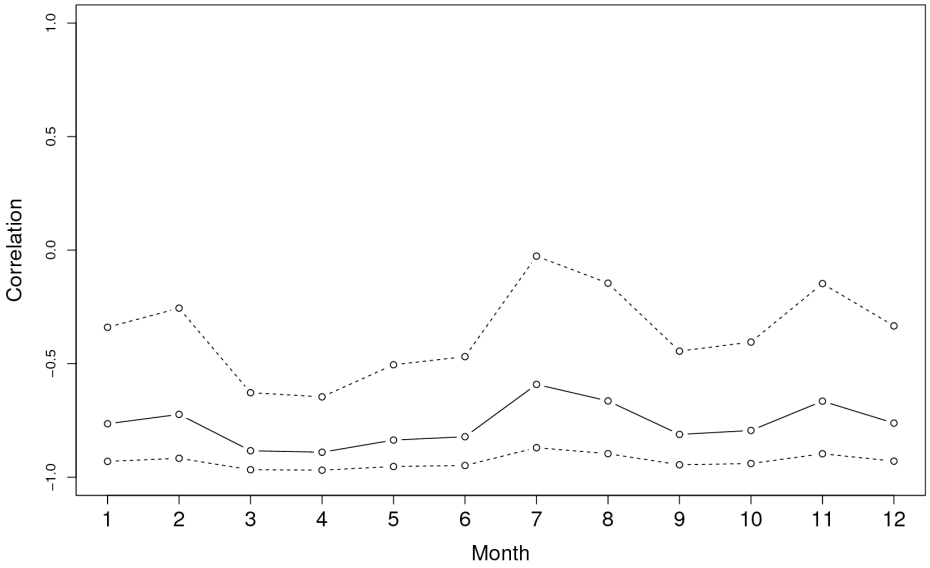


FIGURE 5 | Correlation across time between the tropospheric connectivity and the geographic distance.

(three pairs of sites with directional connectivity larger than 0.04) and the less connected sites (nine pairs of sites with directional connectivity lower than 0.03). The maximum connectivity is about 0.13 between NW and N sites in July (i.e., 13% of air mass trajectories arriving in the N site in July go through the buffer zone around site NW). The connectivity NW→N is the largest for 7 months of the year but the connectivity SW→CW and CW→SW are also occasionally the highest ones. The correlation across time between the tropospheric connectivity and the Earth surface distance is displayed in **Figure 5**. This correlation around -0.75 indicates that the two variables are connected but additional information is included in the tropospheric connectivity, in particular the directionality.

Link Between PIC and Tropospheric Connectivity

Figure 6 provides the result of the log-linear regression between the PIC and the tropospheric connectivity. The regression lines are rather stable across time, reflecting the relative stability of the tropospheric connectivity. They indicate a positive trend in the link between the PIC and tropospheric connectivity, the test of "slope = 0" leading to low *p*-values for June, September, and October. However, the small number of sites does not allow us to definitely state that this link is clearly founded. The regressions could be influenced by extreme points. As a benchmark, we performed the same analysis between the PIC and the Earth surface distance (**Figure 7**), which are expected to be negatively correlated. The limitation due to the number of sites also applies here, but the link seems weaker in **Figure 7** than in **Figure 6**, suggesting that tropospheric connectivity is more appropriate to explain the probability of the incoming microbial component.

Simulation-Based Assessment of Different Sampling Schemes

First, we compared the sampling schemes characterized by $(I, M) = (4, 250)$ and $(I, M) = (10, 100)$, when the maximum proportion of exogenous strains at the site level is rather high, i.e., $\tau_{max} = 0.2$. **Figure 8** shows that the sampling scheme with more sites but less isolates per site leads more systematically to a positive estimate of the slope coefficient in the regression between the connectivity and the log-PIC. Using 10 sites instead of 4 allows, for each repetition, to have $10 \times 9/2 = 45$ points instead of $4 \times 3/2 = 6$ points to fit the regression. Obviously, there is a trade-off between the number of sites and the number of samples per site because a weak number of samples per site could lead to inaccurate estimates of PIC values and, therefore, inaccurate

