SAPIENZA
UNIVERSITÀ DI ROMA

Innovative approaches in spatio-temporal modeling: handling data collected by new technologies

School of Statistical Sciences

Ph.D. program in "Methodological Statistics", XXXIII cycle

Candidate

Pierfrancesco Alaimo Di Loro
ID number 1538953

Thesis Advisors

Prof. Giovanna Jona Lasinio
Prof. Luca Tardella

External Advisor

Prof. Sudipto Banerjee

June 2021

Thesis defended on 13 July 2021
in front of a Board of Examiners composed by:

Prof. Alessandra Luati (chairman)

Prof. Claudio Agostinelli

Prof. Bruno Scarpa

Innovative approaches in spatio-temporal modeling: handling data collected by new technologies

Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: June 2021

Author's email: pierfrancesco.alaimodiloro@uniroma1.it
pierfrancesco.alaimodiloro@gmail.com

To my family, who lighted up my path.
To Giovanna, constant support during my walk.

Abstract

This thesis illustrates and puts in context two of the main research projects I worked on during my Ph.D. program, in collaboration with several national and international co-authors from "La Sapienza" and other prestigious universities. Both research lines concern *spatial and spatio-temporal analysis* of geo-referenced datasets, which is of broad and current interest in the statistical research literature and applications. My focus on such an area of statistics was not meditated before the start of the program. However, while pursuing my original research interests in the broader domain of Bayesian statistics, I realized there was an ever-increasing demand for viable and efficient statistical methods to analyze spatial and spatio-temporal data. That is a consequence of the extraordinary technological development that interested data collection systems during the last few decades. The innovative, cutting-edge technologies conceive new devices that can record and store data and information about the most diverse phenomena, possibly at a fine spatial scale and with high temporal resolution. Such capabilities were just a dream up to 20 or 30 years ago. Spatial statistics methods are rapidly evolving to face this surge of novel data structure in various application fields: geology, meteorology, ecology, epidemiology, economics, politics, and more.

The first chapter of this thesis introduces the general idea behind *spatial statistics*, that is the branch of statistics devoted to analyzing and modeling temporal and spatial structure in time and/or geo-referenced datasets. A brief historical introduction of its developments is provided, starting from the first (sometimes unwitting) applications of its logic to practical and theoretical problems at the end of the XIX century. Many methods and techniques in this domain evolved independently, driven by the specific needs of the application fields in which they were developed. The historical *ex-cursus* leads to a coarse (but reasonable) distinction in three main areas: *continuous spatial variations*, *discrete spatial variations*, *spatial point patterns*. These areas present further facets within themselves, making spatial statistics an incredibly diverse and rich topic. A really comprehensive review would require an entire book to be written and maybe a lifetime to be thoroughly studied. Therefore, in the following Chapters, the discussion is focused on specific areas and techniques used in the studies. Only those tools that proved valuable for the analysis performed in Alaimo Di Loro et al. (2021) and Kalair et al. (2020) are extensively treated.

The second chapter focuses on analyzing continuous spatial variation, which is the modeling of outcomes varying continuously over some space. First, the most relevant properties for continuous spatial processes are introduced; second, some of the most common methodologies for performing spatial interpolation of the mean trend and stochastic modeling of the residuals are listed and sketched. In particular, the chapter digresses on *Spline Regression* as a valid technique to catch the first-order structure in spatial data. Soon after, the *Geo-Statistical* methods and the *Bayesian Hierarchical* framework are claimed as invaluable tools to attain the simultaneous estimation of the first and second-order structure of a process. Extension to spatio-temporal contexts is not as trivial as it may seem but must be approached with due care. An extensive discussion about the possible pitfalls and viable solutions is included in the same chapter. Finally, the problems arising in the analysis of *Big spatial data* are highlighted in the last section, where The *Nearest Neighbor Gaussian Process* (NNGP, Datta et al. (2016a,b)) model is introduced as a highly scalable framework for providing full inference on massive spatial and spatio-temporal datasets.

The third chapter includes an extended version of the paper Alaimo Di Loro et al.

(2021), currently under-review and published as a pre-print. It describes how the aforementioned technological development has strongly affected human tracking and monitoring capabilities, generating substantial interest in monitoring human activity. New non-intrusive wearable devices, such as wrist-worn sensors that monitor gross motor activity (miniature accelerometers), can continuously record individual activity levels, producing massive amounts of high-resolution measurements. Analyzing such data needs to account for spatial and temporal information on trajectories or paths traversed by subjects wearing such devices. Inferential objectives include estimating a subject's physical activity levels along a given trajectory, identifying trajectories that are more likely to produce higher levels of physical activity for a given subject, and predicting expected levels of physical activity in any proposed new trajectory for a given set of health attributes. We argue that the underlying process is more appropriately modeled as a stochastic evolution through time while accounting for spatial information separately. Building upon recent developments in this field, we construct temporal processes using directed acyclic graphs (DAG) on the line of the NNGP, include spatial dependence through penalized spline regression, and develop optimized implementations of the collapsed Markov chain Monte Carlo (MCMC) algorithm. The resulting Bayesian hierarchical modeling framework for the analysis of spatial-temporal actigraphy data proves able to deliver fully model-based inference on trajectories while accounting for subject-level health attributes and spatial-temporal dependencies. We undertake a comprehensive analysis of an original dataset from the Physical Activity through Sustainable Transport Approaches in Los Angeles (PASTA-LA) study to formally ascertain spatial zones and trajectories exhibiting significantly higher physical activity levels. Suggestions for further extensions and improvements on the currently adopted methodology are discussed in the last section of the chapter.

Chapter four undergoes a paradigm shift and introduces the basic theory and tools of spatial point patterns analysis. Some common probabilistic models for point processes are briefly discussed, with some of their properties and limitations highlighted. The rest of the chapter is instead entirely focused on the Hawkes process and its spatio-temporal extension. It is a particular kind of self-exciting point process that presents a strong inter-dependence structure. While conceived in Hawkes (1971a), its use in the statistical application has been for a long time limited to the analysis of earthquakes dynamic. The recent escalation of data at the high temporal resolution, sometimes accompanied by spatial information, has favored its use in modeling events dynamics in diverse fields: finance, society, biology, etc. In particular, its defining properties are presented and state-of-the-art estimation methods of the spatio-temporal version are introduced.

In the fifth chapter, the semi-parametric Hawkes process with a periodic background originally introduced in Zhuang and Mateu (2019) is outlined. While very recent, it has already revealed itself very useful to model phenomena that are likely to present a cyclic pattern. It assumes that primary events occur as an effect of the background intensity, while secondary events are associated with the self-excitation effect. There are sound motivations that justify its utilization in the context of road accident dynamics, e.g.: excitation may occur when a driver, reacting to the disruption of one accident, triggers a subsequent accident upstream of the first one. The proposed framework is tested on two original applications on two original sets of data: the first one, somewhat preliminary, involves the modeling and analysis of road accidents that occurred on the urban road network of Rome, in Italy; the second is instead a conclusive analysis recently published in (Kalair et al., 2020), conducted on a collection of road accidents occurred on the M25 London Orbital, in the United Kingdom. Adaptations of the original methodology to the road accident setting were

deemed necessary in both cases to consider specific features of car accidents and the geometry of the underlying space. The final results permit a fruitful interpretation of the temporal and spatial background that detects the typical commuting behavior in the Roman and Londoners communities. The self-excitation component appears to have slightly different intensities in the two contexts, suggesting excitation mechanisms that vary between urban networks and motorways.

Finally, the sixth chapter summarizes all the main passages in the thesis, highlighting the previous chapters' original contributions. It also tries to summarize a take-home message about statistical modeling's fundamental importance as a scientific tool to formulate and verify hypotheses that must not be discouraged by new challenges and technological advancements.

Contents

Chapter 1

Motivation and introduction

> Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data
>
> Stanford Encyclopedia of Philosophy

The above definition does not put statistics among other sciences but defines it with a slightly different term: discipline. Generally speaking, science can be defined as "*the study of the world which surround us, pursued for the sake of knowledge and development*" and, if we agree with such definition, statistics cannot be a science by itself. Indeed, it is a discipline that *serves* science. It is the language of information, aimed at extracting relevant features from the messy and noisy data collected during experimental observations. Statistics is in an ambiguous position with respect to science. It does not pursue progress and development for its own sake, but it is powered by scientific and technological progress momentum, which it favors in its turn. As a matter of fact, advancements in the statistical field are driven by the scientific and technological evolution in other fields, which pose statisticians in front of new challenges and obstacles to tackle mainly for the sake of the specific problem at hand. This is, somehow, the luck and disgrace of statistics: it will persist and evolve with sciences, but it is doomed to adapt to other sciences' needs, having only partial control over its development directions.

We may say that all the statistical problems and questions arise from data: their nature, their structure, and their complexity. Therefore, if the data collection tools evolve, statistical methods must evolve in order to keep up the pace with such changes. Recent decades have seen a rapid technological development of tracking and monitoring devices (i.e. smartwatch, smartphones), remote sensing technologies (i.e. satellite measurements), and geographic information systems (i.e. GPS, GLONASS). This caused a surge of real-time data of all sorts, recorded at high frequency over large time windows and usually accompanied by geographical information. With great information come great responsibilities, and statistical modeling has an obligation to exploit this massive amount of available information and account for the proper evolution of phenomena over space and time. The study of such data is the subject of **spatial statistics**, which tries indeed to take into account temporal and spatial structure in the observed patterns (and verify its presence, whenever possible). However, while spatial statistics is not a recent topic, the high-cost of collecting temporal and spatial data in older times made it a niche sector dedicated to small ad-hoc studies in the environmental and biological fields.

Methods developed in those contexts were particular to their application sectors. They were never thought to deal with the uncontrolled form and large size of the currently available spatio-temporal data-sets.

1.1 Spatial statistics

Spatial Statistics is the group of statistical tools and techniques devoted to the study of the spatial distribution of the data in order to get more insights about their structural behavior. The spatial or spatio-temporal location may either be of intrinsic interest or just contribute directly to the definition of a suitable stochastic model for the phenomenon under analysis. Eventually, it provides a better understanding and explanation of the data generative process and nature and it aims at improving the prediction accuracy of the outcome of interest at unobserved locations.

From a practical point of view, spatial modeling considers any phenomena evolving over a space as the realization of a stochastic process in the same space. It assumes that the process components are marginally dependent either because of the mutual influence of one on each other and/or because of how the location itself affects the outcome. The literature related to spatial methods is relatively recent but considerably vast, being every possible spatial application of its own kind according to its requirements and peculiarities. For instance, a different type of outcome and/or spatial structure may require utterly different modeling tools. There are also multiple kinds of different dependence structures that inevitably affect the underlying probabilistic assumptions. The challenge of building a meaningful model for all the spatial contexts, also including probabilistically valid and statistically manageable dependence structures, has been tackled from different points of view and produced a great variety of approaches, with mixed fortunes. An extensive review of methods belonging to different perspectives can be found in Gelfand et al. (2010), but other complete and approach-specific accounting of spatial statistics are available in Wahba (1990); Stein (2012); Lawson (2013); Banerjee et al. (2014); Cressie (1993); Cressie and Wikle (2015). When talking about spatial statistics one is naturally inclined to think of the canonical bi-dimensional sub-spaces of $\mathbb{R}^2$. However, computation difficulties aside, theory and methods are usually introduced and stand valid for any euclidean subspace $\mathcal{D} \in \mathbb{R}^d$, with $d \in \mathbb{N}$. From a purely geometric and mathematical perspective, this may also include a time domain. Indeed, the temporal dimensions is a just a peculiar case of uni-dimensional space $\mathcal{T} \in \mathbb{R}$ characterized by an ordered structure (i.e. there is a *before* and an *after*). While the study of phenomena evolving along time and their modeling through temporal processes has its roots in the older and well-established literature of time series and signal processing (Kay (1993), Brillinger (2001)), they can also be considered in the context of *spatial statistics*, which insert those in a general and consistent framework. From a phenomenological and hence modeling point of view, specific care must be taken when dealing with time together with space.

In the sequel, we follow Gelfand et al. (2010) and provide a historical ex-cursus of the milestones that led to the formation of *Spatial Statistics*.

1.1.1 A historical introduction

One of the earliest attempts to keep into account the structure of the space on which a phenomenon is taking place can be dated back to the early geometrical probability literature. The most famous example of this kind is probably the *Buffon's Needle*, named after the Comte de Buffon (1707-1788). The problem supposes that

a needle of length x is thrown at random over a table marked by parallel lines at distance $d > x$. We would like to derive the probability that the needle crosses one of the parallel lines as a function of d. In order to provide an answer to this question, we shall first define what we mean by *at random*. One intuitive definition would be that a pre-specified point of the needle (e.g. the needle center) falls at a random uniformly distributed distance on $(0, d/2)$ from the closest parallel line, and the acute angle formed by the needle and the same line should be as well uniformly distributed but on $(0, \pi/2)$. However, this approach ignores what happens near the edges of the table, whilst remaining valid only for an infinite size surface. A proper spatial approach would require appropriate correction for the table geometry and, if sticking to the infinite size assumption, shall provide a valid justification and/or physical meaning to the experiment. The mathematical resolution to all facets of this problem lies in the theory of *Point Processes*, an extensive review of which can be found in Daley and Vere-Jones (2003) and Daley and Vere-Jones (2007). It provides stochastic models for observed point patterns (i.e. locations of events randomly distributed over space) and the necessity to take into account the spatial structure of the underlying space is intrinsic to it. An example of point process model is the Poisson process (perhaps the most common and widely used), originally defined along time, and later extended to the more general spatial context (dimension $d > 1$). According to Gelfand et al. (2010), the first introduction of such process dates back to 1858, when Rudolf Causius needed to calculate the mean free path of a molecule in a volume of gas in order to defend from criticism the then new molecular theory of heat (*Peter Guttorop, personal communication*). However, the birth of a coherent framework for analyzing statistically spatial patterns had to wait more than one century. It was 1977 when Brian Ripley's paper *"Modeling spatial patterns"*(Ripley, 1977) was published. For more recent reviews of the statistical analysis of spatial point patterns, the reader is referred to Moller and Waagepetersen (2003); Illian et al. (2008); Diggle (2013).

In order to find the first spatial considerations in terms of location-specific effects on the outcome of interest, we need to fast-forward to the twentieth century. Between 1919 and 1933 R.A. Fisher was employed at Rothamsted Experimental station in Hertforshire, England, and was trying to develop an effective and coherent methodology for the analysis of data from agricultural field. In his seminal work (included in the collection Fisher et al. (1960)) he describes how he was looking for a model that could well-represent the data resulting from a uniformity trial conducted at Rothamsted and reported in Mercer and Hall (1911). The experiment consisted of 50 observations referred to the yield of wheat grain from rectangular 3.3×2.59 meters areas, arranged over a rectangular region in 20 rows and 25 columns (see Figure 1.1). Being all the yields from the same sown, variations among the 500 records were presumably due to spatial variations in the soil environment (e.g. soil fertility, slope, etc.). In the first instance, it could sound reasonable to assume that such variations were stochastic by nature. A naive modeling attempt could entail the inclusion of random effects:

$$Y_{ij} = \mu + Z_{ij}, \quad i = 1, \ldots, 20,\ j = 1, \ldots, 25,$$

where μ is the hypothetical population mean yield, i and j denote rows and columns, and the Z_{ij} are independent, zero-mean random perturbations. However, from Figure 1.1, it is evident how near-neighboring rectangles are likely to give similar yields and thus show what we would today define as a *dependence pattern*. Fisher himself noted this difficulty and commented on *"the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than*

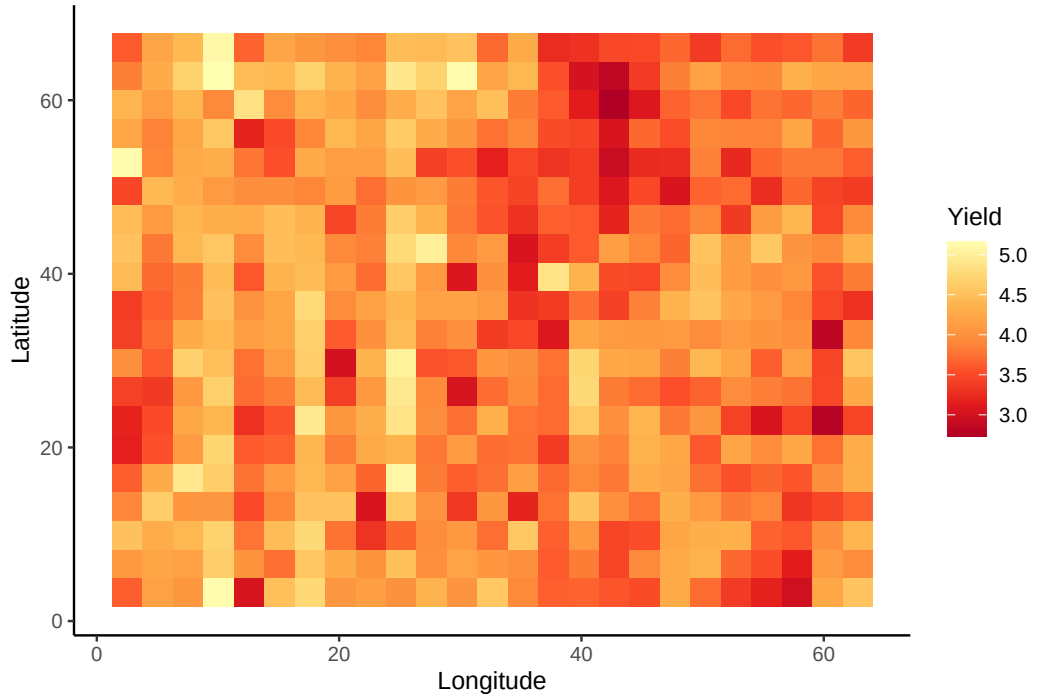


Figure 1.1. Data from a wheat uniformity trial reported in Mercer and Hall (1911). Each square represents a rectangular area of dimension 3.30×2.59 meters. Squares are coded to show the yield of wheat grain from each plot.

those which are farther apart". In order to overcome the distorting influence of an extraneous and unknown source of variability, Fisher advocated the use of *blocking* as a fundamental design principle. In practice, the study region was to be divided in $k = 1, \ldots, K$ portions (containing more than one rectangle) in which the contained rectangles were assumed to partially share a common deviation from the population mean. Such model can be expressed as follows:

$$Y_{ij} = \mu + \beta_{k_{ij}} + Z_{ij}, \quad \forall i,\, j,$$

where the β_k's are constrained to sum to zero and represent the amount by which each portion deviates from the population mean, while $k_{(i,j)}$ links each rectangle (i, j) to the portion it belongs to. It sounded reasonable to define blocks as groups of contiguous arrays, and Fisher himself affirmed that *"it is therefore a safe rule to make the blocks as compact as possible"*(Fisher et al. (1960), p.66). In this sentence, Fisher is implying the existence of a spatial effect that varies systematically over the study region and that, in particular, is assumed to be piece-wise constant within blocks. This is the first documented example of trying to acknowledge spatial covariates with the special role of driving similarity, but as in all revolutionary scientific developments, it does not come alone but is soon followed by others.

Some years later Papadakis (1937) suggests that each yield should be adjusted to take into account the average yield of neighboring rectangles. This is also a form of covariate adjustment, and in some sense, it is a dynamical version of blocking. In a modern interpretation it can be viewed as a conditional model for the plot yield Y_{ij}, given the average yield $\bar{Y}_{N(i,j)}$ over rectangles judged to be neighbors of (i, j):

$$Y_{ij} \mid \{Y_{kl} : (k, l) \neq (i, j)\} = \mu + \beta \cdot (\bar{Y}_{N(i,j)} - \mu) + Z_{ij}, \quad \forall i,\, j. \tag{1.1.1}$$

This model is a primitive example of the more famous Besag's automodels, where spatially lagged values of the observed variable itself (or combinations/transformation of these) were treated as explanatory variables, and that nowadays would be included in the larger context of *Gaussian Markov Random Field* (Rue and Held, 2005). To be accurate, the first example of formalization in a coherent e complete framework in the study of data such the ones in Figure 1.1 according to this logic dates back to Dobruschin (1968). However, it is only With Besag (1974) that this specific branch gained a popularity and a dignity of its own, and is nowadays defined as the study of *lattice systems*.

At the beginning of the 20-th century, also W.F. Gosset was thinking about the problem of spatial correlation, but from a different perspective. He was pursuing the determination of a general model-based solution and in a letter to Karl Pearson in December 1910 he writes: *"Now, in general the correlation weakens as the unit of time or space grows larger and I can't help thinking that it would be a great thing to work-out the law according to which the correlation is likely to weaken with increase of unit"*(Pearson et al., 1990). This is the first statistical approach to the *first law of geography*, decades later made famous by W.Tobler with the statement *"everything is related to everything else, but near things are more related than distant things."*. Differently from Fisher, Gosset was intrinsically considering a continuous domain and the seeds for the definition of spatial covariance function lie in his question. Formalizing it in probabilistic terms, he is looking for a law $f(\cdot)$ that would regulate the covariance between any two realizations of a stochastic process $Y(\cdot)$ at different locations x and y:

$$\mathbb{C}ov\left[Y(x), Y(y)\right] = f(||x - y||)),$$

where $||\cdot||$ denotes a measure of distance. It is now established that looking for some natural, immanent, and unchangeable law of this kind is simply impossible. However, it is possible to define a parsimonious class of models able to appropriately account for these *continuous spatial variations* and approximate reality to the extent of letting researchers reach their inferential goals. Given the complicated nature of the problem, two very commonly used working assumptions are that the process $Z(\cdot)$ is Gaussian and that, given the mean function, it is also stationary over the domain of interest. This last assumption allows specifying the covariance just as the product of a scalar parameter σ^2 and a function of the distance between two points $\rho(\cdot)$:

$$\mathbb{C}ov\left[Z(x), Z(y)\right] = \sigma^2 \cdot \rho(||x - y||),$$

where $\sigma^2 = \mathbb{V}[Z(x)] = \mathbb{V}[Z(y)]$ and $\rho(||x - y||) = \mathbb{C}orr\left[Z(x), Z(y)\right]$. This idea was not fully developed until it was independently applied by Matérn (1960) on forestry data and by Krige (1951) in mining engineering. The independent adoption of this approach in such unrelated and distant fields was the first hint for the revolutionary extent of this new modeling framework.