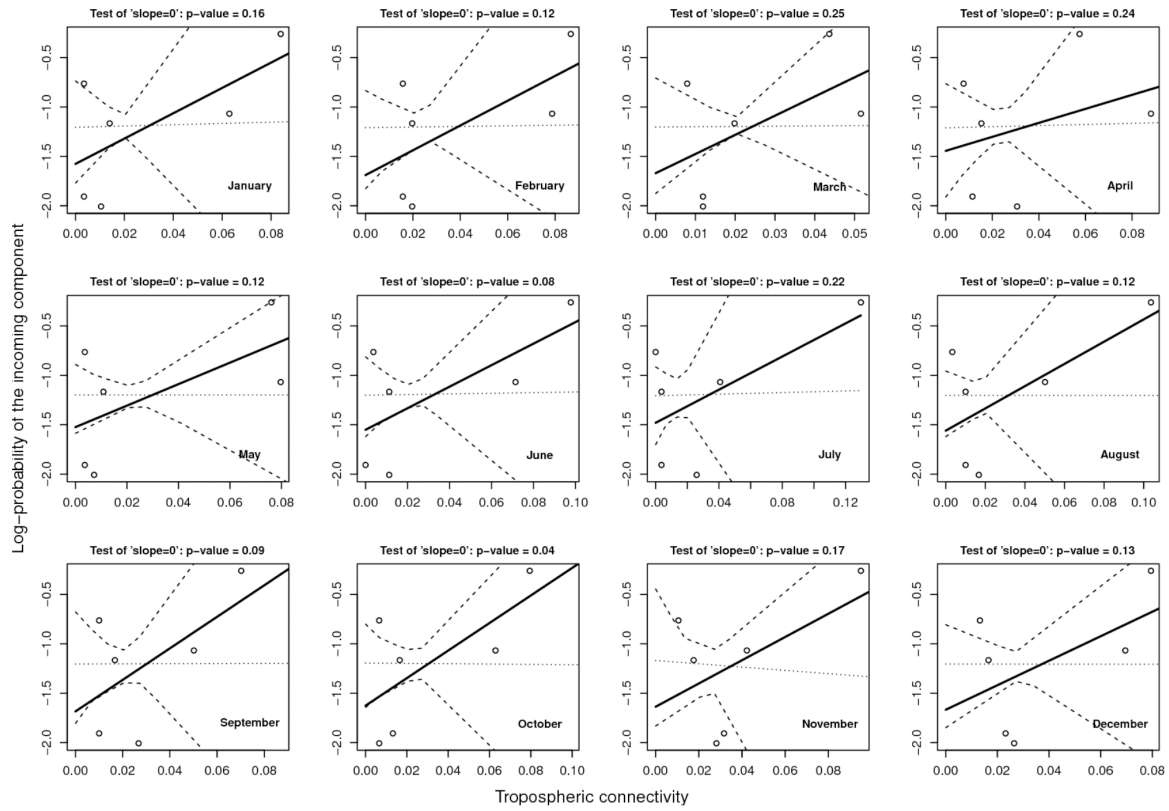


FIGURE 6 | Log-linear regression between PIC and tropospheric connectivity.

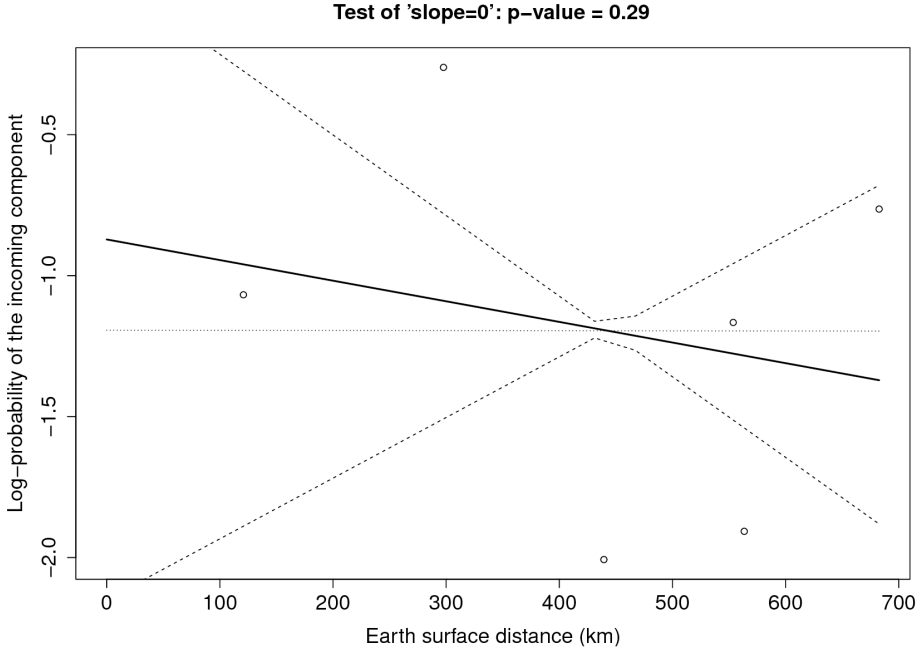


FIGURE 7 | Log-linear regression between the PIC and the Earth surface distance.

estimates of the slope. In addition, from a computational point of view, calculation time increases with the square of the number of sites.

Second, we compared the sampling schemes characterized by $(I, M) = (10, 100)$ and $(I, M) = (10, 1000)$, when the maximum proportion of exogenous strains at the site level is rather low, i.e., $\tau_{max} = 0.2$. **Figure 9** shows that using $(I, M) = (10, 100)$ with a 2% maximum proportion of exogenous strains at the site level is not enough to infer the link between connectivity and PIC. However, the power of the analysis tool significantly increases with the number of sampled isolates per site.

DISCUSSION

Variants of pathogens that migrate from one place to another can reduce the efficiency of strategies to manage plant health if the incoming variants are resistant to or less sensitive to the methods used at the site where they arrive. Furthermore, identification of the origin of incoming inoculum could help in predicting the arrival of inoculum and in reducing, where possible, the quantity of inoculum available from distal sources. For pathogens with a wide range of possible sources, the incoming component of inoculum cannot be monitored in real time in a given site. Therefore, we looked for indicators that were significantly linked to the probability of incoming airborne inoculum. In particular we developed a method to evaluate connectivity between distant sites via the atmosphere and assessed the relationship between aerial connectivity and incoming inoculum. This was based on data for daily air mass trajectories from which the airborne connectivity between sites and the prevailing directions of connectivity were inferred.

The results show that the rate of aerial connectivity between two sites varies according to the direction and to the month considered. For example, it was shown that during the period from 2008 to 2017, North and North–West regions were significantly connected but more strongly in the North-West toward North direction than the inverse. On the contrary, South–West and Center–West had high rates of connectivity but in both directions. Finally some regions had low rates of aerial connectivity (e.g., South–West toward North and North-West). This directional connectivity corresponds to air mass movements that change with the seasons – a notion that is obvious to anyone who is attentive to the weather as seasons change. The originality of this work is that we identified the links between specific sites that were in the trajectories of the air masses and we quantified the rates at which the different directions of the links occurred.

The results also show that aerial connectivity is more informative than geographic distance between two sites to explain the PIC of inoculum at a given site from another one. Aerial connectivity adds information about the direction of potential exchanges via the atmosphere that geographic distance cannot provide. This information is particularly interesting since it makes it possible to identify more precisely the role of each site as a source or sink of spores. It can therefore be hypothesized that isolates with identical haplotypes in two regions are more likely to have been emitted by the one having aerial connectivity directed toward the other (such as NW toward N for example).