1.2 Content of the thesis

The brief historical introduction provided above has recognized, in a nutshell, three major branches of statistics: *spatial point patterns*, *discrete spatial variations* and *continuous spatial variations*. These three branches inevitably share a lot among themselves from a theoretical point of view. However, the (sometimes large) difference in terms of tools and methods applicable in their separate contexts and the great variety of different applications and approaches existing in each group, provide each of these a dignity of its own.

During the three years of my Ph.D. program, I had the chance to apply and produce original work in both the context of *continuous spatial variations* and *spatial point patterns*, while only more recently I also worked on *discrete spatial variations*[1]. Therefore, in this dissertation, I will limit the discussion to the first two areas and, within these, only to the aspects I consider of main importance for understanding the key passages and results of Alaimo Di Loro et al. (2021) and Kalair et al. (2020).

Chapter 2 provides a brief introduction to the problem of modeling spatial variations over continuous domains, highlighting both theoretical and practical motivations on which all modeling attempts in this area are based. This Chapter only scratches the surface of the various possible methods to model continuous data observed over a continuous domain. Nevertheless, it provides context and outlines the most relevant passages for understanding the following Chapter. In particular, Section 2.2 outlines the principles of *Spline Regression* (Wahba, 1990; Hastie et al., 2017), with references to the spatial context; Section 2.3 concerns Bayesian Hierarchical modeling in the Geo-Statistical setting (Gelfand et al., 2010; Banerjee et al., 2014). Section 2.4 provides a brief consideration on some possible extensions of spatial techniques to the spatio-temporal setting. After that, Section 2.5 introduces the issues associated with the analysis of *Big* spatial data, with a focus on the *Directed Acyclic Graph* based approximation known as the *Nearest Neighbor Gaussian Process* method for overcoming all the associated computational problems (Datta et al., 2016a). Finally, Chapter 3 shows a novel application published in the pre-print Alaimo Di Loro et al. (2021) and currently under review. In this, both spline regression and NNGP are used to model physical activity level data collected through modern accelerometer and GPS devices.

As for Chapter 2 in the case of spatial variations, Chapter 4 provides a brief introduction to the problem of modeling spatial point patterns, highlighting both theoretical and practical motivations on which all modeling attempts in this area are based upon. The intention is not to give a complete ex-cursus of all the currently known and adopted methodologies, but the discussion will be limited to the most relevant for proper understanding of the following Chapter. Section 4.2 introduces the general properties of the widely used *Poisson Process*, including the most typical variations on it. Section 4.3 is dedicated entirely to the Hawkes Process (Hawkes, 1971b), a particular kind of self-exciting point process that is revealing itself successful in more and more applications during recent years (Laub et al., 2015; Reinhart et al., 2018). Finally, Chapter 5 introduces the semi-parametric spatio-temporal Hawkes model with periodic background recently proposed by Zhuang and Mateu (2019) for the analysis of crime data. In particular, Section 5.2 and Section 5.3 include two original applications to road accidents data. The first is a work started during the 2019 LML Summer School under the supervision of Prof. J. Zhuang and today is still a work in progress; it presents some important passages which I considered valuable enough to be included in this dissertation. The latter is the result of a fruitful collaboration with two researchers from University of Warwick and the London Mathemathical Laboratory, K. Kalair and C. Connaughton, recently published in Kalair et al. (2020).

[1] (Mingione et al., 2021)

Chapter 2

Continuous spatial variations

This chapter will discuss some statistical approaches commonly used to deal with data that comprise information about their spatial location. It focuses on the case when precise knowledge of the location (spatial coordinates and/or time instant) within a given domain of interest are available and fixed. In contrast, the realizations of the outcome of interest at each location are random. Given the observations at the observed set of locations, we would make inference on the parameters regulating the underlying dependence structure and perform kriging at arbitrary locations (from Krige (1951)), namely provide predictions at unsampled location for the process of interest. Section 2.1 provides the basics ideas and ingredients on the analysis of spatial stochastic processes, while Sections 2.2 and 2.3 describe *Spline Regression* and *Bayesian Hierarchical Modeling* respectively. Finally, Section 2.5 introduces the *big n problem* for the estimation of models accounting for dependent observations. Different strategies are listed, but the section entirely focuses on the Directed Acyclic Graph (DAG) based approximation known as *Nearest Neighbor Gaussian Process* (NNGP) (Datta et al., 2016a). An analogous strategy is indeed used in the application of Chapter 3.

2.1 Analysis of spatial stochastic processes

First of all, some basic notation pervading the whole chapter is introduced. Let us denote the domain of interest of the analysis as $\mathcal{D} \subseteq \mathbb{R}^d$, with $d \in \mathbb{N}^+$, and the set of observed locations as $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^n \subseteq \mathcal{D}$. The vector of observed values is identified by $\boldsymbol{y} = [y_i]_{i=1}^n$, where the i-th component $y_i \in \mathcal{Y}$ corresponds to the outcome value observed at the i-th location.

In approaching the analysis of a spatial dataset with the tools of spatial statistics, we are inevitably dropping the hypothesis of independence of the outcomes. From now on, we will see $\boldsymbol{y}$ as the finite realization of a spatial stochastic process $Y(\cdot,\cdot) : \mathcal{S} \to \mathcal{Y}$, in the sense that:

$$\begin{aligned} &\boldsymbol{y} = Y(\mathcal{S}, \omega), \text{ with} \\ &Y(\mathcal{S}, \omega) = [Y(\boldsymbol{s}_1, \omega), \ldots, Y(\boldsymbol{s}_n, \omega)]^\top, \quad \omega \in \Omega. \end{aligned} \tag{2.1.1}$$

In loose terms, the observed vector of values $\boldsymbol{y}$ is the value assumed by this process over the set $\mathcal{S}$ at the realization of a specific random *trajectory* ω. If ω is not fixed, the vector $Y(\mathcal{S}, \cdot)$ is not determined, but it is a $n-$variate *random* vector with a (possibly) well-defined joint distribution. Under desirable conditions, that we will present hereafter, the multivariate distribution of $Y(\mathcal{S}, \cdot)$ reflects the spatial structure

of the process at any other set of locations $\mathcal{U} \in \mathcal{D}$. Hence, our final objective is to draw inferences on this general structure exploiting the information from the realized trajectory ω that produced the observed set of data. In the sequel, for ease of notation, the dependence of the realization on the trajectory ω will be omitted: we will use $\boldsymbol{y}$ to denote the observed values over $\mathcal{S}$, while $Y(\mathcal{S}) = Y(\mathcal{S}, \cdot)$ to denote the process at $\mathcal{S}$ as a random vector.

The distribution of the spatial stochastic process $Y(\cdot)$ is defined through the specification of all its finite-dimensional joint distributions:

$$F_{Y(\cdot)}(y_1, \dots, y_l;\, \boldsymbol{u}_1, \dots, \boldsymbol{u}_l) = \mathbb{P}(Y(\boldsymbol{u}_1) < y_1, \cdots, Y(\boldsymbol{u}_l) < y_l), \tag{2.1.2}$$

for any l and collection of locations $\mathcal{U} = \{\boldsymbol{u}_i\}_{i=1}^{l}$ in $\mathcal{D}$. In particular, the process is called *strictly stationary* if the family of joint finite dimensional distributions is invariant under spatial shifts, i.e.

$$F_{Y(\cdot)}(y_1, \dots, y_l;\, \boldsymbol{u}_1 + \boldsymbol{h}, \dots, \boldsymbol{u}_l + \boldsymbol{h}) = F_{Y(\cdot)}(y_1, \dots, y_l;\, \boldsymbol{u}_1, \dots, \boldsymbol{u}_l), \tag{2.1.3}$$

with $\boldsymbol{h} \in \mathbb{R}^d$. The *stationarity* property is of the utmost importance since it allows to generalize information gathered from the finite set of observed locations $\mathcal{S}$ to any other set of location $\mathcal{U} \subset \mathcal{D}$. Therefore, we do need a valid definition of (2.1.2), but for the process to be predictable[1] we also need it to be stationary in some sense.

Usually, the analysis of spatial stochastic processes does not encompass the direct design of this finite-dimensional distribution but it focuses on suitable modeling of its first and second moment properties. In practice, we model uniquely the mean and covariance terms of the outcome at any finite set of locations $\mathcal{U}$:

$$\begin{aligned} \boldsymbol{\mu}(\mathcal{U}) &= \mathbb{E}\left[[Y(\boldsymbol{u}_1), \dots, Y(\boldsymbol{u}_l)]^\top\right] \\ \boldsymbol{\Sigma}(\mathcal{U}) &= \mathbb{C}\mathrm{ov}\left[[Y(\boldsymbol{u}_1), \dots, Y(\boldsymbol{u}_l)]^\top\right]. \end{aligned} \tag{2.1.4}$$

Generally speaking, these two terms cannot be modeled independently under the most of valid probability distributions and their arbitrary specification is not guaranteed to produce a probabilistically valid model. Indeed, the Kolmogorov existence theorem states that (2.1.2) is valid if and only if all the finite-dimensional joint distributions are consistent under reordering of the sites. While such requirement is somewhat intuitive, the formal proof is quite burdensome (refer to Billingsley (2008) for technical details) and general sufficient conditions on (2.1.4) are not straightforward to derive.

Nevertheless, an important special case is that of a *Gaussian process* (GP). For Gaussian processes the finite dimensional distributions (2.1.2) are multivariate Normal and thus characterized entirely and independently by their mean and covariance terms. This identifies a straightforward criteria to define a strict stationary Gaussian process. Indeed, strict stationarity is equivalent to *weak* or *second order* stationarity:

$$\begin{aligned} &\mathbb{E}\left[Y(\boldsymbol{u})\right] = \mathbb{E}\left[Y(\boldsymbol{u} + \boldsymbol{h})\right] = \mu \\ &\mathbb{C}\mathrm{ov}\left[Y(\boldsymbol{u}),\, Y(\boldsymbol{u} + \boldsymbol{h})\right] = \mathbb{C}\mathrm{ov}\left[Y(\boldsymbol{0}),\, Y(\boldsymbol{h})\right] = c(\boldsymbol{h}), \end{aligned}$$

where $\boldsymbol{u} \in \mathcal{D}$ is a generic location and the function $c(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is called the *covariance function*. Furthermore, the Kolmogorov condition is easy to verify in

[1] The way the process changes does not vary so that the evolution of the process can be modeled and forecasted at unobserved locations

the Gaussian case as it holds if and only if the covariance matrices corresponding to each finite set are always non-negative definite. For these two reasons, and also the great flexibility and easiness of use of the Gaussian distribution, Gaussianity is a very common working assumptions in the analysis of spatial processes. The non-Gaussian case is considerably more complex and when the nature of the outcome of interest is incompatible with such assumption, one generally resorts to hierarchical specifications depending on latent Gaussian components (Banerjee et al., 2014).

In practice, the validity of a Gaussian process depends uniquely on the suitable specification of the covariance function $c(\cdot)$. With suitable we mean that for any finite collection of sites the resulting covariance matrix:

$$\mathbb{C}\text{ov}\left[[Y(\boldsymbol{u}_1), \cdots, Y(\boldsymbol{u}_l)]^\top\right] = \begin{bmatrix} c(\mathbf{0}) & c(\boldsymbol{u}_1 - \boldsymbol{u}_2) & \cdots & c(\boldsymbol{u}_1 - \boldsymbol{u}_l) \\ c(\boldsymbol{u}_2 - \boldsymbol{u}_1) & c(\mathbf{0}) & \cdots & c(\boldsymbol{u}_2 - \boldsymbol{u}_l) \\ \vdots & \vdots & \ddots & \vdots \\ c(\boldsymbol{u}_l - \boldsymbol{u}_1) & c(\boldsymbol{u}_l - \boldsymbol{u}_2) & \cdots & c(\mathbf{0}) \end{bmatrix} \tag{2.1.5}$$

shall verify Kolmogorov consistency, thus be a non-negative definite matrix. This condition is equivalent to asking for the covariance function to be *positive definite*. By Bochner's theorem (Bochner (1933), Bochner (2005)) a real-valued continuous function $c(\cdot)$ is positive definite if and only if it is the Fourier transform of a symmetric, non-negative measure F on $\mathbb{R}^d$. Assuming that the spectral measure has a Lebesgue density f (i.e. *spectral density*) this reduces to:

$$c(\boldsymbol{h}) = \int_{\mathbb{R}^d} \exp\left(i\boldsymbol{h}^\top \boldsymbol{x}\right) dF(x), \tag{2.1.6}$$

for any $\boldsymbol{h} \in \mathbb{R}^d$ and any measure F. A particular sub-class of second-order stationary processes is when the covariance between different locations depends on the spatial separation $\boldsymbol{h}$ only through its euclidean length $||\boldsymbol{h}||$ (distance between locations):

$$\mathbb{C}\text{ov}\left[\boldsymbol{u}, \boldsymbol{u} + \boldsymbol{h}\right] = c(\boldsymbol{h}) = \sigma^2 \cdot \rho(||\boldsymbol{h}||), \qquad \boldsymbol{h} \in \mathbb{R}^d. \tag{2.1.7}$$

The right-hand decomposition sees it as the product of a correlation function $\rho(\cdot)$ such that $\rho(0) = 1$, and a multiplicative factor σ^2 that controls for the general process dispersion and allows conversion from correlation to covariance. Such processes and the corresponding covariance functions are called *isotropic*, and compose the most commonly used class of spatial processes in the literature.

In general, given the absence of a simple sufficient condition, defining a valid covariance function in the general space is not straightforward at all. Theoretically speaking, it is always possible to obtain one by exploiting the definition in Equation (2.1.6) and deriving the second-order properties of the convolution of independent valid processes; however, there are only rare cases in which analytical calculations are forgiving. Nevertheless, different parametric families of isotropic covariance functions have been proved to be valid (i.e. respect (2.1.5)) in the literature. Among these we find: the *Matérn* class, firstly introduced in Matérn (1960); the *powered exponential family* (Diggle et al., 1998); the Cauchy family (Gneiting, 2002) and others. More details on covariance functions and the Matérn class will be provided in Section 2.1.2, where the text focuses on the basics of the geo-statistical approach.

How can we control for both the first and second-order structure while preserving the validity of our specification? A very convenient representation of any stochastic process $Y(\cdot)$, that is also very common in practice, is the following:

$$Y(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \epsilon(\boldsymbol{u}), \tag{2.1.8}$$

where $\mu(\boldsymbol{u}) = \mathbb{E}[Y(\boldsymbol{u})] \in \mathbb{R}$ is a deterministic and smooth *mean* function, while $\varepsilon(\boldsymbol{u}) \in \mathbb{R}$ is a stochastic residual term. Let us stress the point that the decomposition (2.1.8) does not impose any kind of hypotheses on the process. $\mu(\cdot)$ is a also called *spatial trend* and it characterizes the *global* structure of the process, while $\epsilon(\cdot)$ is the *spatial residual* process and it represents local variations about the mean and characterizes its *local* structure. In a standard modeling framework, we would wish for the spatial trend to disentangle information from pure white noise, so that the spatial residual terms are statistically independent. But the spatial trend term, however complicated and sophisticated its expression, may not be able to properly capture small-scale variations while accounting for the global structure without incorporating noise. Hence, one can expect the residual terms to violate the usual independence assumption and exhibit similar values at locations close together in space. When this is the case, it is essential to allow for positive statistical dependence between values of the spatial residual process at different locations, supposedly decreasing with distance. Considering that this term is first-order stationary by definition (it has 0 mean), weak stationarity (which is also strict in the Gaussian case) is preserved if the original $Y(\cdot)$ satisfies second-order stationarity.

Therefore, the final objective of spatial modeling boils down to estimating the expression of these two terms in order to understand and predict the behavior of the original process $Y(\cdot)$ at arbitrary locations under optimal conditions (e.g. in terms of *Mean-Squared-Error* (MSE)). From a practical view-point, these two terms are convoluted and indistinguishable *a-priori*, but (under suitable hypotheses) their intrinsic structure can be estimated from the observed vector $\boldsymbol{y}$. The general modeling strategy aims at identifying the function $\mu(\cdot)$ in such a way that the resulting residual process $\epsilon(\cdot)$ is indistinguishable from (potentially correlated) *random noise*. However, there is an inner circularity in the estimation of these first-order and second-order structures: in order to get one, we need to know the other. Fixing a suitable parametric (or semi-parametric) form of these two terms, there are different strategies that allow for the contemporary estimation of both components. Nevertheless, if we want to avoid a grueling and potentially infinite model comparison process, there is still the issue of properly defining these forms in advance.

The typical spatial statistics analysis flow usually starts from the estimation of the simplest possible first-order structure, which may be the naive general average or a slightly more sophisticated 1-st degree spatial trend:

$$\boldsymbol{\mu}(\mathcal{S}) = \boldsymbol{S}\boldsymbol{\beta} \;\leftarrow\; \boldsymbol{S}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}}(\mathcal{S}), \tag{2.1.9}$$

where $\boldsymbol{S}$ is the $n \times d$ matrix of observed locations and $\boldsymbol{\beta}$ is a $d \times 1$ vector of coefficients with estimate $\hat{\boldsymbol{\beta}}$. If the resulting residuals:

$$\hat{\boldsymbol{e}}(\mathcal{S}) = \hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\mu}(\mathcal{S}) \tag{2.1.10}$$

do not respect independence but manifest a spatial correlation pattern, then there is a decision to make. Either a more complicated forms for the trend term $\mu(\cdot)$ is considered, hoping to reach independence of residuals without capturing noise together with information; or the modeling of the dependence pattern in the residual process is pursued, introducing a parametric expression for the covariance function $c(\cdot)$. These two approaches can be seen as alternative (and usually provide similar results), but can also be combined in a more exhaustive modeling framework[2]. We refer to the former as *spatial interpolation* and to its parsimonious implementation as *spatial smoothing*; a brief introduction to it is provided in Section 2.1.1 while

[2]Identifiability issues may arise and must be averted

Section 2.2 focuses specifically on its implementation via *Spline Regression*. The latter has its roots in the original geo-statistical literature and we briefly describe its basics in Section 2.1.2; a more thorough description of the geo-statistical modeling framework in the context of Bayesian hierarchical modeling is provided in Section 2.3.

2.1.1 Spatial Interpolation and Smoothing

One possible approach to explain changes in spatial variations considers these as the (potentially noisy) partial observations of an unknown *deterministic* trend function. The underlying assumption is that the residuals shall behave as a stationary and independent process once the *true* expression of $\mu(\cdot)$ is known. Potentially, if there was no unobserved spatial heterogeneity, then this effect could be completely explained by the available covariates (see *land-use regression* (Ryan and LeMasters, 2007; Hoek et al., 2008)). On the other hand, this is practically never the case in real applications.

Let us initially assume the process of interest is observed without error; we are practically asserting that the observed outcome $(y_i, \boldsymbol{s}_i),\ i = 1, \dots, n$ is equal to the value assumed by a true but unknown continuous function $\mu(\cdot) : \mathcal{D} \to \mathbb{R}$ at the same location set:

$$\boldsymbol{y} = \boldsymbol{\mu}(\mathcal{S}) = \left[\mu(\boldsymbol{s}_1), \cdots, \mu(\boldsymbol{s}_n)\right]^\top.$$

To our end, the surface $\mu(\cdot)$ represents the effect of unobserved spatially varying factors that affect our outcome, and therefore induces dependence between observations close in space (these factors vary continuously over $\mathcal{D}$). In general, this function may also depend on other possibly observed covariates $X(\cdot) : \mathcal{D} \to \mathbb{R}^p$ (either geo-referenced or observation specific), but for ease of explanation and without any loss of generality, we now consider dependence upon location $\boldsymbol{u} \in \mathcal{D}$ only. In this setting, the final goal is to estimate a function $\hat{\mu}(\cdot)$ interpolating the observed values at the observed locations but defined on the whole domain $\mathcal{D}$. To this end, the choice of the building blocks for the interpolating function are of the utmost importance, since they will determine the function behavior at all unsampled locations. The estimated function shall be continuous and, desirably, satisfies smoothness up to some degree[3]. For instance, as shown in Equation (2.1.9), the analysis usually starts from a first-order polynomial fit: this is rarely flexible enough for achieving interpolation and capturing any small-scale variation. Nevertheless, starting from there, one may consider the use of a higher order polynomial in the coordinates:

$$\mu(\boldsymbol{u}) \ \leftarrow\ f(\boldsymbol{u}) = \text{Poly}\left(\boldsymbol{u}, q\right)\boldsymbol{\beta}_q,\ \forall \boldsymbol{u} \in \mathcal{D}, \tag{2.1.11}$$

where q is the order of the polynomial and $\text{Poly}\left(\boldsymbol{u}, q\right)$ is the row-vector containing the corresponding spatial coordinates expansion. Its estimation can be tackled as a standard linear regression problem, and fitting to the observations $\boldsymbol{y}$ can be straightforwardly obtained through *Ordinary Least Squares* (OLS). The estimated function $\hat{f}(\cdot)$ will eventually achieve interpolation for large enough q. However, this approach is not efficient from different points of views. First of all, polynomial terms are usually highly correlated, which can jeopardize the proper estimation of the coefficients. Secondly, perfect interpolation is likely to be achieved only for values of q such that the number of polynomial terms $q \times d$ (and hence of coefficients to

[3]The resulting function does not vary wildly but preserves a soft curvature, which may be quantified in terms of its second-order properties (e.g. second derivative in one-dimensions, Hessian in two-dimensions)

estimate) is close to n, making our modeling efforts of limited utility, at the very least. Similar alternatives consider the inclusion of other non-linear terms (e.g. sines and cosines with unknown phase, amplitude and period etc.) for improving on the function flexibility, but the larger the dimension d the least practical such solutions turn out to be.