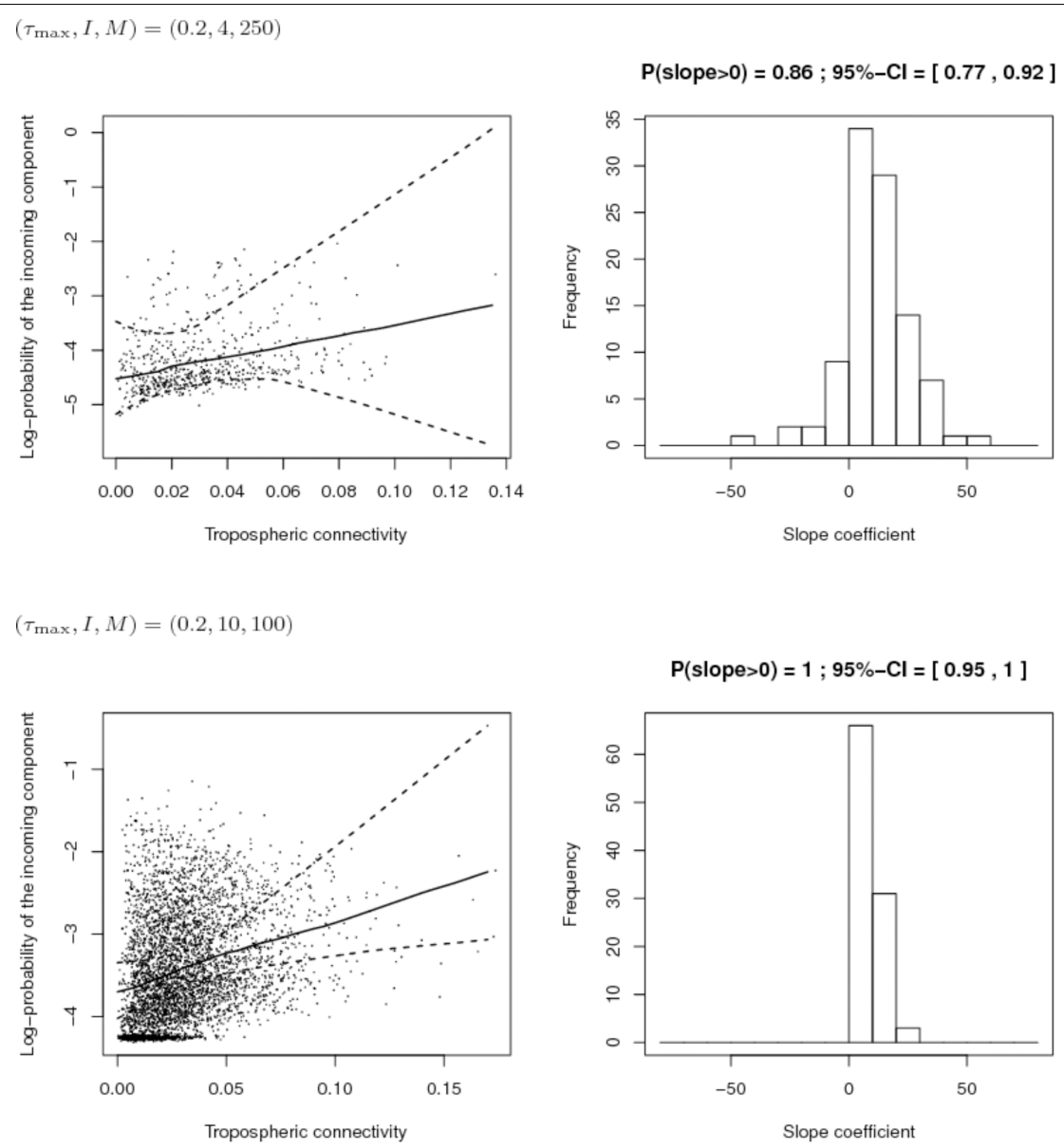


FIGURE 8 | Comparison of the inferred links between connectivity and probability of the incoming component (PIC) for different sampling schemes (top: 4 sites and 250 isolates per site; bottom: 10 sites and 100 isolates per site) when the maximum proportion of exogenous strains at the site level is 20%. Left: Cloud of points "connectivity × PIC" obtained for the 100 simulations and pointwise median and quantiles of order 0.025 and 0.975 of the linear regression line linking connectivity and PIC. Right: histogram of the estimate of the slope coefficient in the regression (a positive slope tends to show a positive link between connectivity and PIC). The proportion of positive estimated slopes and its 95%-confidence interval are given above the histogram.

The fact that aerial connectivity is more informative than geographic distance may be due to the movements of air masses between two sites that are not simply linked to linear distance they travel but also to topography, climate and the global movement of air masses on the terrestrial surface. Taking the example of NW and N regions of the present study: aerial connectivity reflects the general movements of air masses that often enter the French territories by the Atlantic coast and that move inland. In contrast, geographic distance (about 300 km) doesn't give any information about frequency and direction of aerial exchanges between NW and N. In order for connectivity to be as informative as possible, we calculated it over several years to

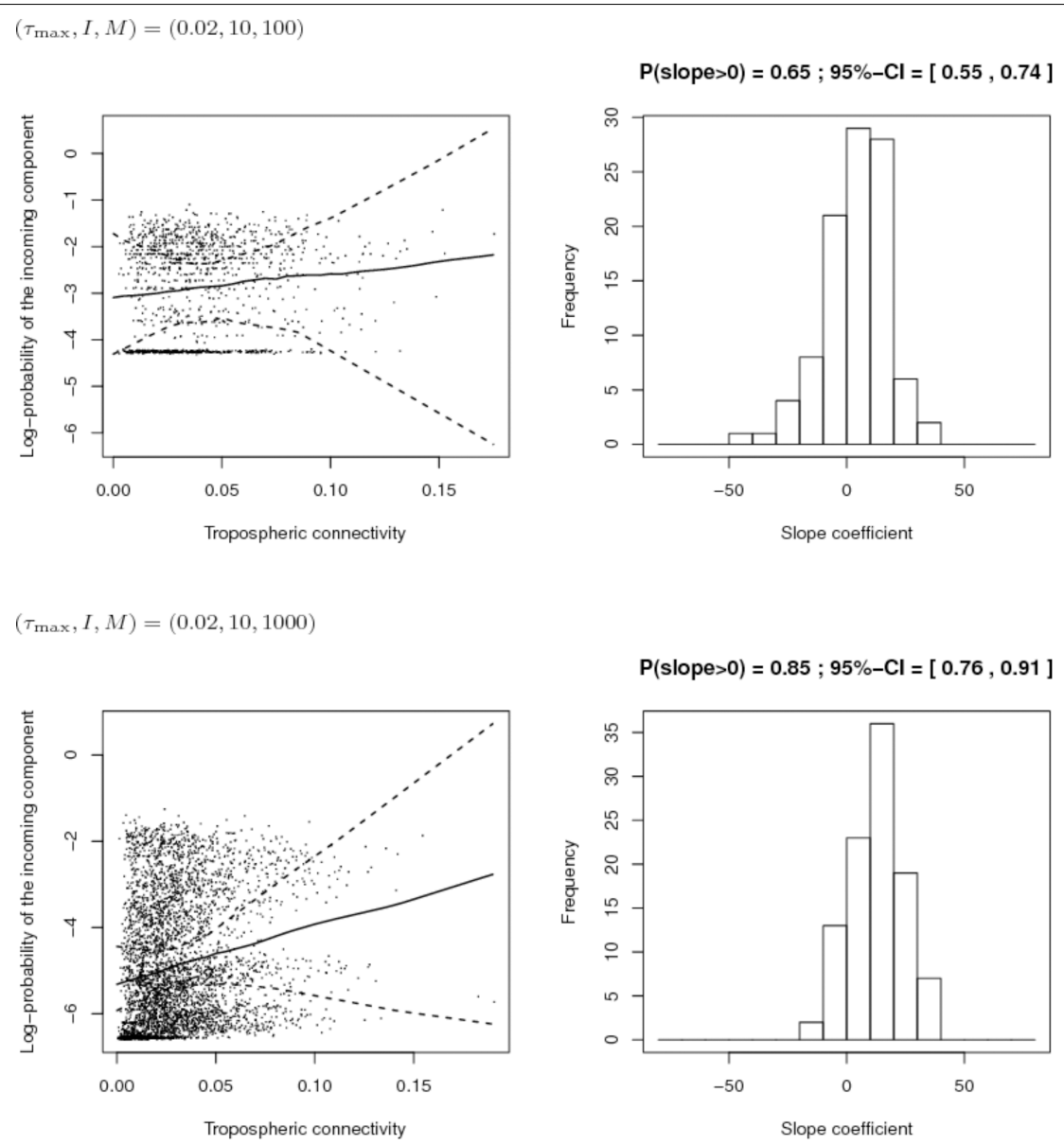


FIGURE 9 | Comparison of the inferred links between connectivity and probability of the incoming component (PIC) for different sampling efforts (top: 10 sites and 100 isolates per site; bottom: 10 sites and 1,000 isolates per site) when the maximum proportion of exogenous strains at the site level is 2%. Left: Cloud of points "connectivity × PIC" obtained for the 100 simulations and pointwise median and quantiles of order 0.025 and 0.975 of the linear regression line linking connectivity and PIC. Right: histogram of the estimate of the slope coefficient in the regression (a positive slope tends to show a positive link between connectivity and PIC). The proportion of positive estimated slopes and its 95%-confidence interval are given above the histogram.

provide a global and robust vision of the prevailing links between two sites.