This interpolation approach's significant limitation is that it is trying to define a *global* function with the requirement of achieving a practically perfect *local* adaptation, which is a particularly tough task for parametric forms. Alternatively, one may consider adopting non-parametric mean functions, whose definition depends on the observed sample's local behavior (Lam, 1983). Among the non-parametric alternatives we find *Kernel smoothing models* (Hastie et al., 2017), which define the function at each location $\boldsymbol{u} \in \mathcal{D}$ as a weighted combination of the values in the observed set:

$$\hat{f}(\boldsymbol{u}) = \frac{\boldsymbol{w}(\boldsymbol{u})^\top \boldsymbol{y}}{\sum_{i=1}^{n} w_i(\boldsymbol{u})}. \tag{2.1.12}$$

$\boldsymbol{w}(\boldsymbol{u})$ is a $n \times 1$ column vector whose $i-$th element $w_i(\boldsymbol{u})$ is the contribution of the $i-$th observation $y_i = Y(s_i)$ to the value of the interpolating function at location $\boldsymbol{u}$. Intuitively, points closer to $\boldsymbol{u}$ shall have more importance. Thus, weights are usually positive decreasing functions of the distance between the location of interest $\boldsymbol{u}$ and the interpolation set $\boldsymbol{s}$:

$$w_i(\boldsymbol{u}) = w(||\boldsymbol{s}_i - \boldsymbol{u}||), \quad \forall\, i = 1, \ldots, n,$$

where $w(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ and $w'(\cdot) \geq 0$. If $w(\cdot)$ is chosen such that the weight grows to infinity as the distance approaches 0, then we achieve exact interpolation at the observed location. This is the case of the well-known *Inverse Distance Weighting* estimator (Shepard, 1968), part of the larger class of *Inverse Power Function* kernel smoothers, but for instance it does not hold for the *Negative Exponential Function* weights. They both provide continuous functions in the end, but the former achieves exact interpolation at the cost of sacrificing smoothness (it is usually very peaked around observed locations), while the latter generally produces surfaces which are smoother than the data itself.

An alternative approach stems from *Finite Element Methods* (FEM) in geometric domains (Zienkiewicz and Taylor, 1989), and started in the field of engineering under a completely algorithmic perspective. Loosely speaking, it consist of partitioning the domain of interest into polygonal sub-domains, usually with all the observed points as vertices. These partitions are called *meshes* and, in the case of 2-D problems, usually have triangular or quadrilateral shapes, while they may well be segments in 1-dimensional problems. For instance, a very popular method for the generation of triangular meshes on spatial domains is the *Delaunay Triangulation* (Shewchuk, 1996; Toth et al., 2017). Interpolation at any location within the sub-domains is obtained as a combination of the vertices values through suitable shape functions (linear, quadratic, polynomials etc.), specifically designed for the mesh geometric shape so to guarantee interpolation and continuity at the bounds.

As for kernel smoothing approaches, there is not any parameter estimation process at play, but only ad-hoc defined combinations of the observed data. A major drawback of such completely non-parametric approaches lies in their *algorithmic* nature: the chosen method (e.g. weighting function, mesh, shape function, etc.) determines uniquely the resulting interpolator, without going through any kind of estimation process. This creates several issues for uncertainty estimation of the interpolated values, which will not be discussed here for the sake of brevity.

In between the two aforementioned alternatives we find a third way, that consists of obtaining the expression of a globally defined surface as a parametrized combination of locally well defined functions. For instance, the global polynomial fit proposed ad the beginning of this section could be replaced by *local-polynomials*. These assume that the *local curvature* of the mean surface can be well-approximated by a polynomial of a given order p; the un-sampled location $\boldsymbol{u}$ is then given a restricted interpolation set $\mathcal{I}(\boldsymbol{u}) \subseteq \mathcal{S}$, usually composed of the neighboring observations, so that the polynomial can be locally estimated through OLS. However, the resulting function is generally not smooth and is also likely to not be continuous[4].

In light of these issues, we seek a semi-parametric procedure that guarantees to yield a locally adaptable surface not only continuous, but in fact *smooth* everywhere. In this sense, a solution is provided by the consideration of suitable basis functions defined on the whole region of interest $\mathcal{D}$, but with a flexible local behavior. The combination of its elements through a set of parameters shall guarantee interpolation while preserving smoothness globally. Similarly to meshes in FEM, the definition of such bases usually requires the choice of a reference set of points called knots $\mathcal{K} = \{\boldsymbol{k}_j\}_{j=1}^{K} \in \mathcal{D}$ that span the considered domain. Interpolation within sections identified by the knots is mostly determined by local adaptation, but it also accounts for the behaviour in neighboring sections in order to retain continuity and smoothness. The finer the grid of knots, the more locally flexible the resulting function. For instance, if exact interpolation is the target, then we can just fix $\boldsymbol{k}_i \equiv \boldsymbol{s}_i$, $\forall\, i$. The mean function is then approximated as:

$$\mu(\boldsymbol{u}) \leftarrow f(\boldsymbol{u}) = \sum_{j=1}^{K^*} \phi_j \cdot \psi_j(\boldsymbol{u}) = \boldsymbol{\psi}(\boldsymbol{u})^\top \boldsymbol{\phi},$$

where $\psi_j(\cdot)$ is the $j-th$ component of the basis, ϕ_j is the corresponding coefficient and K^* is the number of components (depending on the number of knots but also on the chosen basis decomposition). Interpolation can be achieved simply by estimating the vector of coefficients $\boldsymbol{\phi}$ so that:

$$\boldsymbol{y} = \boldsymbol{\Psi}(\mathcal{S})\,\hat{\boldsymbol{\phi}} \iff \hat{\boldsymbol{\phi}} = \boldsymbol{\Psi}(\mathcal{S})^{-1}\boldsymbol{y}.$$

Very common choices are *radial basis functions* (Buhmann, 2003), such as standard multivariate Normal densities and the inverse multi-quadratic functions, and *spline basis models* (Bde Boor, 2001). We will not say more about the former, while we will spend some few words on the latter. As a matter of fact, while any choice of basis may sound arbitrary to some extent, spline basis are so-called because they produce *spline interpolation*, which is very appealing for its theoretical properties. Indeed, it is not only able to achieve interpolation while ensuring continuity and smoothness, but it is also proved to provide the *smoothest* function possible given the assumed structure. Furthermore, it is also proved to have good properties when observations are assumed to be perturbed with random error:

$$y_i = Y(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + \epsilon_i, \qquad \mathbb{E}\left[\epsilon_i\right] = 0,\ \mathbb{V}\left[\epsilon_i\right] = \tau^2 \quad \forall\, i = 1, \dots, n.$$

In this case, the final objective is no more exact interpolation of the observations, but disentanglement of the true underlying function from the random noise perturbing its

[4]Continuity can be attained if observations in the interpolation set are appropriately weighted according to their distance from the interpolating point $\boldsymbol{u}$ (*local weighted linear interpolation function*).

observed values: following other authors, we refer to this as *spatial smoothing* of the observed process (Green and Silverman, 1993; Hastie et al., 2017). Spline functions have a prominent role in the spatial smoothing (Eilers and Marx, 1996; Wood, 2003). Indeed, they are very flexible and reliable, allow for steep local variation without sacrificing smoothness and also have a very interesting interpretation in terms of *kriging predictors* in the geo-statistical settings (Wahba, 1990). A more detailed discussion about splines will be provided in Section 2.2.

2.1.2 Geo-statistics

Notwithstanding the good properties and flexibility of the chosen deterministic trend function, it is often unable to describe the spatial variations at all scales. The resulting fit is often not spatially homogeneous and produces residuals $\hat{v}_i = y_i - \mu(\boldsymbol{s}_i)$ that are spatially correlated. Differently from *spatial interpolation*, the *geo-statistical* approach encompasses the estimation of both the first and second order structure of the process, with special focus dedicated to the proper identification of the second-order one. The residual term, initially neglected in Section 2.1.1, becomes the component of main importance. Without any loss of generality, it is generally seen as the sum of two additional terms:

$$Y(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \epsilon(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \eta(\boldsymbol{u}) + \upsilon(\boldsymbol{u}), \qquad \boldsymbol{u} \in \mathcal{D}, \tag{2.1.13}$$

where $\eta(\cdot)$ is a zero-mean second-order stationary process with covariance function $c_\eta(\cdot)$ characterizing the spatially dependent component, while $\upsilon(\cdot)$ is an uncorrelated random noise with variance τ^2 that represents the measurement error. This representation is known as the *nugget effect* covariance model, where the resulting complete covariance function is:

$$\mathbb{C}\mathrm{ov}\left(Y(\boldsymbol{u}), Y(\boldsymbol{u}+\boldsymbol{h})\right) = \begin{cases} \tau^2 + c_\eta(\boldsymbol{0}) & \boldsymbol{h} = \boldsymbol{0} \\ c_\eta(\boldsymbol{h}) & \boldsymbol{h} \neq \boldsymbol{0} \end{cases} \tag{2.1.14}$$

and τ^2 is called *nugget*. The second order stationarity assumption on $\epsilon(\cdot)$ implies *intrinsic stationarity* on the residual process (i.e. mean and variance are invariant to location). This allows us to define unambiguously the variance of the difference between the error term at two different locations as a function of the spatial separation $\boldsymbol{h}$ only. The half of this quantity is universally known as the *semi-variogram*, and it is computed as:

$$\begin{aligned} \gamma(\boldsymbol{h}) &= \frac{1}{2} \cdot \mathbb{V}\left[\epsilon(\boldsymbol{u}) - \epsilon(\boldsymbol{u}+\boldsymbol{h})\right] = \\ &= \frac{1}{2}\left(\mathbb{V}\left[\epsilon(\boldsymbol{u})\right] + \mathbb{V}\left[\epsilon(\boldsymbol{u}+\boldsymbol{h})\right]\right) - \mathbb{C}\mathrm{ov}\left[\epsilon(\boldsymbol{u}), \epsilon(\boldsymbol{u}+\boldsymbol{h})\right] = \\ &= \begin{cases} 0 & \boldsymbol{h} = \boldsymbol{0} \\ \left(\tau^2 + c_\eta(\boldsymbol{0}) - c_\eta(\boldsymbol{h})\right) & \boldsymbol{h} \neq \boldsymbol{0} \end{cases}, \end{aligned} \tag{2.1.15}$$

where the *nugget* $\tau^2 > 0$ is inducing a discontinuity at the origin.

A stronger but very common assumption, not strictly necessary but important nonetheless, is the one of second-order *isotropy*. We already introduced it for covariance functions in (2.1.7) and it naturally extends to the semi-variogram as:

$$\gamma(\boldsymbol{h}) = \gamma(||\boldsymbol{h}||), \qquad \boldsymbol{h} \in \mathbb{R}^d.$$

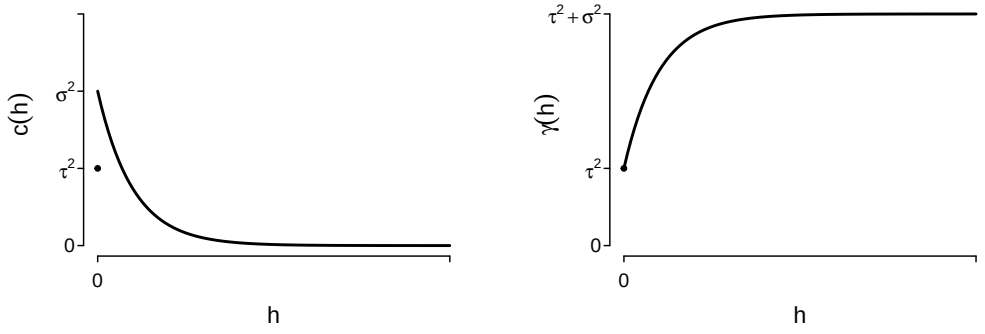


Figure 2.1. Generic covariance function (on the left) and corresponding semivariogram (on the right) with nugget τ^2 and partial sill σ^2 annotated

Isotropy implies that iso-level contours of the semi-variogram (and of the correlation function) are d-dimensional spheres: all points at the same distance present analogous correlation and covariance level. This hypothesis sensibly simplifies the analysis of spatial processes and also looks completely reasonable in terms of spatial dependence intuition. Nevertheless, many natural phenomena present anisotropic correlation patterns[5] but their discussion is not included in this work (Zimmerman, 1993; Christakos et al., 2000; Anderes and Chatterjee, 2009). So, even if isotropy is not a necessary condition for pursuing estimation through the geo-statistical approach, in the sequel it will always be assumed to hold.

Without any loss of generality, denoting with $h = ||\boldsymbol{h}||$ and referring to the same notation used in (2.1.7), we can express any covariance function as $c(h) = \sigma^2 \cdot \rho(h)$, where $\rho(\cdot)$ is a correlation function such that $\rho(0) = 1$ and $\rho(h) \overset{h\to\infty}{\to} 0$. Thus, the semi-variogram expression becomes:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \tau^2 + \sigma^2 (1 - \rho(h)) & h \neq 0 \end{cases}, \tag{2.1.16}$$

where $\tau^2+\sigma^2$ is the effective *sill* (sum of *nugget* and *partial sill* of the nested structure $c(\cdot)$) and is the limiting value of the semi-variogram for h approaching infinity. An example of isotropic covariance function and the corresponding semivariogram is graphically represented as a function of the spatial lag h in Figure 2.1, with the sill and nugget values properly annotated.

The semi-variogram has a key role in the context of geo-statistics. Indeed, the second order structure of $\epsilon(\cdot)$ is not estimated directly but is derived from the corresponding estimated semi-variogram. From a practical point of view, the first step of classical geo-statistics consists of obtaining a *provisional* estimate of the mean function. Typically, it is a constant term or a linear function of the coordinates (as in Equation (2.1.9)). In principles, nothing precludes from using higher order polynomials or fancier modeling as the ones introduced in Section 2.1.1, but some care must be taken in the choice of the trend function. As mentioned in Chapter 1, we should avoid absorbing too many degrees of freedom in the large scale component

[5]The most tractable form of anisotropy is *geometric anisotropy* for which $\gamma(\boldsymbol{h}) = \gamma\left(\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{h}\right)^{\frac{1}{2}}$ where $\boldsymbol{A}$ is a positive definite matrix: in this case iso-level contours are d-dimensional ellipsoids.

to estimate second-order properties. When geo-statistics is at play, small-scale variations are supposed to be captured by the covariance structure. Thus, it may be unwise to use trend surfaces that cannot be justified from prior knowledge on the phenomenon or accurate analysis of the data: the surface trend must capture large-scale variations only, without adapting to local ones and confounding with the dependence structure.

Once the provisional estimate of the mean function $\hat{\mu}_p(\cdot)$ is obtained, we can compute the residuals at each sampled locations as:

$$\hat{\epsilon}_i = \hat{\epsilon}(\boldsymbol{s}_i) = Y(\boldsymbol{s}_i) - \hat{\mu}_p(\boldsymbol{s}_i), \qquad i = 1, \dots, n.$$

These residuals are all referred to different locations $\boldsymbol{s}_i$, and each pair $(\hat{\epsilon}(\boldsymbol{s}_i), \hat{\epsilon}(\boldsymbol{s}_j))$ will correspond to a spatial lag $s_{ij} = ||\boldsymbol{s}_j - \boldsymbol{s}_i||$. Partitioning the lag space $\mathcal{H} = \{\boldsymbol{u} - \boldsymbol{t} : \boldsymbol{u}, \boldsymbol{t} \in \mathcal{D}\}$ into bins $H_1, \dots, H_K$, we can then compute the so-called *empirical semi-variogram* at each bin location as:

$$\hat{\gamma}(h_k) = \frac{1}{2N(H_k)} \sum_{s_{ij} \in H_k} \left(\hat{\epsilon}(\boldsymbol{s}_i) - \hat{\epsilon}(\boldsymbol{s}_j)\right)^2, \qquad k = 1, \dots, K, \tag{2.1.17}$$

where h_k is a representative of the k-th bin H_k and $N(H_k)$ is the number of residuals' pairs with lag s_{ij} falling into the same bin. Partitioning the lag-space also into angle-classes leads to *polar partitioning*, which allows to build omni-directional variograms. This can be very useful to check for anisotropic patterns in the observed data. Empirical estimation of the semi-variogram, whilst being pretty practical, is not devoid of criticism: the binning of the lag-space is quite arbitrary; terms in (2.1.17) are not independent and this inflates variance; the provisional estimate of the mean function (on which the residuals depend on) is inevitably biased; generally speaking (2.1.17) is very sensitive to outliers. The resulting fit is often quite bumpy, and often fails to be conditional non-positive definite (so that the corresponding covariance function would be positive definite). Consequently, whilst being a useful tool for getting an hint of the true underlying second order structure, the empirical semi-variogram is practically never used as the final estimate of the true semi-variogram. Typically, a smoother version is obtained by fitting to its values a function with a pre-specified parametric form $\gamma_\theta(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$, $\theta \in \Theta$. Obviously, the chosen parametric family shall be such that the corresponding covariance function is valid for any admissible value of θ. This validity transfers onto the semi-variogram through the following three conditions on $\gamma_\theta(\cdot)$.

- Vanishing at the origin: $\gamma_\theta(0) = 0$.
- Evenness: $\gamma_\theta(-h) = \gamma_\theta(h)$.
- Conditional negative definiteness:

  $$\boldsymbol{a}^\top \gamma(\boldsymbol{H}) \boldsymbol{a} \leq 0$$

  where $\boldsymbol{H}$ is a $n \times n$ symmetric matrix of pairwise distances between n locations and $\boldsymbol{a}$ is a $n \times 1$ vector such that $\sum_{i=1}^n a_i = 0$.

Assuming also isotropy and adding the intuitive requirement of positive monotonicity (correlation decays with distance and values at distant locations tend to be less alike than values at close ones), there is a large variety of models which satisfies all these criteria. Recalling from (2.1.16) that $\gamma_\theta(h) = \tau^2 + \sigma^2 (1 - \rho_\theta(h))$ for $h > 0$ (it must be equal to 0 for $h = 0$), we here introduce 4 of the most widely used models in terms of the corresponding correlation function $\rho_\theta(\cdot)$.

- **Exponential**. Correlation decreases exponentially towards zero as:

$$\rho_\theta(h) = \exp\{-\rho \cdot h\}, \quad h > 0,$$

where ρ is a positive parameter.

- **Gaussian (squared exponential)**. Correlation decreases towards zero with a Gaussian behavior:

$$\rho_\theta(h) = \exp\left\{-\rho^2 \cdot h^2\right\}, \quad h > 0,$$

where ρ is a positive parameter.

- **Matèrn**. Firstly introduced in Matérn (1960), it is probably the most general expression of a correlation function. It has the form:

$$\rho_\theta(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\sqrt{2\nu} \cdot \frac{h}{\rho}\right) \cdot \mathcal{K}_\nu \left(\sqrt{2\nu} \cdot \frac{h}{\rho}\right), \quad h > 0,$$

where $\Gamma(\cdot)$ is the gamma function, ρ and ν are positive parameters and $\mathcal{K}_\nu$ is the modified Bessel function of second type. The parameter ν is a *smoothness* parameters, in the sense that the corresponding process is m times mean-square differentiable only for $m < \nu$. Exponential and Gaussian are particular cases of this for $\nu = 1/2$ and $\nu \to \infty$, respectively.

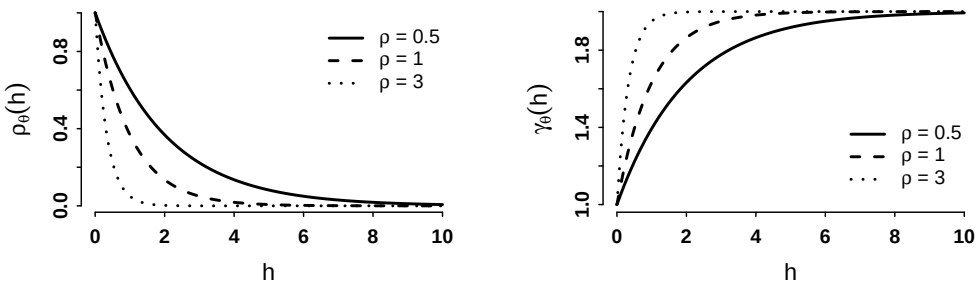
- **Spherical**. Correlation decreases up to some range ρ and vanishes after:

$$\rho_\theta(h) = \begin{cases} 1 - \frac{3 \cdot h}{2 \cdot \rho} + \frac{h^3}{2 \cdot \rho^3} & 0 < h \leq \rho \\ 0 & h > \rho \end{cases},$$

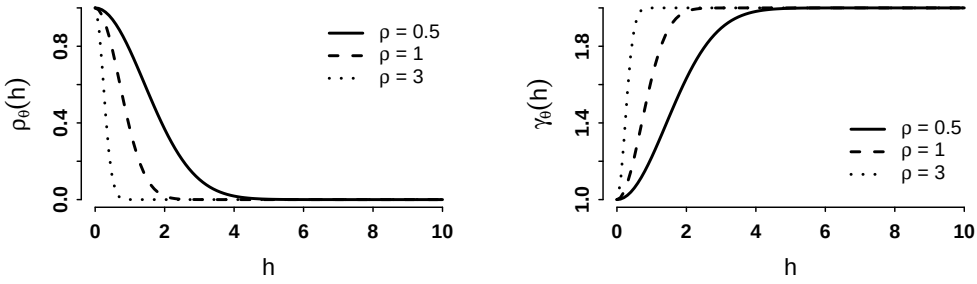
where ϕ is a positive parameter.