Studying and predicting airborne routes of plant pathogens is of general interest because atmospheric highways have very few boundaries. Recently, key airborne dispersal routes of rust between several African and Asian countries were studied by Meyer et al. (2017) using a mechanistic modeling framework. In their study, data about heterogeneous landscapes of wheat fields, turbulent atmospheric spore dispersal and environmental suitability for infection after deposition were computed. This model is well adapted to a host specific pathogen like rust but not for broad-host range fungi because sources cannot be easily

identified. The method developed in the present study bypasses this pitfall by allowing us to assess aerial interconnectivity between sites that do not necessarily harbor crops.

The present study describes a method in its preliminary state. We used simulations to investigate how to improve this method and in particular how to optimize the balance between the number of sites considered and the number of strains collected at each site. These simulations revealed that to determine links between sites it is better to have more sites with few strains than few sites with many strains. However, a balance has to be found because an insufficient number of strains per site could lead to an inaccurate estimates of PIC values. This balance has to be considered according to the practical aspects of sampling (number of samplers to deploy, number of isolates to isolate and to genotype. . .). These simulations could be performed to establish the experimental design for future studies while taking into account the number of potential sampling sites and the effort needed to isolate and characterize strains of the targeted pathogen.

Moreover, the way the PIC was computed could be tuned to reflect more realistic processes. In this study, the PIC was obtained by introducing a filter f_i that implies that all strains in j have an equal propensity to be dispersed from j to i and an equal propensity to establish in i. To weaken this assumption, we tested an alternative specification for the filter that incorporates the possibility for some strains of j to not be dispersed or not be adapted to site i: the filter, denoted by $f_{j\to i}$, is a vector of dimension n whose component s is equal to 1 if strain s was collected both in site j and i; otherwise it is 0. With such a specification, we, however, assume that all the strains in j that can be dispersed to and establish in i have an equal propensity to make it to i. The use of this alternative filter led to different PIC values in the real-case study tackled in this article, but did not yield qualitatively different results. Further investigation should be carried out to consider more relevant filters and obtain more accurate inferences about the long-distance dispersal of pathogens.

CONCLUSION

Knowledge about connectivity (its rate and its direction between sites) at a regional and global scale is useful for establishing strategies to set up networks of epidemio-surveillance for wind-dispersed fungi. Practically, spore monitoring devices could be placed at existing sites of meteorological networks, for example, or on elevated buildings (i.e., lighthouses for example), or at stations for measurement of flux of atmospheric particles[1]. Placing trapping devices upstream of the North–West site described in this study, for example, along the French oceanic border could help to identify spores coming from outside the French territory. Information about airborne inoculum potentially arriving in a territory is important to obtain since the traits of exogenous strains may differ from endogenous ones and sometimes may lead to more damaging epidemics. For example, countries where the inoculum originated could have policies concerning the use of fungicides that are different from French policies. In this way, incoming fungal strains may have developed resistance to fungicides that have not yet been deployed in France or that are reserved for use in cases of extreme potential for crop damage. If such strains manage to infect and multiply in France, they may lead to the modification of the endogenous fungal population thereby rendering obsolete the use of certain fungicides. Different cropping systems or climates may also result in more aggressive or more robust strains than endogenous ones. Finally if an incoming fungal population of a given species is more diverse than the local endogenous population, the choice of cultivars might need to be modified otherwise resistance may be overcome by the first arrival of exogenous airborne inoculum. The tools and analytical methods that we have presented here can be further developed to address the impact of aerial inoculum from long distance sources and the capacity to account for this in strategies of crop protection.

[1] https://daac.ornl.gov/FLUXNET/guides/Fluxnet_site_DB.html

AUTHOR CONTRIBUTIONS

CL carried out microsatellite genotyping and genetic data analysis. CM and SS developed the initial idea of tropospheric interconnectivity. SS and MC developed and conducted the calculations concerning tropospheric interconnectivity and pathogen composition decomposition. CL coordinated the writing of the manuscript. CL, CM, MC, and SS wrote the manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported in part by a CASDAR grant from the French Ministry of Agriculture together with the Scientific Interest Group "GIS PICLég" ("ScleroLeg" project - https://www.picleg.fr/Les-Projets-en-cours/Scleroleg). The salary of MC was provided by the SPREE project from the French National Research Agency (Contract No. ANR-17-CE32-0004-01). Part of this work was carried out by using the resources of the Molecular Biology Platform of INRA-PACA center.