A very relevant attribute of any variogram, and by then of any correlation function, is the so-called *range* r. It is defined as the smallest value of the lag h for which the variogram reaches the sill $\tau^2 + \sigma^2$ (and the correlation function reaches 0). In case it is only reached in the limit, we can still define the *practical range* $\widetilde{r}$ as that lag for which the 95% of $(\tau^2 + \sigma^2)$ is reached. In all the introduced families, the parameter ρ determines the range. We have $\widetilde{r} = \frac{1}{3\rho}$ for the Exponential family; $\widetilde{r} = \frac{1}{\sqrt{3}\rho}$ for the Gaussian family; $r = \rho$ for the spherical family. In the Matérn model the practical range $\widetilde{r}$ depends also on the smoothness parameter ν. However, there is no general and simple analytical way to express it, and it is usually computed numerically. A graphical representation of the semi-variograms and corresponding correlation functions of the four proposed models (with typical nugget and sill) is shown in Figure 2.2.

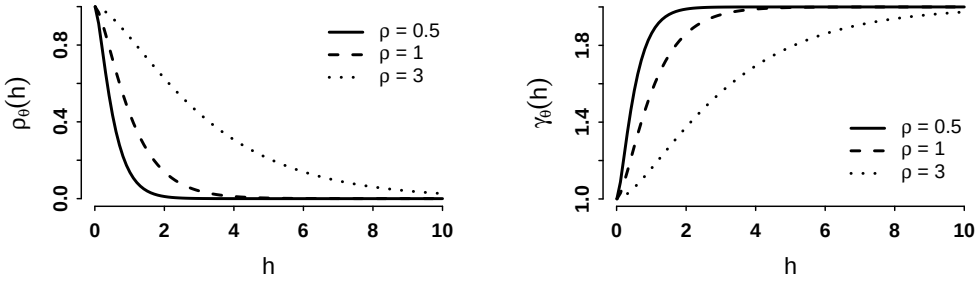
Different procedures allow fitting a parametric family to the empirical variogram, obtaining estimates $\hat{\theta}$ of the whole set of parameters θ (including the nugget τ^2 and sill $\sigma^2 + \tau^2$). The most common are *Weighted Least Squares*, *Maximum Likelihood* and *Restricted Maximum Likelihood* (Mardia and Marshall, 1984; Cressie and Lahiri, 1993; Zhang and Zimmerman, 2005). The choice of the parametric family to fit may be performed in advance (according to visual inspection of the empirical variogram or exploiting prior knowledge on the phenomenon) or by comparing the resulting fit of different alternatives through goodness of fit measures (mean squared error, likelihood, cross-validation, etc.). Once the parametric estimate $\gamma_{\hat{\theta}}(\cdot)$ has been



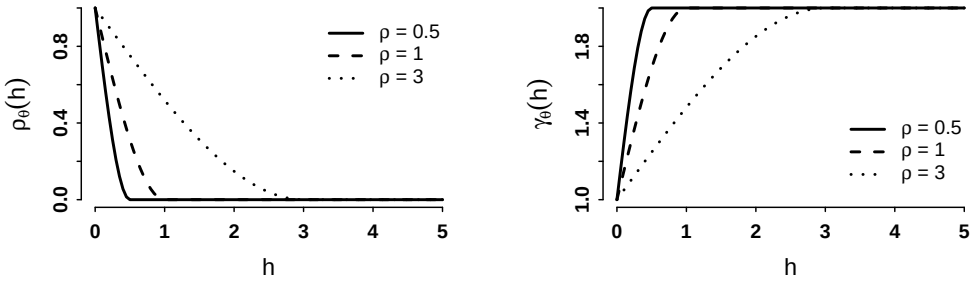
(a) Exponential model



(b) Gaussian model



(c) Matérn model for $\nu = 1$



(d) Spherical model

Figure 2.2. Correlation functions (on the left) and semi-variograms (on the right) for the 4 proposed models, with common $\tau^2 = 1$ and $\sigma^2 = 1$ and different values of the range parameter ρ

obtained, the value of $\hat{\theta}$ determines the shape of the corresponding correlation function $\rho_{\hat{\theta}}(\cdot)$, and the nugget and partial sill. Thus, we can get an estimate of the covariance function of the process:

$$c_{Y,\hat{\theta}}(h) = \begin{cases} \hat{\tau}^2 + c_{\eta,\hat{\theta}}(0) \\ c_{\eta,\hat{\theta}}(h) \end{cases} = \begin{cases} \tau^2 & h = 0 \\ \hat{\sigma}^2 \cdot \rho_{\hat{\theta}}(h) & h > 0 \end{cases},$$

and then proceed to the re-estimation of the mean function while keeping into account the estimated covariance structure. For instance, in a *linear regression* setting, $c_{Y,\hat{\theta}}(\cdot)$ is used to get an estimate of the covariance function at the observed set of locations $\mathcal{S}$:

$$\begin{aligned} \boldsymbol{\Sigma}(\mathcal{S}) \;&\leftarrow\; \hat{\boldsymbol{\Sigma}}(\mathcal{S}) = \boldsymbol{C}_{Y,\hat{\theta}}(\boldsymbol{H}) \\ [\boldsymbol{H}]_{ij} &= h_{ij} = ||\boldsymbol{s}_j - \boldsymbol{s}_i||, \quad \boldsymbol{s}_i,\, \boldsymbol{s}_j \in \mathcal{S},\; \forall\, i, j \end{aligned}$$

where $\left[\boldsymbol{C}_{Y,\hat{\theta}}(\boldsymbol{H})\right]_{ij} = c_{Y,\hat{\theta}}(h_{ij})$. Denoting $\hat{\boldsymbol{\Sigma}}(\mathcal{S}) = \hat{\boldsymbol{\Sigma}}$, we can then use it as a plug-in estimator for obtaining *Generalized Least Square* (GLS) estimation of the spatial trend coefficients:

$$\hat{\boldsymbol{\beta}}_{\text{EGLS}} = (\boldsymbol{S}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{S})^{-1} \boldsymbol{S}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{Y}, \qquad \hat{\mu}_{\text{EGLS}} = \boldsymbol{S}\hat{\boldsymbol{\beta}}_{\text{EGLS}}$$

where EGLS means *Estimated Generalized Least Squares*.

Theoretically, the procedure may be re-iterated by recomputing the residuals, refitting the variogram and finally re-estimating the mean function until some convergence criteria is met. However, in practice, such procedure usually stops after the first estimation.

Once the final mean surface $\hat{\mu}(\cdot)$ and covariance structure $\hat{c}_{\hat{\theta}}(\cdot)$ have been estimated, kriging may be performed at arbitrary sampled and/or un-sampled locations according to the model hypothesis.

However, we cannot help but notice how this approach is intrinsically flawed. The provisional identification of the mean function, which is mainly arbitrary in both the choice of the mean specification and the estimation technique, impacts the residual values' sensibly and thus the resulting variogram behavior, affecting the mean re-estimation and the final model outcome. We would rather use techniques that allow for the simultaneous estimation of both the first and second-order structure, looking for the best joint combination of the two and not conditionally on the *provisional* other. In this sense, *likelihood-based* methods provide us with a viable solution. These methods adhere to the *likelihood principle* and can be proved to satisfy certain optimality properties under mild conditions (see Chapter 4 of Gelfand et al. (2010)). As usual, the likelihood principle may be tackled both from a frequentist and a Bayesian approach. The general principles of the hierarchical Bayesian modeling framework are discussed in Section 2.3.

2.2 Spline regression

In this section, we investigate more deeply the method known as *spline regression*. This technique has been developed in the more general framework of *functional data analysis* (Ramsay and Silverman, 2007), where interest lies in inferring the unknown, true, underlying function from noisy observations.

Let us consider the function $f(\cdot) : \mathcal{U} \to \mathbb{R}$ as the objective of our estimation problem, where $\mathcal{U} \subseteq \mathbb{R}^p$. Even when $f(\cdot)$ is non-linear in its arguments, standard linear regression modeling would initially represent it as a linear function of these:

$$f(\boldsymbol{u}) \to \widetilde{f}(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{\beta}, \quad \boldsymbol{u} \in \mathcal{U}, \tag{2.2.1}$$

where $\boldsymbol{\beta}$ is a vector of coefficients to estimate in order to get $\widetilde{f}(\cdot)$ as the best *linear* approximation of to $f(\cdot)$. There is no need for $f(\cdot)$ to be actually linear in order to justify the linearity assumption, which is in any case a convenient representation for many reasons: estimation of the coefficients is straightforward in a linear regression setting; it allows for direct interpretation of the effect of $\boldsymbol{u}$ on $f(\cdot)$; it is avert of over-fitting issues because of its reduced *degrees of freedom*.

Nevertheless, in many cases, linearity is so far from reality that it does not provide a satisfactory approximation, and the need to express $f(\cdot)$ through a non-linear representation arises. This is particularly true in the case of spatial trends, that usually present (even steeply) varying behaviour in their domain and for which direct interpretation of the coordinates coefficients is not of primary interest. Indeed, in the spatial context, while an accurate estimation and representation of the surface is key, northing and easting effects are not of interest of their own (they represent unobserved spatial variability).

In Section 2.1.1 we introduced different methods that allow to move beyond linearity (parametric, semi-parametric and non-parametric). The core idea behind *spline regression* is to move beyond linearity by replacing the vector of arguments $\boldsymbol{u}$ with M transformations $h_m(\cdot) : \mathcal{U} \to \mathbb{R},\ m = 1, \dots, M$, and stick to the linear regression model setting in order to perform inference in the augmented space of input features:

$$f(\boldsymbol{u}) \leftarrow \widetilde{f}(\boldsymbol{u}) = \boldsymbol{h}(\boldsymbol{u})^\top \boldsymbol{\beta}, \qquad \boldsymbol{u} \in \mathcal{U}, \tag{2.2.2}$$

where $\boldsymbol{h}(\boldsymbol{u}) = [h_1(\boldsymbol{u}), \dots, h_M(\boldsymbol{u})]$.

Therefore, the representation problem reduces to choosing the most suitable set of function $\boldsymbol{h}(\cdot)$ for the problem at hand. According to reasonable assumptions on the behavior of $f(\cdot)$, we may want to use basis functions that span the functional space $\mathcal{F}$ to which it belongs. For instance, the Taylor expansions provides a rationale for the use of polynomials. However, as mentioned in Section 2.1.1, they have limitations both in theoretical and practical terms: their global nature causes local adaptation to have potentially large effects on the global behavior and the typically large correlation between the various terms make standard linear regression estimation unreliable. In order to overcome these issues, piece-wise polynomials may be considered; nevertheless, we argued how the resulting approximations are not guaranteed to satisfy continuity and, unless some tweaks are contemplated, are not smooth. A viable solution is instead represented by the so-called *spline functions* and their basis decomposition.

2.2.1 Spline functions

The origin of the term *spline* dates back to before World War II and has its roots in the aircraft and shipbuilding industries. Indeed, during that time, templates for airplanes were usually designed through a technique known as *lofting*, which consisted in passing thin wooden strips (called *splines*) through points laid down on the floor of a large design loft. These stripes provided an interpolation of the key points (called "ducks" in the engineering vocabulary) into smooth curves that, in particular, would produce the minimum strain energy. Splines became mathematical objects

later on, after proper formalization of their concept, and it is commonly accepted that the term spline in reference to smooth, piece-wise polynomial approximation of an objective function was first used by Schoenberg in Schoenberg (1946). They will then prove themselves useful in a number of other mechanical applications, among which the modeling of automobile design (Birkhoff and De Boor, 1965).

First, let us introduce what a spline function is in the uni-variate context. A uni-variate polynomial spline $h^o(\cdot)$ of order o (or degree $o-1$) is a real-valued function on a compact interval $[k_0, k_{K+1}] \subseteq \mathbb{R}$ defined through K so-called knots:

$$\mathcal{K} = \{k_1, \ldots, k_K : \; k_i < k_j, \; \forall \, i < j\},$$

where the terms k_0 and k_{K+1} are usually called *boundary knots*. Globally, it is a piece-wise polynomial function over the sub-intervals $\{[k_{i-1}, k_i]\}_{i=1}^{K-1}$, satisfying the following properties.

1. It is a polynomial of degree $(o/2-1)$ at the boundary sections, i.e. for $x \in \{[k_0, k_1], [k_K, k_{K+1}]\}$.
2. It is a polynomial of degree $(o-1)$ at each internal section, i.e. for $x \in \{[k_i, k_{i+1}] : \; i = 1, \ldots, K-1\}$.
3. It has continuous derivatives up to order $(o-2)$ on the whole domain $[k_0, k_{K+1}]$, i.e. $h^o(\cdot) \in \mathcal{C}^{o-2}$.

From properties 1. and 2. follows that a total of $(o \cdot K)$ coefficients are needed in order to define the whole spline function:

- $(o/2)$ to define it at the left of k_1;
- $(o/2)$ to define it at the right of k_K;
- $((K-1) \cdot o)$ are instead necessary to define it in the $(K-1)$ internal intervals.

The continuity conditions in 3. identify $((o-1) \cdot K)$ of the coefficients, while the remaining K can be uniquely identified only if the values assumed by the function at the K knots $f(k_1) = f_1, \ldots, f(k_K) = f_K$ (or at other K distinct points) are fixed.

In particular, as proved by the work by Schoenberg (Schoenberg (1964), Schoenberg (1988)), the use of spline functions has strong theoretical justification in the context of interpolation techniques. Let us consider the problem of finding:

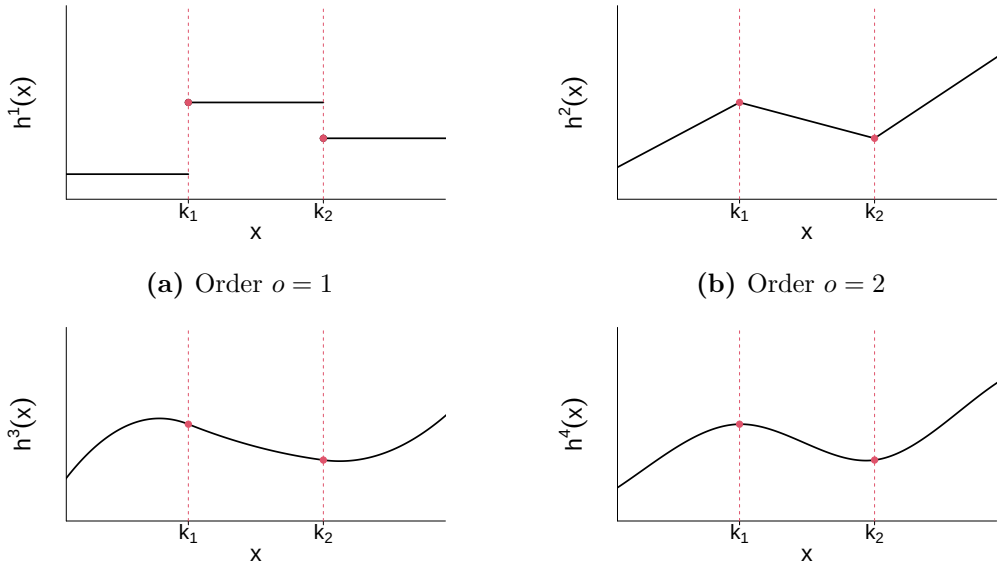
$$\hat{f}(\cdot) = \arg \min_{f(\cdot) \in \widetilde{\mathcal{F}}_o} \int_{k_0}^{k_{K+1}} f^{(o)}(x) dx \tag{2.2.3}$$

under:

$$f(k_i) = f_i, \quad \forall \, i = 1, \ldots, K,$$

where $\widetilde{\mathcal{F}}_o$ is the Sobolev space of functions with $o-1$ continuous derivatives and square integrable o-th derivative . He showed that, provided $K \geq o$, this minimizer actually is the *unique* polynomial spline of corresponding order satisfying the interpolation constraint. This is the property we referred to when at the end of Section 2.1.1 we mentioned that splines provide *the smoothest function possible given the assumed structure*.

However, for a statistician, this apparently amazing property is not of great use. Indeed, when the approximation of a function from partial observations is pursued in statistical applications, the exact values over a set of its arguments is rarely available.



(a) Order $o = 1$ **(b)** Order $o = 2$

(c) Order $o = 3$ **(d)** Order $o = 4$

Figure 2.3. Behavior in the vicinity of the two knots k_1 and k_2 of spline functions interpolating the same values, but with increasing order o.

More likely, the available observations $\boldsymbol{y} = [y_i]_{i=1}^n$ at locations $\boldsymbol{X} = [x_i]_{i=1}^n$ are perturbed with random noise. This makes exact interpolation undesirable, but the smoothness property of splines still stands as heavily appealing. Generally speaking, in the statistical framework observations are modeled in the following way:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where the ϵ_i's are independent zero-mean terms with common variance τ^2 and $f(\cdot)$ is known (or at least assumed) to be smooth. If we consider an analogous problem to the one of (2.2.3), but replace the interpolation constrain with:

$$f(\cdot) = \arg \min_{f(\cdot) \in \widetilde{\mathcal{F}}_o} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2, \tag{2.2.4}$$

it was again shown by Schoenberg that the minimizer is again a polynomial spline. In this setting, data are not interpolated anymore, but smoothed with approximation accuracy and smoothness degree depending on the chosen degrees of freedom of the spline (i.e. order and knots). In the sequel, we will introduce some practical tools to define and represent spline functions so that statistical estimation of the coefficients can be pursued in a linear regression setting. For further technical details about spline function properties and extensions (e.g. their connection to *Reproducing Kernel Hilbert Spaces*) the reader is referred to Wahba (1990).

2.2.2 Fixed-knot spline

If the knots are fixed in advance (e.g. at the observations locations), the resulting *fixed-knot splines* are also known as *regression splines*. This name is due to the fact that straightforward estimation of the coefficients can be obtained in a standard

linear regression setting. In order to define a regression spline one needs to select the number of knots, their location, and also the spline order. Figure 2.4 shows different examples of splines interpolating the same values over the same two knots, but with increasing orders. We can see how the order $o = 1$ spline is just a step-function, that does not even satisfy continuity at the knots (by property (4) it does not belong to $\mathcal{C}^0$). The order $o = 2$ spline of Figure 2.3b instead is piece-wise linear and satisfies continuity, but is not smooth. On the other hand, Figures 2.3c and 2.3d show two spline functions with orders $o = 3$ and $o = 4$, and the two satisfy both continuity and smoothness in terms of 1-st derivative continuity. The latter is known as *cubic-spline* (spline of order $o = 4$) and is usually claimed as the lowest-order spline for which the knot-discontinuity is not visible to the human eye. Therefore, there is rarely the need to use splines of order $o > 4$, unless continuity of higher order derivatives is deemed necessary for theoretical reasons. In fact, *cubic splines* are the most widely used in the literature (Hastie et al., 2017). But how can we represent a fixed-knot spline in a convenient way in order to perform the coefficient estimation?

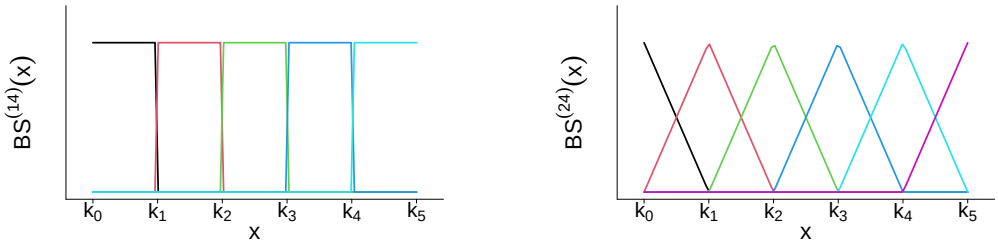
It can be shown that the space of spline functions of a particular order o and knot-sequence $k_0, \dots, k_{K+1}$, denoted as $\mathcal{F}_S^{o,K}$, is a vector space spanned by a basis of $J = (K+o)$ elements. From this perspective, splines are low-rank smoothers, because they are constructed from regression basis of smaller dimension than the number of observations. There are many equivalent bases that can be used to represent them. Among those we find: the *truncated power basis* (TPB), conceptually simple but not so simple from a numerical point of view; the Bernstein Polynomials (BP), computable through the *De Casteljau's algorithm* and way more numerically stable than the TPB; the B-Spline (BS) basis, that combines numerical stability with a structure that allows for efficient computations even when the number of knots K is large. Indeed, differently from its alternatives, the BS are built in such a way that the resulting spline is discontinuous at the boundary knots, thus undefined beyond the boundaries, and each element of the basis spans at most the support covered by o knots. Denoting with $\boldsymbol{B}^{o,k}(\boldsymbol{U})$ the $l \times J$ matrix of the basis elements evaluated at l points $\boldsymbol{U} = [\boldsymbol{u}_i]_{i=1}^{l}$, this *local support* property implies that many of its elements will be equal to 0. This can be profitably exploited to reduce the computational burden when the matrix $\boldsymbol{B}^{o,K}(\boldsymbol{U})$ is subject to matrix operations. Figure 2.3 shows different examples of BS basis functions with increasing order. In the sequel, including the Application of Chapter 3, the B-Spline basis is the standard choice as basis representation of a spline function. For more technical details about alternative spline basis, and in particular the B-Splines, the reader is referred to De Boor (1978); Bde Boor (2001).

Therefore, in order to reconstruct $f(\cdot)$ as a spline function given our observations $(\boldsymbol{y}, \boldsymbol{X})$, where we recall that $\boldsymbol{X}$ is a $n \times 1$ matrix (uni-dimensional case), we can simply express it as a linear combination of the BS basis elements evaluated at the $x_i,\ i = 1, \dots, n$ locations. Recalling Equation (2.2.2), this boils down to expressing at each support value x the approximation function as $\widetilde{f}(x) = \boldsymbol{B}^{(0,k)}(x)^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of unknown coefficients. Setting up the typical linear regression problem:

$$\widetilde{f}(\boldsymbol{X}) = \boldsymbol{B}^{(o,K)}(\boldsymbol{X})\,\boldsymbol{\beta} \;\leftarrow\; \boldsymbol{B}^{(o,K)}(\boldsymbol{X})\,\hat{\boldsymbol{\beta}} = \hat{f}(\boldsymbol{X}) \tag{2.2.5}$$

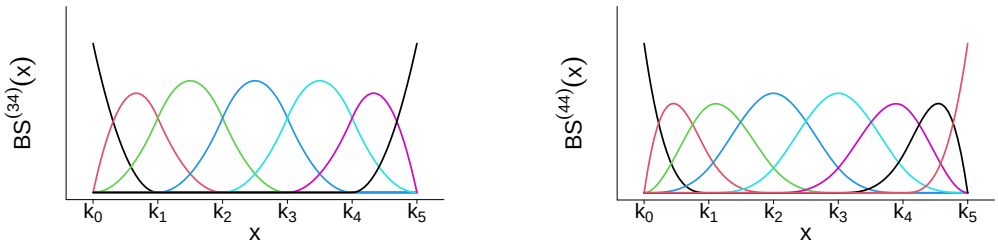
where $\boldsymbol{B}^{(o,K)}(\boldsymbol{X}) = \left[\boldsymbol{B}^{(o,K)}(x_1)^\top, \dots, \boldsymbol{B}^{(o,K)}(x_1)^\top\right]^\top$ and $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$ obtained through any valid fitting procedure such as *Ordinary Least squares* (OLS), or *Maximum Likelihood Estimation* (MLE).

The order of the basis and the number and positioning of the knots determine the flexibility and smoothness of the resulting approximating function. We already



(a) Order $o = 1$

(b) Order $o = 2$



(c) Order $o = 3$

(d) Order $o = 4$

Figure 2.4. BS basis functions at 4 evenly spaced knots between the boundaries k_0 and k_5, with increasing orders o

argued in favor of the use of an order $o = 4$ basis. However, there is not a generally valid criteria to determine the optimal choice of the knots. If knots overlapping with the observations set $\boldsymbol{X}$ were chosen, then perfect interpolation would be achieved, capturing noise together with information. On the other hand, a too little number of knots would produce a too rough approximation with behavior between the knots governed more by the super-imposed mathematical structure than by the information in the data. In general, the choice of the number of knots is really problem dependent and concerns the hypothesis of the researcher on the behavior of the true underlying function $f(\cdot)$. Nevertheless, there is some general guidance on their positioning. For instance, one may consider evenly spaced knots between the two boundaries; this choice is reasonable when the function is assumed to have similar variations on its whole domain. Alternatively, one may consider placing the knots at quantiles of evenly spaced levels between 0 and 1 of the location set; the implied assumption is that the function is allowed to present a wigglier behavior where more information is available, while it is forced to have less freedom in sections where there is less information.

All the aforementioned recommendations are nothing more than *thumb rules*, quite arbitrary and especially not automatic. The knot selection issue is still an open problem, usually addressed resorting to model comparisons via validation approaches. Some automatic strategies have been introduced in order to reduce its importance and relevance to estimation purposes. One possible solution is to consider the execution of the fitting procedure (2.2.5) via suitable/*ad-hoc* penalized regression approaches.

2.2.3 Bayesian P-Spline

The standard approach to linear regression is to estimate $\widetilde{f}(\cdot)$ simply by minimizing the squared deviations of the approximating function from the observations.

More generally, including also *Generalized Linear Model* settings, this encompasses the maximization of the log-likelihood $l(\widetilde{f}|\boldsymbol{y}, \boldsymbol{X})$. Thus, $\hat{f}$ is usually determined as:

$$\hat{f} = \arg \max_{\widetilde{f} \in \mathcal{F}_S^{o,K}} l(\widetilde{f} \,|\, \boldsymbol{y}, \boldsymbol{X}). \tag{2.2.6}$$

However, while theoretically well-grounded, this estimation procedure is very prone to over-fitting and this issue is exacerbated in the context of *spline regression*. Indeed, if a large enough set of knots is chosen, $\hat{f}(\cdot)$ is flexible enough to account for all the variations in the data and interpolates all the observed values $\boldsymbol{y}$ incorporating information and noise. Different methods have been proposed in the literature to help in the choice of degree and knots, mainly based on cross-validation approaches (Wood, 2017; Hastie et al., 2017).

Here, we introduce a method that controls for the complexity of the fit through regularization, where the penalty is specifically designed to penalize for steep variations of the resulting spline function. This allows to avoid the knot selection problem and, if desired, even to choose the maximal set of knots without achieving interpolation. A penalty term is added to the optimization criteria, as in the case of the *Akaike Information Criterion* (AIC), the *Bayesian Information Criterion* (BIC) or the *Deviance Information Criterion* (DIC) (Burnham and Anderson, 2004; Gelman et al., 2013; Spiegelhalter et al., 2002). More in general, penalized/regularized regression replaces the fitting criterion of Equation (2.2.6) with:

$$\hat{f} = \arg \max_{\widetilde{f} \in \mathcal{F}_S^{o,K}} \left\{ l\left(\widetilde{f} \,|\, \boldsymbol{y}, \boldsymbol{X}\right) - \lambda \cdot \text{pen}\left(\widetilde{f}\right) \right\}. \tag{2.2.7}$$

where $\text{pen}(\cdot)$ is a penalty function on the complexity of $\widetilde{f}(\cdot)$, and $\lambda > 0$ establishes the trade-off between goodness of fit (first term) and penalization (second term). For $\lambda = 0$ the criterion (2.2.7) is equivalent to (2.2.6), while for $\lambda \to \infty$ the estimation process always returns the simplest possible version (e.g. linear).

In the spline regression setting, the driving idea is to penalize large variations of the spline functions at the knots. Therefore, the penalty term shall represent a roughness measure for the function estimate and complexity shall be measured in terms of function *non-smoothness*. This is practically equivalent to the problem of Equation (2.2.4) under condition (2.2.3) solved by Schoenberg (1988), with the difference that instead of looking for the minimum error solution in the infinite-dimensional Sobolev function space $\mathcal{F}$, we look for the MLE solution in the restricted space of spline functions $\mathcal{F}^{o,K}$. When using B-splines as a basis, this kind of penalty can be defined simply in terms of differences of coefficients referred to adjacent basis elements. This yields the following:

$$\lambda \cdot \text{pen}(\widetilde{f}) = \lambda \cdot \sum_{j=m+1}^{J} \left(\delta^m \beta_j\right)^2, \quad \widetilde{f} \in \mathcal{F}_S^{(o,K)}, \tag{2.2.8}$$

where λ is called *smoothing parameter*, m is the derivative order for which deviations from smoothness are penalized and:

$$\begin{aligned}
\delta^1 \beta_j &= \beta_j - \beta_{j-1}, \\
\delta^2 \beta_j &= \delta^1\left(\delta^1 \beta_j\right) = \beta_j - 2\beta_{j-1} + \beta_{j-2}, \\
&\vdots \\
\delta^m \beta_j &= \delta^{m-1} \beta_j - \delta^{m-1} \beta_{j-1}.
\end{aligned}$$

Writing the whole penalty term in matrix notation, the following formulation of (2.2.8) is obtained:

$$\lambda \cdot \text{pen}(\widetilde{f}) = \lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{D}_m^\top \boldsymbol{D}_m \boldsymbol{\beta} = \lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{M}^{(m)} \boldsymbol{\beta}, \qquad \widetilde{f} \in \mathcal{F}_S^{(o,K)}, \tag{2.2.9}$$

where $\boldsymbol{M}^{(m)}$ is the *penalization matrix* derived from the differences matrices $\boldsymbol{D}_m$. Each of them can be derived through the recursive formula $\boldsymbol{D}_m = \boldsymbol{D}_{m-1}\boldsymbol{D}_1$. Replacing it in the maximization criterion of (2.2.7) the following is obtained:

$$l(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) - \lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{M}^{(m)} \boldsymbol{\beta}. \tag{2.2.10}$$

For a given value of the smoothing parameter λ, inference on the coefficients $\boldsymbol{\beta}$ and at any arbitrary set of points $\boldsymbol{Z}$ can be performed in terms of this penalized criterion using:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{M}^{(m)}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\
\boldsymbol{H}_{\lambda,m} &= \boldsymbol{Z}\left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{M}^{(m)}\right)^{-1} \boldsymbol{Z}^\top \\
\hat{f}_{\lambda,m} &= \boldsymbol{H}_{\lambda,m} \boldsymbol{y},
\end{aligned}$$

where $\boldsymbol{H}_{\lambda,m}$ is known as the *hat matrix*. However, this solution is feasible only when least square estimation is the objective (that coincides with MLE estimation only for Gaussian outcomes), and for λ fixed to some specific value. Usually, this value is chosen through cross-validation on some metric (or any other model comparison methods) after estimation has been performed on a grid of reasonable values for it (Lewis and Stevens, 1991; Wood, 2017; Hastie et al., 2017).

However, looking carefully to Equation (2.2.10), we can notice how this is equivalent to performing Bayesian inference on the $\boldsymbol{\beta}$ coefficients, while using m-order Normal random walk priors on these. Indeed, we may ascribe to the coefficients a prior:

$$\pi(\boldsymbol{\beta}) = \exp\left\{-\lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{M}^{(m)} \boldsymbol{\beta}\right\} \propto \mathcal{N}_J\left(\boldsymbol{\beta} \,\middle|\, \boldsymbol{0},\, \frac{1}{\lambda}\left(\boldsymbol{M}^{(m)}\right)^{-1}\right)$$

where $\lambda \cdot \boldsymbol{M}^{(m)}$ is the precision matrix of the Gaussian prior, and get inference in the same penalized framework of (2.2.7). For instance, the first order penalty matrix is:

$$\boldsymbol{M}^{(1)} = \begin{bmatrix}
1 & -1 & 0 & \cdots & 0 \\
-1 & 2 & -1 & & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & & -1 & 2 & -1 \\
0 & \cdots & 0 & -1 & 1
\end{bmatrix}$$

and it would correspond to the following set of conditional priors:

$$\begin{aligned}
&\pi_{\beta_1}(\cdot) \propto const \\
&\pi_{\beta_j}\left(\cdot \,|\, \boldsymbol{\beta}_{1:j-1}\right) = \mathcal{N}\left(\cdot \,\middle|\, \beta_{j-1},\, \frac{1}{\lambda}\right), \qquad j = 2, \ldots, J.
\end{aligned}$$

The resulting joint prior is a partially improper multivariate Gaussian distribution since, for any choice of λ and m, we have that $\text{rank}\left(\boldsymbol{M}^{(m)}\right) < \dim(\boldsymbol{\beta})$. This is practically never an issue in terms of posterior inference, since the posterior will be

proper nonetheless. Another great advantage of the Bayesian approach is that we may also ascribe a suitable prior to the smoothing parameter λ, that in this context is a hyper-parameter representing the overall precision of our prior on $\boldsymbol{\beta}$, and make inference on its value inside the same *Markov Chain Monte-Carlo* machinery used to get posterior samples of the other parameters. A practical implementation of this approach in a more general model is presented in Chapter 3, Section 3.2.4.

2.2.4 Multi-dimensional splines

The initial part of Section 2.2 is focused on spline models for data on a one-dimensional domain. In the next section, a brief introduction to spline models for multidimensional domains is provided.

The theoretical generalization of spline function to domains of arbitrary dimension $d > 1$ is somewhat natural. Multidimensional splines on $\mathbb{R}^d$ are indeed determined by the straightforward extension of the smoothness criteria in (2.2.3). For example, for $\mathcal{U} \in \mathbb{R}^2$, it becomes:

$$\int\int_{\mathbb{R}^2} \left[\left(\frac{\delta^2 f(u)}{\delta u_1^2} \right)^2 + \left(\frac{\delta^2 f(u)}{\delta u_1 \delta u_2} \right)^2 + \left(\frac{\delta^2 f(u)}{\delta u_2^2} \right)^2 \right] du_1 du_2, \tag{2.2.11}$$

that combined with condition (2.2.4) has a solution in the two dimensional surfaces known as *thin-plate splines* (Wood, 2003). As their one-dimensional counterpart, these can be represented through linear combination of basis functions:

$$\widetilde{f}(\boldsymbol{u}) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{u} + \boldsymbol{h}(\boldsymbol{u})^\top \boldsymbol{\alpha}, \quad \boldsymbol{u} \in \mathcal{U}, \tag{2.2.12}$$

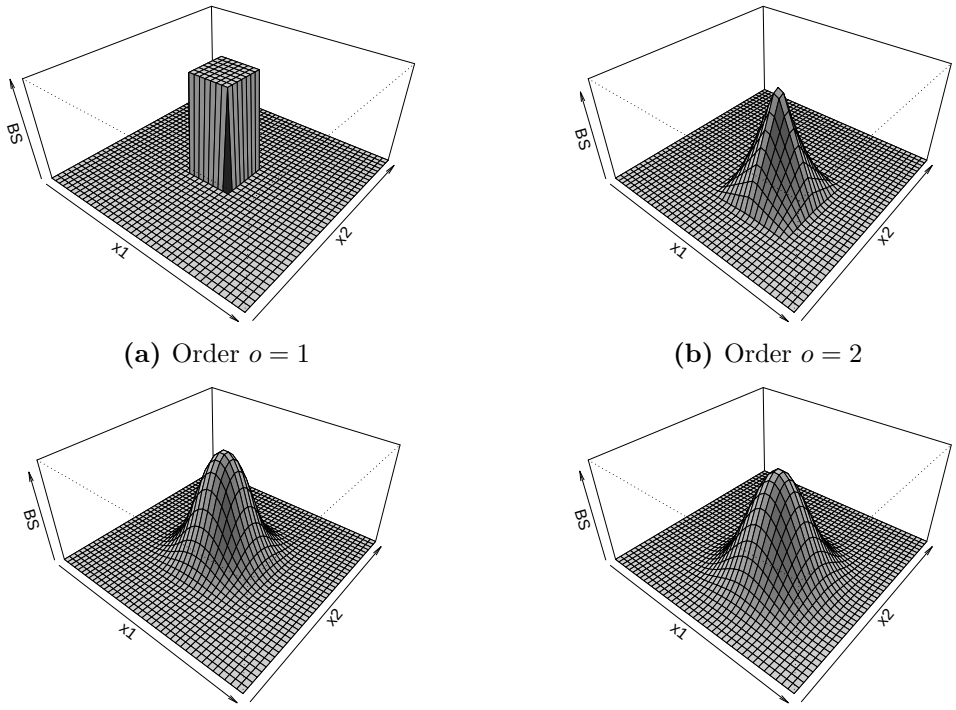
where β_0 is a constant real term and $\boldsymbol{\beta}$ is a $d \times 1$ vector of coefficients and $\boldsymbol{h}(\boldsymbol{u}) = [h_1(\boldsymbol{u}), \ldots, h_M(\boldsymbol{u})]$ the vector of basis elements in $\boldsymbol{u}$. In particular, $\boldsymbol{h}(\boldsymbol{u})$ is a set of *radial basis functions* built on the set of knots $\{\boldsymbol{u}_j\}_{j=1}^M$. It can actually be defined for any arbitrary dimension d, and one possible example for $d = 2$ has as generic element $h_j(\boldsymbol{u}) = ||\boldsymbol{u} - \boldsymbol{u}_j|| \log(||\boldsymbol{u} - \boldsymbol{u}_j||)$.

Once the thin-plate-spline structure has been imposed by expressing the surface through Equation (2.2.12), the estimation reduces to the usual finite-dimensional (eventually penalized) least-squares problem. However, differently from B-spline basis[6], the elements of a radial basis do not present any exploitable sparse structure. In particular, the complexity cost of their evaluation scales with the order $\mathcal{O}(M^3)$, making the problem computationally intensive also for M of moderate size. On top of that, approximation performances decay rapidly for a knot sequence too little for the sample size, deterring from choosing a $M << n$ and hindering the potential of such representation.

Tensor product B-splines Unfortunately, there is not an exact correspondent to the B-Spline basis in the multi-dimensional setting, and the sole reliance on thin-plate splines is gruesome for applications with moderate to large data sizes.

An appealing alternative to *thin-plate spline* stems from the concept of *tensor product basis*. Let us introduce this basis construction method for $d = 2$. Suppose $\mathcal{U} \in \mathbb{R}^2$ and let $\boldsymbol{h}_1(\cdot)$ of size J_1 and $\boldsymbol{h}_2(\cdot)$ of size J_2 be two (arbitrary) sets of basis functions defined independently on the two dimensions. Then, the $J = J_1 \times J_2$ dimensional tensor product of the two basis $\boldsymbol{H}^{(2)}(\cdot) = \boldsymbol{h}_1(\cdot) \otimes \boldsymbol{h}_2(\cdot)$ defines the

[6]Recall that most of the elements of the basis are null for large sections of the domain



(a) Order $o = 1$ **(b)** Order $o = 2$

(c) Order $o = 3$ **(d)** Order $o = 4$

Figure 2.5. Generic element of the *tensor product B-Spline* for increasing orders $o_1 = o_2 = o$.

corresponding *tensor product basis*. For a fine enough grid, it is possible to represent any smooth two-dimensional function on $\mathcal{U}$ as:

$$\widetilde{f}(\boldsymbol{u}) = \sum_{j=1}^{J_1}\sum_{l=1}^{J_2}\beta_{jl}H_{jl}^{(2)}(\boldsymbol{u}) = \sum_{j=1}^{J_1}\sum_{l=1}^{J_2}\beta_{jl}h_{1j}(\boldsymbol{u})h_{2l}(\boldsymbol{u}), \quad \boldsymbol{u}\in\mathcal{U},$$

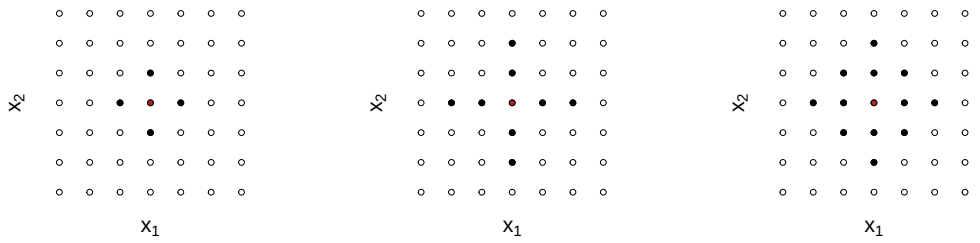
where $H_{jl}^{(2)}(\boldsymbol{u}) = \left[\boldsymbol{H}^{(2)}(\boldsymbol{u})\right]_{jl}$. Applying the same reasoning to B-Spline basis on the bidimensional domain, we get the so called *tensor product B-Spline*. This allows to keep the useful and desirable B-spline properties in more than one dimension, while approximating well at will the thin-plate spline solution for fine grids on rectangular domains. On the downside, the resulting basis is not radial (i.e. contours are not given by circles), and therefore is not invariant under rotation of the space coordinates. Nevertheless, for a large enough degree, the non-radiality almost disappears and (generally) does not represent much of an issue for degree > 2 (order > 3; Jie et al. (2016); Hastie et al. (2017)). Examples of the generic element of the 2-dimensional *tensor product B-Spline* for different orders are provided in Figure 2.5.

It is then possible to represent (with good approximation) the spline function as:

$$\widetilde{f}(\boldsymbol{X}) = \boldsymbol{B}^{(2)}(\boldsymbol{X})\,\boldsymbol{\beta} \;\leftarrow\; \boldsymbol{B}^{(2)}(\boldsymbol{X})\,\hat{\boldsymbol{\beta}} = \hat{f}(\boldsymbol{X}) \tag{2.2.13}$$

where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$ and $\boldsymbol{B}^{(2)}(\cdot) = \boldsymbol{B}^{(o_1,K_1)}(\cdot)\otimes\boldsymbol{B}^{(o_2,K_2)}(\cdot)$ and:

$$\boldsymbol{B}^{(2)}(\boldsymbol{X}) = \left[\boldsymbol{B}^{(2)}(\boldsymbol{x}_1),\ldots,\boldsymbol{B}^{(2)}(\boldsymbol{x}_n)\right]^\top.$$



(a) Order $v = 1$ **(b)** Order $v = 2$ **(c)** Order $v = 2$

Figure 2.6. Neighborhood examples for knots spanned on regular 2-dimensional lattices, where the last includes diagonal neighbors

Estimation can be performed in the usual linear regression setting and the *Generalized Array Method* of Currie et al. (2006) offers an efficient implementation that exploits all the sparsity in $\boldsymbol{B}^{(2)}(\boldsymbol{X})$. Obviously, the choice of the orders o_1 and o_2 and the number and placement of the knots $\{k_{ij} : i = 1, 2,\ j = 0, \ldots, K_i\}$ play a key role in the structure of the resulting basis. An order $o \geq 3$ ($o = 4$ especially) is strongly recommended, while general guidance for knots placement is analogous to the uni-dimensional case. In particular, the choice of evenly spaced knots along the two dimensions correspond to regular lattices over $\mathbb{R}^2$ and presents a very convenient geometric structure.