ACKNOWLEDGMENTS

We thank especially M. Duffaud, C. Troulet, and F. Villeneuve for their contribution to the fungal strain collection. We also thank all the field experimenters of the technical institutes (CTIFL, ACPEL, APEF, CEFEL, INVENIO, SILEBAN, and UNILET) who collaborated on the project and who collected the strains used in the present study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02257/full#supplementary-material

REFERENCES

Arnaud-Haond, S., and Belkhir, K. (2007). GENCLONE: a computer program to analyse genotypic data, test for clonality and describe spatial clonal organization. *Mol. Ecol. Notes* 7, 15–17. doi: 10.1111/j.1471-8286.2006.01522.x

Arnaud-Haond, S., Duarte, C. M., Alberto, F., and Serrao, E. A. (2007). Standardizing methods to address clonality in population studies. *Mol. Ecol.* 16, 5115–5139. doi: 10.1111/j.1365-294X.2007.03535.x

Aylor, D. E., Taylor, G. S., and Raynor, G. S. (1982). Long-range transport of tobacco blue mold spores. *Agric. Meteorol.* 27, 217–232. doi: 10.1016/0002-1571(82)90007-3

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., and Bonhomme, F. (1996–2004). *Genetix 4.05, Logiciel Sous Windows TM Pour la Génétique Des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171*. Montpellier: Université de Montpellier.

Boland, G. J., and Hall, R. (1994). Index of plant host of *Sclerotinia sclerotiorum*. *Can. J. Plant Pathol.* 16, 93–108. doi: 10.1080/07060669409500766

Bowden, J., Gregory, P. H., and Johnson, C. G. (1971). Possible wind transport of coffee leaf rust across Atlantic Ocean. *Nature* 229, 500–501. doi: 10.1038/229500b0

Brown, J. K., and Hovmoller, M. S. (2002). Epidemiology-aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. *Science* 297, 537–541. doi: 10.1126/science.1072678

Damialis, A., Kaimakamis, E., Konoglou, M., Akritidis, I., Traidl-Hoffmann, C., and Gioulekas, D. (2017). Estimating the abundance of airborne pollen and fungal spores at variable elevations using an aircraft: how high can they fly? *Sci. Rep.* 7:44535. doi: 10.1038/srep44535

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform.* 1, 47–50. doi: 10.1177/117693430500100003

Golan, J. J., and Pringle, A. (2017). Long-distance dispersal of fungi. *Microbiol. Spectr.* 5. doi: 10.1128/microbiolspec.FUNK-0047-2016

Goudet, J. (1995). FSTAT (Version 1.2): a computer program to calculate F-statistics. *J. Heredity* 86, 485–486. doi: 10.1093/oxfordjournals.jhered.a111627

Hirst, J. M., Stedman, O. J., and Hurst, G. W. (1967). Long-distance spore transport: vertical sections of spore clouds over sea. *J. Gen. Microbiol.* 48, 357–377. doi: 10.1099/00221287-48-3-357

Hovmoller, M. S., Justesen, A. F., and Brown, J. K. (2002). Clonality and long-distance migration of *Puccinia striiformis* f.sp tritici in north-west Europe. *Plant Pathol.* 51, 24–32. doi: 10.1046/j.1365-3059.2002.00652.x

Isard, S. A., Gage, S. H., Comtois, P., and Russo, J. M. (2005). Principles of the atmospheric pathway for invasive species applied to soybean rust. *Bioscience* 55, 851–861. doi: 10.1641/0006-3568(2005)055[0851:POTAPF]2.0.CO;2

Lehner, M. S., Paula, T. J., and Mizubuti, E. S. G. (2016). Does hyphal-tip ensure the same allelic composition at SSR loci as monosporic isolates of *Sclerotinia sclerotiorum*? *J. Phytopathol.* 164, 417–420. doi: 10.1111/jph.12429

Leyronas, C., Bardin, M., Berthier, K., Duffaud, M., Troulet, C., Torres, M., et al. (2018). Assessing the phenotypic and genotypic diversity of *Sclerotinia sclerotiorum* in France. *Eur. J. Plant Pathol.* doi: 10.1007/s10658-018-1493-9 doi: 10.1007/s10658-018-1493-9

Maldonado-Ramirez, S. L., Schmale, D. G., Shields, E. J., and Bergstrom, G. C. (2005). The relative abundance of viable spores of *Gibberella zeae* in the planetary boundary layer suggests the role of long-distance transport in regional epidemics of *Fusarium* head blight. *Agric. For. Meteorol.* 132, 20–27. doi: 10.1016/j.agrformet.2005.06.007

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edn. New York, NY: Chapman and Hall.