Penalized approaches Also in this case the problem of knots placement has not a general answer and resorting to penalized approaches is highly recommended (Wood, 2006). The extension of B-Spline penalization in the bi-variate case requires a proper definition of the neighborhood structure of the knots and how a penalty based on such neighborhoods can be computed. When the knots are placed on regular lattices, neighborhoods of different orders v can be defined by taking the v closest knots along horizontal and vertical directions or even considering the diagonal. A graphical example is provided in Figure 2.6.

Let us first introduce the extension of first-order differences based on the four nearest neighbors (Figure 2.6a). Let $\boldsymbol{D}_1^{(1)}$ and $\boldsymbol{D}_2^{(1)}$ be the univariate first-order differences matrices in the two directions. The sum of the row-wise differences in the lattice of coefficients can be computed by applying the blown-up difference matrix $\boldsymbol{I}_{J_2} \otimes \boldsymbol{D}_1^{(1)}$ to the vector $\boldsymbol{\beta}$, i.e.

$$\boldsymbol{\beta}^\top \left(\boldsymbol{I}_{J_2} \otimes \boldsymbol{D}_1^{(1)}\right)^\top \left(\boldsymbol{I}_{J_2} \otimes \boldsymbol{D}_1^{(1)}\right) \boldsymbol{\beta} = \sum_{j=1}^{J_1} \sum_{l=1}^{J_2} (\beta_{jl} - \beta_{j-1,l})^2,$$

where $\boldsymbol{I}_\cdot$ is the identity matrix of size $\cdot$ and $\otimes$ is the Kronecker product. From basic matrix algebra $\left(\boldsymbol{I}_{J_2} \otimes \boldsymbol{D}_1^{(1)}\right)^\top \left(\boldsymbol{I}_{J_2} \otimes \boldsymbol{D}_1^{(1)}\right) = \left(\boldsymbol{I}_{J_2} \otimes \boldsymbol{D}_1^{(1)\top} \boldsymbol{D}_1^{(1)}\right)$ where $\boldsymbol{D}_1^{(1)\top} \boldsymbol{D}_1^{(1)}$ would be the univariate penalty matrix. Similarly, the column-wise differences can be obtained as:

$$\boldsymbol{\beta}^\top \left(\boldsymbol{D}_2^{(1)} \boldsymbol{D}_2^{(1)\top} \otimes \boldsymbol{I}_{J_1}\right) \boldsymbol{\beta} = \sum_{j=1}^{J_1} \sum_{l=1}^{J_2} (\beta_{jl} - \beta_{j,l-1})^2 .$$

Finally, the two penalties along the two directions can be combined element by element by taking their matrix sum. This yields the following form for the penalty

matrix:

$$\boldsymbol{M}_{\lambda}^{(1)} = \lambda \left(\boldsymbol{I}_{J_2} \otimes \boldsymbol{M}_1^{(1)} + \boldsymbol{M}_2^{(1)} \otimes \boldsymbol{I}_{J_1} \right), \tag{2.2.14}$$

where $\boldsymbol{M}_1^{(1)} = \boldsymbol{D}_1^{(1)^\top} \boldsymbol{D}_1^{(1)}$ and $\boldsymbol{M}_2^{(1)} = \boldsymbol{D}_2^{(1)^\top} \boldsymbol{D}_2^{(1)}$ are the univariate penalty matrices and λ has been collected inside the final penalty matrix for convenience. Allowing for more than first-order difference penalty on the horizontal and vertical directions can be straightforwardly obtained just by replacing the univariate penalty matrices with higher-order penalty matrices $\boldsymbol{M}_1^{(m_1)}$ and $\boldsymbol{M}_2^{(m_2)}$, yielding $\boldsymbol{M}_{\lambda}^{(m_1,m_2)}$.

The final penalty is then equivalent to the one of (2.2.9) in the uni-variate case, where the resulting penalty matrix must be pre-multiplied by a smoothing parameter λ. Thus, all remarks about the estimation of the coefficients according to the maximization criteria in (2.2.10) hold unchanged.

As an addendum, nothing precludes the consideration of different smoothing parameters on different directions, which would introduce *anisotropy* in the degree of smoothness. This can be achieved by replacing the simple sum of Equation (2.2.14) by a *weighted sum* of the two penalty matrices with weights $\lambda_1, \lambda_2 \in \mathbb{R}^+$ (smoothing parameters), as follows:

$$\boldsymbol{M}_{\lambda}^{(m_1,m_2)} = \left(\lambda_1 \cdot \boldsymbol{I}_{J_2} \otimes \boldsymbol{M}_1^{(m_1)} + \lambda_2 \cdot \boldsymbol{M}_2^{(m_2)} \otimes \boldsymbol{I}_{J_1} \right). \tag{2.2.15}$$

That can be easily extended to dimensions $d > 2$ by taking multiple Kroenecker products on various bases and then summing the penalizations over the various neighborhood structures. However, smoothing in several dimensions is susceptible to runaway problems with storage and computational time. Arranging the long basis vectors into matrices/arrays allows for the implementation of efficient estimation algorithms known as *Generalized Linear Array Models*, first developed in Currie et al. (2006); Eilers et al. (2006), that strongly reduce the required computational time. That implies a straightforward and insightful interpretation of penalized tensor-product B-splines in terms of mixed-effect ANOVA models (Gu, 2002).

2.3 Bayesian Hierarchical modeling of Spatial Processes

An accurate and comprehensive analysis of spatial processes requires the ability to describe properly their convoluted structure, characterized by a multitude of uncertainty factors and various degrees of scientific knowledge. In the attempt of considering the various components from a *joint* perspective, the complexity of the problem at hand may seem overwhelming and too challenging for designing a well-suited but manageable statistical model. When this is the case, it may be much easier to tackle it from a conditional perspective, approaching the model specification at multiple levels inserted in a nested structure. Indeed, if the intricate dependence structure among the various elements at play is expressed through a convenient graphical model (Wainwright and Jordan, 2008), the originally complex joint distribution can be straightforwardly factored in (typically) simpler conditional densities. For instance, it may be very difficult to specify a meaningful and valid joint multivariate dependence structure between the spatial distribution of rain amount and forestry in a specific area. However, if we can assume the rain to impact on the forestry and not viceversa, it is way simpler to consider the forestry distribution conditionally on the rain amount rather than dealing with the two processes jointly.

This is the essence of hierarchical modeling that, theoretically, is not exclusive of either the classical or the Bayesian approach. However, the Bayesian paradigm founds itself on the ability to integrate information from different sources (i.e. the

sub-models) while accounting for all the uncertainty accumulated at each level. Doing the same in a frequentist framework is cumbersome, or even unfeasible, if one wants to stay true to the frequentist principles. The Bayesian perspective becomes a necessity if the level of complexity is high enough, which is usually the case in spatial statistics application. Its wide adoption, originally limited by the lack of computational power, has been growing steadily ever since the revolution in Bayesian computation linked to the development of effective and efficient *Markov Chain Monte Carlo* (MCMC) techniques (Gelfand and Smith, 1990; Robert and Casella, 2013). Over time the most of the literature has been developed in the Bayesian framework (Robert, 2007; Gelman et al., 2013), and this tendency does not seem to be slowing down.

In this section, we provide a brief explanation of the general principles of Bayesian hierarchical modeling, with focus on its use in the context of spatial modeling as for the course plotted by Berliner (1996) and Wikle and Cressie (1999). An extensive review of the *Bayesian hierarchical modeling for spatial data* with application in `R` and `WinBugs`[7] is available in Banerjee et al. (2014). Many models are also directly implemented in the `R` package `spBayes` (Finley et al., 2007).

2.3.1 Bayesian Hierarchical Modeling

While the hierarchical specification of a model is a bit of a stretch under the frequentist paradigm, the *hierarchical structure* is something intrinsic to the Bayesian approach, even in its simplest form. When specifying a Bayesian model, the parameters governing the *data generative process* are *random variables* and the whole uncertainty of the phenomenon must be quantified in terms of the joint distribution of data ($data$) and parameters ($pars$). In particular, inference is based on the so-called *posterior distribution* of the $pars$ given the $data$. By applying the Bayes' Theorem, this can always be expressed as:

$$\pi(pars \,|\, data) = \frac{J(data,\, pars)}{m(data)} \propto \mathcal{L}(data \,|\, pars) \cdot \pi(pars), \tag{2.3.1}$$

where $J(data,\, pars)$ is the *joint distribution* of $data$ and $pars$; $m(data)$ is the *marginal likelihood* of $data$ given the model (irrelevant to the end of parameters' inference); $\mathcal{L}(data \,|\, par)$ is the *data likelihood*; $\pi(par)$ is the *parameters' prior*.

The left-hand side of Equation (2.3.1) highlights the strong and indissoluble link between the posterior distribution $\pi(cdot \,|\, cdot)$, which is the final inferential objective, and the joint distribution $J(\cdot, \cdot)$. The direct specification of a joint distribution coherent with any experimental setting is simply impractical, and generally has little to no guidance in any Bayesian setting. Nevertheless, from the right-hand side of Equation (2.3.1), we can see how it can be factored in two relevant terms: $\mathcal{L}(data \,|\, par)$ and $\pi(pars)$. The former regulates the generative process of the data given the parameters, while the latter contains all our *a-priori* knowledge on the parameters defining the model. In contrast with the joint distribution, the separate specification of these two components is straightforward: the *likelihood* leads the choice as to what best represents the data; the prior usually follows according to convenient combinations with the Likelihood, while keeping into account eventually available prior information. In practice, these can be specified in a hierarchy of (at least) two levels:

[7] WinBUGS is a software developed as part of the BUGS project, which aims to make practical MCMC methods available to applied statisticians. URL: `https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/`

- **Level 1**: $\mathcal{L}(data \,|\, pars)$;
- **Level 2**: $\pi(pars)$;

where each level may have additional sub-levels.

The consideration of such a hierarchical structure has been a real game-changer in statistical modeling for complicated process, among which spatial ones. Indeed, specifying elaborate dependence structure directly on the outcome of interest itself (*data*) has several drawbacks: it makes it hard disentangling structured variance and covariance terms from the unstructured random noise; further it sensibly complicates the data likelihood specification, especially when the *Gaussianity* assumption cannot hold for the outcome of interest.

A *hierarchical* specification of the statistical model allows to keep the independence assumption at the *data* level, while transferring the dependence structure onto a latent process *proc* at a deeper level of the hierarchy (allowing for a potentially arbitrary choice of its distribution form). This paves the way to a conceptual scheme that greatly simplifies the problem, well-summarized by the following three levels skeleton (e.g. Berliner (1996)):

- **Level 1**: $\mathcal{L}(data \,|\, proc, pars)$;
- **Level 2**: $\pi(proc \,|\, pars)$;
- **Level 3**: $\pi(pars)$.

This framework decomposes a complicated generative process into three primary components that are all linked by simple probability rules. This partitioning allows to specify way simpler models at each stage that, when combined, can describe very complex joint data, process and parameters distributions.

Obviously, the ultimate interest lies in performing inference on the model parameters, on the latent process (not always) and on the distribution of the outcome at un-sampled locations of the *data* space. Generally speaking, all this can be pursued in terms of the parameters' and process' *posterior distributions*:

$$\pi(proc\,,\, pars \,|\, data) \propto \mathcal{L}(data \,|\, proc,\, pars) \cdot \pi(proc \,|\, pars) \cdot \pi(pars) \tag{2.3.2}$$

and of the *posterior predictive distribution*:

$$\pi(\widetilde{data} \,|\, data) = \int \pi(\widetilde{data} \,|\, proc,\, pars,\, data) \cdot \pi(proc\,,\, pars \,|\, data)\, dpars\, dproc, \tag{2.3.3}$$

where $\widetilde{data}$ denotes new, unobserved data. This comes with computational concerns, which usually suggest convenient specifications of the latent process and parameters' priors. Nevertheless, aside from special and rare cases, Equations (2.3.2) and (2.3.3) lack a closed form solution and their computation relies on simulation techniques known as *Markov Chain Monte Carlo* (Gelfand and Smith, 1990; Robert and Casella, 2013). Hereby, we will not provide a thorough description of these methods in a general framework, and this concludes the very brief introduction to Bayesian statistics principles and hierarchical modeling. We merely scratched the surface of the Bayesian potential and it has to be intended just as an introduction to the remainder of this Chapter, which will be focused on the Bayesian hierarchical specification of strictly spatial models. For more details about this approach and its theoretical bases the author recommends Robert (2007); Gelman et al. (2013).

2.3.2 Hierarchical Spatial Modeling

Let us denote with $Y(\cdot) : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D} \in \mathbb{R}^d$, a uni-variate spatial process[8] of which we have observed a realization at a finite set of locations. As argued at the beginning of Section 2.1, we would like to use the information in $\boldsymbol{y}$ in order to perform inference on the inherent stochastic structure of $Y(\cdot)$. In order to do so, we need to specify a model (hence a distribution) for $Y(\cdot)$ at any set of locations $\mathcal{U} \in \mathcal{D}$. This model shall take into account at the same time a great variety of factors: the nature of the observed data (e.g. their support) and the shape of their distribution, the effects of known external factors, the presence of variability and measurement errors, the spatial dependence structure, etc.

The marginal specification of a model for $Y(\cdot)$ through $f_Y(\cdot)$ is a very complicated (if not impossible) task that requires some sort of simplification if a viable solution is deemed. The hierarchical approach introduced in the previous section provides a straightforward way to produce a general model entangling all the needed complexity, but based on the specification of simpler distributions at different levels of the hierarchy. Indeed, is is possible to introduce an additional latent process $\epsilon(\cdot)$ defined on the same space and unload the most of the complicated dependence structure present in $Y(\cdot)$ on it. As a result, the distribution of the data given the latent process $f_Y(\cdot \,|\, \epsilon)$ is simplified by the conditional independence assumption, and its specification shall be focused on reflecting the data nature and distribution. At the same time, modeling of $\epsilon(\cdot)$ is not bound to any constraint and the choice of $f_\epsilon(\cdot)$ shall be led just by the convenient specification of the underlying dependence structure.

Typically, the two chosen distributions depend on a set of parameters ω (in a common parametric or semi-parametric setting). As for the Bayesian principles, also the parameters shall be ascribed a distribution $\pi_\omega(\cdot)$ (i.e. the *prior* distribution). This distribution is usually specified under the reasonable independence assumption between data parameters and process parameters, so that the joint distribution can be factored as $\pi_\omega(\cdot) = \pi_{\omega_Y}(\cdot) \cdot \pi_{\omega_\epsilon}(\cdot)$, where ω_Y and ω_ϵ are respectively the subset of parameters on which $f_Y(\cdot \,|\, \epsilon, \omega_Y)$ and $f_\epsilon(\cdot \,|\, \omega_\epsilon)$ depend on. However, aside from that, the prior specification is really model-dependent and has few general guidelines. In many cases, these are chosen mainly (or partly) to facilitate computations and, conditionally on that, to represent eventually available prior knowledge on the phenomenon under analysis.

This last piece fills the third and final level in our hierarchical spatial framework, leading to the following three levels hierarchy.

- **Level 1**. Observable model $Y(\cdot) \sim f_Y(\cdot \,|\, \epsilon, \omega_y)$;
- **Level 2**. Process model $\epsilon(\cdot) \sim f_\epsilon(\cdot \,|\, \omega_\epsilon)$;
- **Level 3**. Parameters model $\theta \sim \pi_{\omega_y}(\cdot) \cdot \pi_{\omega_\epsilon}(\cdot)$.

This general framework can accommodate data with different support and/or alignment than the underlying process $\epsilon(\cdot)$ (Gelfand et al., 2001; Wikle et al., 2001; Wikle and Berliner, 2005). Actually, the real strength and power of the hierarchical approach resides especially in the first of these two considerations. Indeed, for the reasons given in Section 2.1, the specification of spatial dependence structure is way easier under the Gaussianity assumption. Therefore, reliance on a latent Gaussian

[8] For ease of notation and explanation, we will always refer to uni-variate spatial processes. Generalization to q-variate processes is straightforward under suitable conditional independence assumptions and for further details the reader is referred to Chapter 7 of Banerjee et al. (2014)

process is very common (if not necessary) in most applications. Luckily, this does not preclude from specifying alternative distributions for the observable process $Y(\cdot)$ conditionally on the latent process.

As a matter of fact, there are many real-world processes in which data cannot certainly be assumed to be Gaussian, neither well-approximated by it, but yet exhibit spatial dependence. Let us think to count data, very common in many biological, ecological and environmental studies, that are defined only on discrete and positive values and present an indiscernible mean-variance relationship. Ignoring these structures and blindly considering the Gaussian distribution can lead to inefficiency and bias in both estimation and prediction. Instead, it would be natural to consider a Poisson or Negative Binomial distribution in such situations. Therefore, once the dependence structure has been properly accounted for in the Gaussian latent process, non-Gaussian spatial data may be formally analyzed within this same framework simply by specifying the proper conditional model for the data. Dependence in the latent process can then be transferred to the observables through a suitable link function (on the mean for instance), as within the context of *Generalized Linear Mixed Models* (McCulloch and Searle, 2001; Bradley et al., 2016b, 2020). Here, for the sake of brevity, we will not navigate into further details on the spatial modeling of non-Gaussian data. Starting from the next section, we will delve into the hierarchical specification of the *Geo-statistical model* for Gaussian data. However, for further details about the *generalized* version of the same model, the reader is referred to Diggle and Ribeiro Jr (2007) and Banerjee et al. (2014).

2.3.3 The hierarchical Gaussian Geo-Statistical model

The hierarchical modeling of Gaussian variations over a continuous domain $\mathcal{D}$ usually starts from the well-known *geo-statistical* model. As for the decomposition in Equation (2.1.8), it expresses the target process $Y(\cdot)$ through the following additive formula:

$$Y(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \eta(\boldsymbol{u}) + \upsilon(\boldsymbol{u}), \quad \boldsymbol{u} \in \mathcal{D}, \tag{2.3.4}$$

where $\mu(\cdot)$ is the *large-scale variation* term and represents the mean-structure/spatial trend of the process, while the residual is partitioned into the sum of a *small scale variation* term $\eta(\cdot)$ and a *measurement error* term $\upsilon(\cdot)$.

As already discussed in Section 2.1.2, the mean term is a continuous deterministic function defined on the same domain of the process, which is also usually assumed to be smooth. In the geo-statistical setting, it is typically expressed as the linear combination of a set of p geo-referenced covariates $\boldsymbol{x}(\cdot) = [x_1(\cdot), \dots, x_p(\cdot)]^\top$ (thus known without error) through a vector of coefficients $\boldsymbol{\beta}$:

$$\mu(\boldsymbol{u}) = \boldsymbol{x}(\boldsymbol{u})^\top \boldsymbol{\beta}, \quad \boldsymbol{u} \in \mathcal{D},\ \boldsymbol{\beta} \in \mathbb{R}^p.$$

Suppose large scale variations due to unobserved spatial features are deemed possible. In that case, the set of covariates can also include the spatial coordinates themselves and/or basis expansions of these (see Section 2.1.1 and 2.2).

The small scale variation term, instead, is stochastic in its nature. It entwines all the spatial dependence structure in the process and is typically assumed to be a stationary *zero-mean* Gaussian process. Other very common assumptions are that of isotropy and that the covariance can be expressed through parametric forms such as the ones introduced in Section 2.1.2. Under these two last assumptions, we can then generally write that:

$$\eta(\cdot) \sim \mathcal{GP}\left(0,\, \sigma^2 \cdot \rho_\theta(\cdot)\right),$$

where $\sigma^2 > 0$ is the *sill* and $\rho_\theta(\cdot)$ is an isotropic correlation function depending on the set of parameters θ.

The *pure error* term $\upsilon(\cdot)$ is a zero-mean Gaussian process again, but with independent components:

$$\upsilon(\boldsymbol{u}) \sim \mathcal{N}(0, \tau^2), \quad \forall\, \boldsymbol{u} \in \mathcal{D},$$

where $\tau^2 > 0$ is known as the *nugget*. It represents random variations of the outcome which are not related to the spatial effect. It induces a discontinuity in the covariance structure of the main process (i.e. the nugget effect). Theoretically, it is not necessary in order to define a valid process and may be excluded from the model specification. However, practically and philosophically speaking, it does have a good deal of importance. Indeed, while:

$$\eta(\boldsymbol{u} + \boldsymbol{h}) - \eta(\boldsymbol{u}) \overset{\boldsymbol{h} \to \boldsymbol{0}}{\to} 0,$$

at any $\boldsymbol{u} \in \mathcal{D}$, this never happens to:

$$(\eta(\boldsymbol{u} + \boldsymbol{h}) + \upsilon(\boldsymbol{u} + \boldsymbol{h})) - (\eta(\boldsymbol{u}) + \upsilon(\boldsymbol{u}))\,.$$

This envisions additional variability associated with the observed process $Y(\cdot)$, that can be explained as the noise associated with eventual replications of the measurement at the same location. Practically, this encourages the *means* of the spatial variables at proximate locations to be close (or equal) to each other, without forcing the process values to be actually close. The absence of the nugget would instead transform a dependence *on average* into a *punctual* dependence, forcing proximate values of the process to be similar. This is not necessarily an issue, but if explanation of the mean is of interest, the inclusion of the *nugget* effect becomes a necessity.