Meyer, M., Cox, J. A., Hitchings, M. D. T., Burgin, L., Hort, M. C., Hodson, D. P., et al. (2017). Quantifying airborne dispersal routes of pathogens over continents to safeguard global wheat supply. *Nature Plants* 3, 780–786. doi: 10.1038/s41477-017-0017-5

Ojiambo, P., and Holmes, G. (2011). Temporal and spatial spread of cucurbit downy mildew in the eastern United States. *Phytopathology* 101, S131–S131. doi: 10.1094/PHYTO-09-10-0240

Prospero, J. M., Blades, E., Mathison, G., and Naidu, R. (2005). Interhemispheric transport of viable fungi and bacteria from Africa to the Caribbean with soil dust. *Aerobiologia* 21, 1–19. doi: 10.1007/s10453-004-5872-7

Purdy, L. H. (1979). *Sclerotinia sclerotiorum*: history, diseases and symptomatology, host range, geographic distribution, and impact. *Phytopathology* 69, 875–880. doi: 10.1094/Phyto-69-875

Purdy, L. H., Krupa, S. V., and Dean, J. L. (1985). Introduction of sugarcane rust into the Americas and its spread to Florida. *Plant Dis.* 69, 689–693. doi: 10.1094/PD-69-689

Rolph, G., Stein, A., and Stunder, B. (2017). Real-time environmental applications and display system: ready. *Environ. Model. Softw.* 95, 210–228.

Singh, R. P., Hodson, D. P., Huerta-Espino, J., Jin, Y., Bhavani, S., Njau, P., et al. (2011). "The emergence of ug99 races of the stem rust fungus is a threat to world wheat production," in *Annual Review of Phytopathology*, Vol 49, eds N. K. Vanalfen, G. Bruening and J. E. Leach, 465–481.

Singh, R. P., Hodson, D. P., Jin, Y., Lagudah, E. S., Ayliffe, M. A., Bhavani, S., et al. (2015). Emergence and spread of new races of wheat stem rust fungus: continued threat to food security and prospects of genetic control. *Phytopathology* 105, 872–884. doi: 10.1094/PHYTO-01-15-0030-FI

Sirjusingh, C., and Kohn, L. M. (2001). Characterization of microsatellites in the fungal plant pathogen, *Sclerotinia sclerotiorum*. *Mol. Ecol. Notes* 1, 267–269. doi: 10.1046/j.1471-8278.2001.00102.x

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139, 1463–1463.

Smith, D. J., Jaffe, D. A., Birmele, M. N., Griffin, D. W., Schuerger, A. C., Hee, J., and Roberts, M. S. (2012). Free tropospheric transport of microorganisms from Asia to North America. *Microb. Ecol.* 64, 973–985. doi: 10.1007/s00248-012-0088-9

Steadman, J. R., Marcinkowska, J., and Rutledge, S. (1994). A semi-selective medium for isolation of *Sclerotinia sclerotiorum*. *Can. J. Plant Pathol.* 16, 68–70. doi: 10.1080/07060669409500791

Stein, A. F., and Ngan, F. (2015a). Potential use of transport and dispersion model ensembles for forecasting applications. *Weather Forecast.* 30, 639–655. doi: 10.1175/WAF-D-14-00153.1

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F. (2015b). NOAA'S HYSPLIT atmospheric transport and dispersion modeling system. *Bull. Am. Meteorol. Soc.* 96, 2059–2077. doi: 10.1175/BAMS-D-14-00110.1

Tao, Z. N., Malvick, D., Claybrooke, R., Floyd, C., Bernacchi, C. J., Spoden, G., et al. (2009). Predicting the risk of soybean rust in Minnesota based on an integrated atmospheric model. *Int. J. Biometeorol.* 53, 509–521. doi: 10.1007/s00484-009-0239-y

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.