If we consider the evaluation/realization of all the introduced components on any finite set of locations $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l\} \in \mathcal{D}$, these would present the following structures:

$$\begin{aligned} &\boldsymbol{\mu}(\mathcal{U}) = \boldsymbol{X}(\mathcal{U}) \cdot \boldsymbol{\beta} : \ \left(\mathbb{R}^l \times \mathbb{R}^d\right) \to \mathbb{R}^l \\ &\boldsymbol{\eta}\,(\mathcal{U}) \sim \mathcal{N}_l\left(\boldsymbol{0},\, \sigma^2 \cdot \boldsymbol{R}_\theta(\mathcal{U})\right) \qquad\qquad \mathcal{U} \in \mathcal{D}, \\ &\boldsymbol{\upsilon}\,(\mathcal{U}) \sim \mathcal{N}_l\left(\boldsymbol{0},\, \tau^2 \cdot \boldsymbol{I}_l\right) \end{aligned} \tag{2.3.5}$$

where $\boldsymbol{X}(\mathcal{U}) = [\boldsymbol{x}(\boldsymbol{u}_1), \ldots, \boldsymbol{x}(\boldsymbol{u}_l)]^\top$ is the *design matrix* associated with the set of locations $\mathcal{U}$, $\boldsymbol{R}_\theta(\mathcal{U})$ is a $l \times l$ correlation matrix such that $[\boldsymbol{R}_\theta(\cdot)]_{ij} = \rho_\theta(\cdot),\ \forall\, i, j$ and $\boldsymbol{I}_l$ is the $l \times l$ identity matrix.

Let us now denote with $\mathcal{S} = \{s_1, \ldots, s_n\}$ the set of locations on which the vector $\boldsymbol{y}$ of the outcome of interest has been observed. It is the observed realization of the random vector $\boldsymbol{Y}(\mathcal{S}) = [Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)]^\top = \boldsymbol{Y}$, to which also correspond the unknown $\boldsymbol{\eta}(\mathcal{S}) = \boldsymbol{\eta}$ and the design matrix $\boldsymbol{X}(\mathcal{S}) = \boldsymbol{X}$. From now on, for ease of notation, any time the location set is not specified as the argument of spatially varying terms, we mean their values at the observed location set $\mathcal{S}$.

Assigning to the parameter set $\omega = \{\boldsymbol{\beta}, \sigma^2, \theta, \tau^2\}$ a prior distribution $\pi_\omega(\cdot)$, we can pull all the ingredients of 2.3.5 in the additive formula of Equation 2.3.4 and build-up two alternative (but equivalent) hierarchical model formulations: the *conditional model* and the *marginal model*.

Conditional model It is probably the most natural way to define a Bayesian hierarchical model starting from the previous assumptions. The Gaussian pure error term is combined with the (deterministic) large scale variation term, leading to a Gaussian distribution at the data level. As a whole, it results in the following three-level hierarchy:

- **Level 1:** $\boldsymbol{Y} \mid \boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2 \sim \mathcal{N}_n\left(\boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\eta},\, \tau^2 \cdot \boldsymbol{I}_n\right)$
- **Level 2:** $\boldsymbol{\eta} \mid \sigma^2, \theta \sim \mathcal{N}_n\left(\boldsymbol{0},\, \sigma^2 \cdot \boldsymbol{R}_\theta\right)$
- **Level 3:** $(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2) \sim \pi_\omega(\cdot)$.

The data likelihood is defined conditionally on the latent process and hence presents independent components. All the dependence is captured in its mean, stochastic because of its dependence on $\boldsymbol{\eta}$ specified at the second level.

As pointed out at the end of Section 2.3.1, we ultimately seek to perform inference on the parameters and (possibly) on the latent process in terms of their posterior distributions. By Equation (2.3.2) we have:

$$\pi(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2, \boldsymbol{\eta} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\eta}, \tau^2) \cdot f(\boldsymbol{\eta} \mid \sigma^2, \theta) \cdot \pi_\omega(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2), \tag{2.3.6}$$

where typically independent priors are chosen for the parameters:

$$\pi_\omega(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2) = \pi(\boldsymbol{\beta}) \cdot \pi(\sigma^2) \cdot \pi(\theta) \cdot \pi(\tau^2).$$

Generally speaking, the normalizing constant of (2.3.6) does not present a closed-form solution. Hence the posterior is not analytically computable. Therefore, one must resort to MCMC methods to solve and simulate from it (Robert and Casella, 2013). In the considered framework, the MCMC method would update sequentially the latent states $\boldsymbol{\eta}$ given the parameters and the parameters $\boldsymbol{\beta}$, σ^2, θ, τ^2 given the latent states.

The efficiency of such simulation methods depends on many factors, among which a suitable choice of the prior distributions. We already picked the $\mathcal{GP}$ as a prior for the latent process. However, we did not discuss the choice of priors for the parameters. A useful candidate for $\boldsymbol{\beta}$ is the multivariate Normal $\mathcal{N}_p\left(\boldsymbol{0},\, \boldsymbol{V}_\beta\right)$. In particular, the $\mathcal{GP}$ prior on the latent process $\boldsymbol{\eta}(\cdot)$ and the multivariate Normal prior on $\boldsymbol{\beta}$ in the Gaussian setting have the great advantage of leading to closed form solution for the respective full-conditionals:

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \cdot\,) &= \mathcal{N}_p\left(\boldsymbol{\beta} \,\middle|\, \boldsymbol{B}^{-1}\boldsymbol{b},\, \boldsymbol{B}^{-1}\right) \\ f(\boldsymbol{\eta} \mid \cdot\,) &= \mathcal{N}_n\left(\boldsymbol{\eta} \,\middle|\, \boldsymbol{E}^{-1}\boldsymbol{e},\, \boldsymbol{E}^{-1}\right) \end{aligned} \tag{2.3.7}$$

where $\mid\cdot$ is short for *all the rest* and

$$\begin{aligned} \boldsymbol{B} &= \left(\frac{1}{\tau^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{V}_\beta{}^{-1}\right), & \boldsymbol{b} &= \frac{1}{\tau^2}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\eta}) \\ \boldsymbol{E} &= \left(\frac{1}{\tau^2}\boldsymbol{I}_n + \frac{1}{\sigma^2}\boldsymbol{R}_\theta{}^{-1}\right), & \boldsymbol{e} &= \frac{1}{\tau^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \end{aligned} \tag{2.3.8}$$

This enables the envision of a Gibbs-step for their update inside a Metropolis-within-Gibbs algorithm, with sensible improvements in terms of mixing and convergence (Gelfand and Smith, 1990; Casella and George, 1992).

As a general rule, one may even adopt non-informative priors for the mean parameters $\boldsymbol{\beta}$, for which improper priors will always provide proper posteriors. However, this choice is discouraged in the case of the remaining variance-covariance parameters. These are usually very poorly identified (Zhang, 2004; Gelfand et al., 2010; Tang et al., 2019), and such choice may lead to improper posteriors with consequent MCMC convergence failure. Shape-wise, inverse-gamma is the typical choice for σ^2 and τ^2, while the specification of the prior on θ is dependent on the chosen correlation function form. If appropriate reparametrization are not considered, these are usually updated through a block Metropolis-Hastings step (Roberts et al., 1997).

The same Bayesian recipe can be used to get predictions at both sampled and unsampled locations. Let us again denote with $\mathcal{U} = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_l\}$ an arbitrary set of un-sampled locations on which the outcome posterior distribution is of interest. Denoting with $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}(\mathcal{U})$ and $\widetilde{\boldsymbol{\eta}} = \boldsymbol{\eta}(\mathcal{U})$, inference is based on the so-called posterior predictive distribution. According to Equation (2.3.3), this is:

$$\begin{aligned} f_{\widetilde{\boldsymbol{Y}}}\left(\widetilde{\boldsymbol{y}}|\boldsymbol{y}\right) = \int\int & f(\widetilde{\boldsymbol{y}}|\,\widetilde{\boldsymbol{\eta}}, \boldsymbol{\beta}, \tau^2, \boldsymbol{y}) f(\widetilde{\boldsymbol{\eta}}, \boldsymbol{\eta}|\,\sigma^2, \theta, \boldsymbol{y}) \cdot \\ & \cdot \pi(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2|\,\boldsymbol{y})\; d\boldsymbol{\eta}\, d\boldsymbol{\beta}\, d\sigma^2\, d\theta\, d\tau^2, \end{aligned} \tag{2.3.9}$$

where $f(\widetilde{\boldsymbol{\eta}}, \boldsymbol{\eta}|\,\sigma^2, \theta, \boldsymbol{y}) = f(\widetilde{\boldsymbol{\eta}}|\,\boldsymbol{\eta}, \sigma^2, \theta, \boldsymbol{y}) \cdot f(\boldsymbol{\eta}|\,\sigma^2, \theta, \boldsymbol{y})$. As for the posterior distribution of the parameters, neither the posterior predictive distribution (2.3.9) usually presents a closed analytical expression. Nevertheless, we can use the samples from the posterior distribution of the parameters and the latent process provided by the MCMC algorithm to solve/simulate it. Indeed, we recall that the conditional distributions of multivariate Normal vectors are again multivariate Normal vectors. This implies that given $\boldsymbol{\eta}$, σ^2, and θ distributed according their posterior distribution, we can simulate from:

$$\widetilde{\boldsymbol{\eta}}|\,\boldsymbol{\eta}, \sigma^2, \theta \sim \mathcal{N}_l\left(\boldsymbol{R}_\theta(\mathcal{S},\mathcal{U})\widetilde{\boldsymbol{R}}_\theta^{-1}\boldsymbol{\eta},\; \sigma^2 \cdot \left(\boldsymbol{R}_\theta - \boldsymbol{R}_\theta(\mathcal{S},\mathcal{U})\widetilde{\boldsymbol{R}}_\theta^{-1}\boldsymbol{R}_\theta(\mathcal{U},\mathcal{S})\right)\right) \tag{2.3.10}$$

where $\boldsymbol{R}_\theta(\cdot,\cdot)$ is the cross-correlation matrix of its two arguments and $\widetilde{\boldsymbol{R}}_\theta = \boldsymbol{R}_\theta(\mathcal{U})$. Hence, by *composition* sampling, we can easily sample at each MCMC iteration first from $\widetilde{\boldsymbol{\eta}}|\,\boldsymbol{\eta}, \sigma^2, \theta, \boldsymbol{y}$ and after from $\widetilde{\boldsymbol{Y}}|\,\widetilde{\boldsymbol{\eta}}, \boldsymbol{\beta}, \tau^2, \boldsymbol{y}$. This can be done during the MCMC main run, or recycling the already simulated parameters' chains in a second moment.

Marginal model Theoretically speaking, one can always combine two subsequent levels of a hierarchical model by marginalizing the stochastic components of the lower into the upper level. The marginal model is indeed built by integrating the latent spatial process (level 2) out of the data likelihood (level 1):

$$f(\boldsymbol{Y}|\,\boldsymbol{\beta}, \sigma^2, \theta, \tau^2) = \int_{\mathbb{R}^n} f\left(\boldsymbol{Y}\;\middle|\;\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2\right) \cdot f\left(\boldsymbol{\eta}\;\middle|\;\sigma^2, \theta\right) d\boldsymbol{\eta}. \tag{2.3.11}$$

The Gaussianity assumption on the latent process $\eta(\cdot)$ and on the errors $\epsilon(\cdot)$, reflected onto the data, allows for a straightforward solution of (2.3.11). Indeed, by basic conjugacy properties of the Normal-Normal model, the solution is again a multivariate Normal distribution with updated parameters. Let us denote the set of parameters that govern the whole variance and covariance structure in the process with $\phi = \{\sigma^2, \theta, \tau^2\}$. The marginal model can then be expressed as a hierarchy of the following two levels:

- **Level 1:** $\boldsymbol{Y}|\,\boldsymbol{\beta}, \sigma^2, \theta, \tau^2 \sim \mathcal{N}_n\left(\boldsymbol{X}\cdot\boldsymbol{\beta},\, \boldsymbol{\Sigma}_\phi\right), \qquad \boldsymbol{\Sigma}_\phi = \sigma^2\cdot\boldsymbol{R}_\theta + \tau^2\cdot\boldsymbol{I}_n.$
- **Level 2:** $(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2) \sim \pi_\omega(\cdot).$

Here, the data likelihood is defined marginalizing on the latent process, and therefore the dependence structure is no longer captured by the mean but by the covariance structure. Considerations on this model are very similar to the ones of the conditional model, with the main difference being that dependence on the latent process is not apparent in this case. Indeed, the expression of the parameters posterior distribution now is:

$$\pi(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2|\,\boldsymbol{y}) \propto f(\boldsymbol{y}|\,\boldsymbol{\beta}, \sigma^2, \theta, \tau^2)\cdot\pi_\omega(\boldsymbol{\beta}, \sigma^2, \theta, \tau^2). \tag{2.3.12}$$

Posterior samples of the parameters are not drawn conditionally on the latent process, but marginally. If a multivariate Normal distribution is assigned as prior to the coefficients $\boldsymbol{\beta}$, then partial conjugacy (closed-form full-conditional) still holds. Its marginal expression is again a multivariate Normal, but the terms defining its mean and covariance are slightly changed with respect to the ones in (2.3.8), i.e.:

$$\boldsymbol{B} = \left(\boldsymbol{X}^\top\boldsymbol{\Sigma}_\phi^{-1}\boldsymbol{X} + \boldsymbol{V}_\beta{}^{-1}\right), \qquad \boldsymbol{b} = \frac{1}{\tau^2}\boldsymbol{X}^\top\boldsymbol{\Sigma}_\phi^{-1}\boldsymbol{y}.$$

Simultaneously, the explicit absence of the latent process $\boldsymbol{\eta}$ implies that we can perform inference on the parameters without sampling values from its distribution, with potential computational advantages. However, if we are interested in its estimation, the marginal model does not preclude the possibility of sampling from it. Indeed, recalling that $f(\boldsymbol{\eta}|\boldsymbol{y}) = \int f(\boldsymbol{\eta}|\boldsymbol{\beta}, \phi)\cdot\pi(\boldsymbol{\beta}, \phi|\boldsymbol{y})\; d\boldsymbol{\beta}\, d\phi$, we can exploit composition sampling and use the posterior sample of the parameters at each iteration of the MCMC to sample from $\boldsymbol{\eta}|\boldsymbol{y}$. The same can be done to obtain samples (and thus predictions) at un-sampled locations using (2.3.10).

While the two models are theoretically equivalent to each other, the marginal model is usually preferable from a practical perspective. This rationale stems from a general consideration about MCMC techniques, whose efficiency increases the more one can marginalize a model analytically. Marginalizing reduces the dimensionality of the space that the sampler must explore, favoring good mixing and reducing correlations. Moreover, as discussed in Banerjee et al. (2014), the covariance structure of $\boldsymbol{\Sigma}_\phi$ is way more numerically stable than the one of $\sigma^2\cdot\boldsymbol{R}_\theta$ because of the additive constant on the diagonal. Consequently, the determinant and inverse computations on which the likelihood and full conditionals depend on are better behaved in the marginal model than in the conditional model.

This last paragraph concludes the brief introduction to the Bayesian hierarchical modeling of spatial processes. In these last pages, we explained the necessity to rely on MCMC techniques to perform inference on the parameters, the latent process, and the observable process distribution at unsampled locations. Nevertheless, we never discussed the topic of the computational efficiency of such methods. The latter is a pressing issue in the Big-Data era we live in, and it is one of the principal motives of this research's attention. In Section 2.5, just after some brief considerations on the framework extension to the analysis of spatio-temporal processes in Section 2.4, we will delve more deeply into this problem.

2.4 Continuous Spatio-Temporal Processes

Until now, we have considered spatial stochastic processes $Y(\cdot) : \mathcal{D} \to \mathbb{R}$, where the domain $\mathcal{D} \subseteq \boldsymbol{R}^d$ is assumed to be euclidean. We now briefly discuss the problem

of analyzing stochastic processes defined over a spatio-temporal domain, i.e.:

$$\left\{Y(\boldsymbol{u},\tau) \,:\, (\boldsymbol{u},\tau) \in \mathcal{D}\times\mathcal{G} \subseteq \mathbb{R}^d \times \mathbb{R}\right\},$$

that vary both as a function of the spatial location $\boldsymbol{u} \in \mathcal{D} \subseteq \mathbb{R}^d$, and time $\tau \in \mathcal{G} \subseteq \mathbb{R}$. Spatio-temporal problems arise in the analysis of dynamic processes that evolve both in space and time, and are actually pretty common in many scientific and engineering fields: environmental sciences (Oehlert, 1993; Vyas and Christakos, 1997; Lindström et al., 2014), climate predictions and meteorology (Armstrong et al., 1993; Faghmous and Kumar, 2014), disease mapping (Christakos and Hristopulos, 1998; Yanosky et al., 2008; Meliker and Sloan, 2011), etc.

We have widely seen how statistical analysis of natural and artificial phenomena encompasses the adoption of a stochastic model for the observables. Differently from physical models, statistical ones are not (necessarily) based on the causal mechanistic interactions occurring at the microscopic level and generating the data. They provide inference and predictions of unobserved outcome of the system by learning the joint spatial and temporal dependence of the underlying process from the observed data, under suitable probabilistic hypotheses on their structure. From a purely mathematical perspective, spatio-temporal processes are not much different from a $(d+1)$-dimensional spatial process, whose domain is within $\mathbb{R}^{d+1} = \mathbb{R}^d \times \mathbb{R}$. After all, time is nothing else but an additional coordinate and all the techniques and all the results discussed above clearly extend and apply to space-time problems defined on the augmented space. However, from a physical and practical perspective, this approach is generally misleading and intrinsically flawed. Stochastic models must approximate (accurately) the outcomes of true physical dynamics, and therefore cannot neglect completely the original nature of the phenomena under study. The temporal dimension is structurally different from spatial ones. First of all, time moves only forward, while there is no preferred direction in space (usually). Secondly, spatial lags are difficult, if not impossible, to compare with temporal lags since they come in distinct units. Finally, the process may behave in completely different ways at the same lags on the spatial and temporal dimension, e.g.: seasonality/periodicity makes sense along time, while it is generally absent along space; isotropy is well defined in space, while it has no meaning in a space–time context due to the intrinsic ordering and non-reversibility of time. The specification of the combined space–time domain $\mathcal{D}\times\mathcal{G}$ acknowledges these differences, and should be regarded only as a very specific coordinate system, where observations are tagged by a spatial coordinate vector $\boldsymbol{u}$ and a separate temporal coordinate τ.

In many contexts, spatio-temporal processes are not observed continuously along time, but can be discretized to customary integer-spaced intervals $\mathcal{G} = \{\tau_1, \tau_2, \ldots\}$. In this cases, the problem reduces to the analysis of a collection of temporally varying spatial processes $Y_t(\cdot),\ t \in \mathcal{G}$: different picture of the same area at different occasions. If time is assumed to not play a direct role in the explanation of the observed phenomenon, each spatial process may be assumed to evolve independently, perturbed with different random errors[9]. Therefore, all the independent realizations can be used to inform about the intrinsic structure of the same underlying spatial process. Otherwise, also the potential temporal dependence between the points belonging to different surfaces at subsequent times must be taken into account. The analysis of such data encompasses the use of standard time series models with spatial covariance structure (Pfeifer and Deutsch, 1980; Pfeifer and Jay Deutsch,

[9]Potentially, taking a switch of perspective, one may want to talk about collections of spatially varying time-series.

1980; Stoffer, 1986), or of the *dynamic spatiotemporal* models. These describe the temporal evolution of spatial points by specifying suitable transition equations in the space of latent features that induce temporal dependence between residuals at different times (Huang and Cressie, 1996; Tonellato, 1997; Sanso and Guenni, 1999; Stroud et al., 2001; Gelfand et al., 2003; West and Harrison, 2006).

Instead, we here discuss extensions of the concepts and methodologies introduced in Section 2.1 to the spatio-temporal modeling of (univariate) continuous data, with continuous time. Inference is thus sought at arbitrary scales, that usually are even finer than the observed data. The corresponding spatio-temporal process is nothing else but a random variable $Y(\cdot,\cdot) : \mathcal{D} \times \mathcal{G} \to \mathcal{Y}$ having realizations at any spatial location $\boldsymbol{u} \in \mathcal{D} \subseteq \mathbb{R}^d$ and instant in time $\tau \in \mathcal{G} \subseteq \mathbb{R}$, according to a probability distribution. As in Equation 2.1.2, the distribution is uniquely determined by the collection of finite dimensional joint distribution, i.e.:

$$F_{Y(\cdot,\cdot)}\left(y_1,\ldots,y_l;(\boldsymbol{u}_1,\tau_1),\ldots,(\boldsymbol{u}_l,\tau_l)\right), \quad \forall\ \{(\boldsymbol{u}_i,\tau_i)\}_{i=1}^{l} \subset \mathcal{D}\times\mathcal{G},$$

where the realizations at different space-time locations are usually dependent random variables. Also the concepts of strict and second-order stationarity directly translate in the spatio-temporal context as:

$$\begin{aligned} F_{Y(\cdot,\cdot)}\left(y_1,\ldots,y_l;\ (\boldsymbol{u}_1,\tau_1),\ldots,(\boldsymbol{u}_l,\tau_l)\right) = \\ = F_{Y(\cdot,\cdot)}\left(y_1,\ldots,y_l;(\boldsymbol{u}_1+\boldsymbol{h},\tau_1+\delta),\ldots,(\boldsymbol{u}_l+\boldsymbol{h},\tau_l+\delta)\right) \end{aligned} \tag{2.4.1}$$

and

$$\begin{aligned} &\mathbb{E}\left[Y(\boldsymbol{u},\tau)\right] = \mathbb{E}\left[Y(\boldsymbol{u}+\boldsymbol{h},\tau+\delta)\right] \\ &\mathbb{C}\mathrm{ov}\left[Y(\boldsymbol{u},\tau),Y(\boldsymbol{u}+\boldsymbol{h},\tau+\delta)\right] = \mathbb{C}\mathrm{ov}\left[Y(\boldsymbol{0},0),Y(\boldsymbol{h},\delta)\right] \end{aligned} \tag{2.4.2}$$

for any $\{(\boldsymbol{u}_i,\tau_i)\}_{i=1}^{l} \subset \mathcal{D}\times\mathcal{G}$ and $(\boldsymbol{h},\delta) \in \mathbb{R}^d \times \mathbb{R}$. We recall that in the case of Gaussian processes, second order stationarity implies strict stationarity.

Whilst convenient, stationarity is hardly matched by natural processes observed over large regions or span of times. Usually, this non-stationarity is limited to the presence of a mean/trend component that may be dependent on the location or the time instant (for instance because of space-time varying unobserved heterogeneity). The stationarity assumption is then not entirely abandoned, but used to properly model any residual variation from such trend. Without any loss of generality, we can easily extend the representation of Equation (2.1.8) to the spatio-temporal case as:

$$Y(\boldsymbol{u},\tau) = \mu(\boldsymbol{u},\tau) + \upsilon(\boldsymbol{u},\tau), \quad \boldsymbol{u} \in \mathbb{R}^d,\ \tau \in \mathbb{R} \tag{2.4.3}$$

were $\mu(\cdot) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ denotes the deterministic space-time mean structure and $\upsilon(\cdot) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ the residual term. Analysis and inference on the underlying process from a set of realizations $\boldsymbol{y} = [y_i]_{i=1}^{n}$ at $\mathcal{O} = \{(\boldsymbol{s}_i,t_i)\}_{i=1}^{n} \subset \mathcal{D}\times\mathcal{G}$ goes through suitable modeling of either or both the deterministic mean function and residual terms covariance structure. As for the spatial case, we will refer to the former as *spatial interpolation* or *spatial smoothing*; we briefly introduce some spatio-temporal methods in Section 2.4.1, with some additional care devoted to adaptation of *B-spline regression*. We refer to the latter as *geo-statistical modeling*, while being aware that this last category evokes a larger set of techniques; Section 2.4.2 explores the difficulties and peculiarities of defining a valid and meaningful spatio-temporal covariance function.

2.4.1 Interpolation, smoothing and spline regression

If the spatio-temporal changes could be explained in terms of large scale variations only, then the underlying assumption is that the residuals $\upsilon(\boldsymbol{u}, \tau)$ of Equation (2.4.3) shall behave as a stationary and independent process once the deterministic expression of $\mu(\boldsymbol{u}, \tau)$ is known. If space-time heterogeneity was completely observed and known, then this effect could be completely explained by the available space-time covariates, as in the case of *land-use regression* (Ryan and LeMasters, 2007; Hoek et al., 2008). On the other hand, this is practically never the case in real applications. Alternatively, we could look for a mathematical expression of the space-time trend function $\mu(\cdot, \cdot)$ that well interpolates or smooths the observed data, leaving uncorrelated residuals (or zero residuals under the assumption that data are observed without error). Theoretically speaking, when dealing with the expression of the trend function, techniques do not vary much with respect to what we already introduced for the spatial context in Section 2.1.1.

In the GIS literature, it is usually made a distinction between two classes of approaches: the *reduction* and the *extension* approach (Li and Revesz, 2002, 2004). The former defines a hierarchy between the spatial and the temporal components. It encompasses the subsequent use of two interpolating techniques, where the second is used conditionally on the results of the first. The whole space-time is either sliced into as many d-dimensional spatial sub-spaces as many different time-points are observed, or viceversa (i.e. as many temporal 1-dimensional spaces as many different space-points). A first interpolating techniques is used the smooth *independently* all the slices and, conditionally on that, the slices are finally joined altogether by using a second interpolation technique on the initially excluded temporal dimension. The interpolators may either be *Kernel-Smoothers* (such as the *Inverse Distance Weighting*), or FEM based shape functions, but also local polynomials, splines etc. (Mauser and Prasch, 2015; Xiao et al., 2016). From a practical point of view, this corresponds to a sequence of spatial interpolation problems occurred in different time-series (or viceversa). For instance, we first exclude the temporal dimension and build a collection of independently interpolated surfaces. Then, the resulting surfaces are connected point by point by interpolating those belonging to different surfaces along their common temporal dimension. This approach is clearly only viable if samples are always taken in the same locations, all at the same times. When this is not the case, each slice may contain few (or none) points and interpolation would become unfeasible.

In the *extension* setting time is instead treated as an additional dimension. Therefore, spatio-temporal interpolation is simply handled as a one higher dimension spatial interpolation, directly *extending* any of the methodologies introduced in Section 2.1.1. This does not require the observed data to be recorded over regular space-time grids but, on the downside, can result computationally expensive to be computed. Furthermore, it requires additional care in defining an appropriate time-scale that would (at least approximately) adapt to the different behaviors of the spatial and temporal dimensions. Being these methods non-parametric, hence not model-based, this is usually attained via brute force by cross validating results obtained on different time scales (Li et al., 2020).

These non-parametric interpolation methods have been mainly produced in the engineering and GIS literature. For a review we refer to (Eldrandaly and Abdelmouty, 2017) and references therein, that include different works in applied science that have seen the application and comparison of many of these. Generally speaking, they all present much focus on the practical aspects of obtaining a point-wise accurate interpolated surface, able to provide estimates at unsampled locations, but only

little care to the statistical and inferential properties of the estimation process. As a matter of fact, such surfaces are *computed* from the given data, and not really *estimated*.

As for the spatial case, we would rather switch perspective and transform the interpolation/smoothing task into an estimation problem. In the simplest case, we can assume the so-called *space-time* separability to hold at the mean function level (Loader and Switzer, 1992). This means that the trend function decomposes into the sum, or product, of a purely spatial and purely temporal component:

$$\begin{aligned} \mu(\boldsymbol{u}, \tau) &= \mu_s(\boldsymbol{u}) + \mu_t(\tau) \\ \mu(\boldsymbol{u}, \tau) &= \mu_s(\boldsymbol{u}) \cdot \mu_t(\tau) \end{aligned} \qquad \boldsymbol{u} \in \mathbb{R}^d,\ \tau \in \mathbb{R},$$

where the additive composition of the two effects implies total neutrality between the two components, while the multiplicative one (additive on the log-scale) denotes an augmentative/diminutive mutual effect. This allows modeling space and time effects independently, respecting their own peculiarities. The spatial component can be modeled using any of the methodologies introduced in Section 2.1.1, e.g. *Spline Smoothing* from Section 2.2. The same holds for the temporal component, that on the other hand may also include the contribution of trigonometric functions in order to reflect diurnal or seasonal effects. The estimation is not much different from what is usually done in the larger context of *Generalized Additive Models* (MacNab and Dean, 2001; Fahrmeir et al., 2004; Kneib and Fahrmeir, 2006; Hastie et al., 2017)

This kind of decomposition is very convenient from a model specification and estimation point of view. However, it assumes that the temporal effect is constant over the space, while the spatial effect is constant trough time. It neglects any structural interactions between the spatial and temporal components, often characterizing observed data that manifest different spatial behaviors at different times and/or viceversa. In such cases, space-time separability does not hold and the deterministic trend component must be modeled using non-separable structures. This can be attained by modeling the trend as:

$$\mu(\boldsymbol{u}, \tau) \leftarrow \widetilde{\mu}(\boldsymbol{u}, \tau) = \sum_{j=1}^{J} \sum_{l=1}^{L} \beta_{jl} \phi_{jl}(\boldsymbol{u}, \tau) = \boldsymbol{\Phi}(\boldsymbol{u}, \tau)^\top \boldsymbol{\beta}, \quad \forall (\boldsymbol{u}, \tau) \in \mathcal{D} \times \mathcal{G},$$

where $\{\phi_{jl}(\cdot, \cdot)\}_{j,l=1}^{J,L}$ are elements of a space-time basis function chosen to fit the average variation of the data, then collected over the vector $\boldsymbol{\Phi}(\cdot, \cdot)$, and combined through coefficients $\{\beta_{jl}\}_{j,l=1}^{J,L}$ that must be estimated by a fitting procedure (Kyriakidis and Journel, 1999; De Luna and Genton, 2005; Wang and Wang, 2009). This can include periodic terms along the time axis and polynomial terms to model smooth variations in space. For instance, Dimitrakopoulos and Luo (1997) proposes three different forms: traditional polynomial functions along all dimensions, Fouriers expressions (i.e. periodic functions) along all dimensions, a combination of both. Obviously, each dimension can be modeled using different shapes and/or orders, allowing flexibility in its specification. Without diving into any more details, this encompasses the expression of the space-time trend through basis elements of globally defined functions. Section 2.1.1 discussed how spatial data, hence spatio-temporal data too, present local behaviors that can harshly vary among different regions of the domain. Seeking a good fit by mean of globally defined functions is a path full of pitfalls, that usually leads to over-parametrized models which are rarely able to provide homogeneous fitting performance on the whole considered domain. This often produces residuals that violate the stationarity hypothesis.

Therefore, considering functions that benefit from local flexibility, but still guarantee desirable global properties, is a more convenient strategy: *local-polynomials* (Haas, 1995; Wang and Wang, 2009), thin-plate splines (Aberg et al., 2005; Hancock and Hutchinson, 2006), spline basis and penalized splines (Brezger and Lang, 2006; Lee and Durbán, 2011; Rodriguez-Alvarez et al., 2018; Spiegel et al., 2020) etc.

B-Spline Space-Time regression We here focus on the specification of a suitable space-time B-spline regression approach, hence express the space-time trend function as:

$$\mu(\boldsymbol{u}, \tau) \leftarrow \widetilde{\mu}(\boldsymbol{u}, \tau) = \boldsymbol{B}^{s,t}(\boldsymbol{u}, \tau)^\top \boldsymbol{\beta},$$

where $\boldsymbol{B}^{s,t}(\cdot, \cdot)$ is the vector of the elements of a spatio-temporal B-spline basis and $\boldsymbol{\beta}$ is a vector of coefficients. Let us denote with $\boldsymbol{y} = [y_i]_{i=1}^n$ the column vector of observed values at the spatial locations $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^n \subset \mathcal{D}$ and time-points $\mathcal{T} = \{t_i\}_{i=1}^n \subset \mathcal{G}$. We already introduced the straightforward extension of B-spline basis to the multidimensional case in Section 2.2.4, focusing on the 2-dimensional case. This encompasses building two independently defined B-spline bases $\boldsymbol{B}^{(o_1,K_1)}(\cdot)$, $\boldsymbol{B}^{(o_2,K_2)}(\cdot)$ with J_1 and J_2 elements (depending on the order and number of knots of each). Their Kroenecker (tensor) product $\boldsymbol{B}^{(2)}(\cdot) = \boldsymbol{B}^{(o_1,K_1)}(\cdot) \otimes \boldsymbol{B}^{(o_2,K_2)}(\cdot)$ yields the combined B-spline basis. The same procedure can be used to extend the 2-dimensional basis to further dimensions, and hence time. Without any loss of generality, the spatial basis $\boldsymbol{B}^{(2)}(\cdot)$ can be extended to the additional temporal dimension by taking the following tensor product:

$$\boldsymbol{B}^{(3)}(\boldsymbol{u}, \tau) = \boldsymbol{B}^{(2)}(\boldsymbol{u}) \otimes \boldsymbol{B}^{(o_3,K_3)}(\tau),$$

where $\boldsymbol{B}^{(3)}(\cdot, \cdot)$ is the resulting spatio-temporal basis and $\boldsymbol{B}^{(o_3,K_3)}(\cdot)$ is a B-spline basis of order o_3 and with knots $\{k_{3j} : j = 0, \dots, K_3\}$ placed over $\mathcal{T}$, with J_t elements. The independent choice of degree and knots over the three dimensions gives much space in the definition of an anisotropic basis. This is particularly useful over spatio-temporal domains. Indeed, as mentioned above, while it may be reasonable to assume isotropy along space and consider analogous basis over the two spatial dimensions (i.e. $o_1 = o_2$, $K_1 = K_2$), a different structure may be necessary in order to model the behavior along the temporal dimension. Evaluating the basis at the observed locations, the vector of coefficients $\boldsymbol{\beta}$ can then be estimated by setting up the usual optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ l\left(\boldsymbol{y} \,|\, \boldsymbol{B}^{(3)}(\mathcal{S}, \mathcal{T}),\, \boldsymbol{\beta}\right) \right\}, \tag{2.4.4}$$

where:

$$\boldsymbol{B}^{(3)}(\mathcal{S}, \mathcal{T}) = \left[\boldsymbol{B}^{(3)}(\boldsymbol{s}_1, t_1), \dots, \boldsymbol{B}^{(3)}(\boldsymbol{s}_n, t_n) \right]^\top.$$

Efficient implementations in this setting are discussed in Currie et al. (2006); Spiegel et al. (2020). The choice of the orders o_1, o_2 and o_3 and the number and placement of the knots $\{k_{ij} : i = 1, 2, 3,\ j = 0, \dots, K_i\}$ play a key role in the structure of the resulting basis. An order of 2 or 3 is usually sufficient for achieving sufficient flexibility, while placing too many knots or too few may result in under or over-fitting. A general guidance is to consider evenly spaced knots along all dimensions, with finer resolutions at areas (or dimensions) over the one the mean function is expected to vary more wildly. Usually, different candidates from a shortlist are compared through cross-validation approach, but in the spatio-temporal setting this may entail an incredibly long computational time.

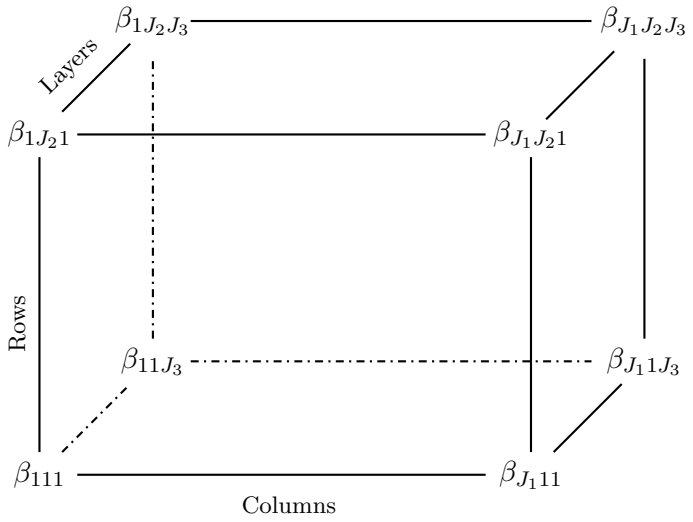


Figure 2.7. Neighboring structure of the 3 dimensional B-Spline coefficients.

A viable and automatic solution is to not waste time in the knots placement decision, but pick the upper bound given the available computational resources while resorting to penalized approaches. This corresponds to replacing the criteria of Equation (2.4.4) with:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ l\left(\boldsymbol{y} \,\middle|\, \boldsymbol{B}^{(3)}(\mathcal{S}, \mathcal{T}),\, \boldsymbol{\beta}\right) - \text{pen}_{\lambda}\left(\boldsymbol{\beta}\right) \right\},$$

where $\text{pen}_{\lambda}(\cdot)$ puts a penalty on the vector of parameters, modulated by one (or more) shrinkage parameter λ. We have already introduced this approach in Section 2.2.3, adopting a Bayesian perspective in which the penalty is represented by a suitable choice of covariance structure in the prior:

$$\pi(\boldsymbol{\beta}) = \exp\left\{-\boldsymbol{\beta}^{\top}\boldsymbol{M}_{\lambda}^{(m)}\boldsymbol{\beta}\right\} \propto \mathcal{N}_J\left(\boldsymbol{\beta}\,\middle|\,\boldsymbol{0},\, \left(\boldsymbol{M}_{\lambda}^{(m)}\right)^{-1}\right).$$

If knots are placed on regular lattices along space and time, the concept of *neighborhood* is easily extendable to the spatio-temporal context. Indeed, every knot will also be a neighbor of itself at the previous and following times. This can include also higher order neighborhood structures along time and space, or crossing the two dimensions (i.e. being a neighbor of the spatial neighbors at previous and following times). 2.2.3.

When cross-neighborhoods are not included, the overall penalization matrix can be easily computed using the same passages of Section Let m_s and m_t be the chosen order differences over space and time, and let us denote the corresponding univariate difference matrices along the three dimensions as $\boldsymbol{D}_1^{(m_s)}$, $\boldsymbol{D}_2^{(m_s)}$ and $\boldsymbol{D}_3^{(m_t)}$. We can then compute the sum of row-wise (spatial dimension 1), column-wise (spatial dimension 2) and layer-wise (temporal dimension) differences in the lattice of coefficients, placed as in Figure 2.7. The matricial expression goes through the definition of the blown-up difference matrices, summed all together in the following penalty matrix:

$$\begin{aligned}\boldsymbol{M}_{\lambda}^{(m_s,m_t)} =& \lambda_s\left(\boldsymbol{I}_{J_3}\otimes\boldsymbol{I}_{J_2}\otimes\boldsymbol{M}_1^{(m_s)} + \boldsymbol{I}_{J_3}\otimes\boldsymbol{M}_2^{(m_s)}\otimes\boldsymbol{I}_{J_1}\right) + \\ &+ \lambda_t\left(\boldsymbol{I}_{J_3}\otimes\boldsymbol{I}_{J_2}\otimes\boldsymbol{M}_3^{(m_t)}\right),\end{aligned}$$

where $\boldsymbol{M}_1^{(m_s)} = \boldsymbol{D}_1^{(m_s)^\top}\boldsymbol{D}_1^{(m_s)}$, $\boldsymbol{M}_2^{(m_s)} = \boldsymbol{D}_2^{(m_s)^\top}\boldsymbol{D}_2^{(m_s)}$ and $\boldsymbol{M}_3^{(m_t)} = \boldsymbol{D}_3^{(m_t)^\top}\boldsymbol{D}_3^{(m_t)}$ are the univariate penalty matrices. In the penalized scenario, the choice of knots and order cannot anymore be addressed toward inducing anisotropy, but the same can be achieved by considering different shrinkage parameters to the penalty along space λ_s and along time λ_t. The upside is that the penalization parameter can be estimated together with any other parameter (especially in a Bayesian context), and therefore the degree of anisotropy between space and time can be automatically detected by the model. Obviously, as for the purely spatial case, nothing precludes from considering different penalties on the two different spatial directions.

The *smooth-ANOVA* representation in Gu (2002) would also allow further decomposition of the mean trend function as:

$$\widetilde{\mu}(\boldsymbol{u},\tau) = \boldsymbol{B}^s(\boldsymbol{u})^\top\boldsymbol{\beta}_s + \boldsymbol{B}^t(\tau)^\top\boldsymbol{\beta}_t + \boldsymbol{B}^{s,t}(\boldsymbol{u},\tau)^\top\boldsymbol{\beta}_{s,t},$$

where the first is a purely spatial term, the second a purely temporal term and the last an interaction term (Lee and Durbán, 2011). Particular care must be adopted in this setting, since there are overlapping components in the pure and interaction terms. Nevertheless, the mixed model representation of Gu (2002) allows for the identification such components and remove them, or apply constraints to guarantee identifiability. The reader is referred to the original paper for further details.

2.4.2 The geo-statistical model and Hierarchical Modeling

The fit of a deterministic space-time trend function is often unable to describe the spatio-temporal variations present in the data at all scales. As a consequence, the difference between observations and the estimated trend, computed as:

$$\hat{v}_i = y_i - \mu(\boldsymbol{s}_i, t_i), \quad \forall\, i = 1,\dots,n,$$

does not results in correlated residuals. In the *geo-statistical* setting we do not exclude the existence of a deterministic trend function, but unload part of the fitting burden to the second-order structure of the process residuals. Following Equation (2.1.13) from Section 2.3, we decompose the residual term in two components and express the spatio-temporal process as the following sum:

$$Y(\boldsymbol{u},\tau) = \mu(\boldsymbol{u},\tau) + \eta(\boldsymbol{u},\tau) + \nu(\boldsymbol{u},\tau), \quad \forall\, (\boldsymbol{u},\tau) \in \mathcal{D}\times\mathcal{G},$$

where $\eta(\cdot,\cdot)$ is known as the *small-scale* variation term and represents the dependence structure of the process, while $\nu(\cdot,\cdot)$ is an independent pure error term (zero-mean, independent components).

The small scale variation $\eta(\cdot,\cdot)$ is usually assumed to be a zero-mean second-order stationary spatio-temporal process, with continuous paths on $\mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$. The covariance between observations at different space-time locations can be expressed in terms of a space-time covariance function $c_\eta(\cdot,\cdot)$. In order to guarantee continuous paths and second-order stationarity, this function must belong to $\mathcal{C}^0$ and its value must depend only on the separation vector between locations:

$$\mathbb{C}\mathrm{ov}\left[\eta(\boldsymbol{u},\tau), \eta(\boldsymbol{u}+\boldsymbol{h},\tau+\delta)\right] = c_\eta(\boldsymbol{h},\delta), \quad \forall\, \boldsymbol{u},\boldsymbol{h} \in \mathbb{R}^d,\ \tau,\delta \in \mathbb{R}.$$

Its margins $c_\eta(\boldsymbol{0},\cdot)$ and $c_\eta(\cdot,0)$ are purely spatial and purely temporal covariance functions, respectively. A very common additional assumption is that $\eta(\cdot,\cdot)$ is a spatio-temporal Gaussian process $\mathcal{GP}\,(0,\, c_\eta(\cdot,\cdot))$, for which second-order stationarity implies strict stationarity (see Equation (2.4.1)). This does not necessarily preclude

the modeling of non-Gaussian outcomes, that can be related to a latent Gaussian process that drives dependence using suitable link functions (McCulloch and Searle, 2001; Bradley et al., 2018, 2020).

The pure error term $\nu(\cdot, \cdot)$ represents uncorrelated measurement errors that increase the variability of the collected outcome. It induces a discontinuity in the paths of the overall residuals $\upsilon(\boldsymbol{u}, \tau) = \eta(\boldsymbol{u}, \tau) + \nu(\boldsymbol{u}, \tau)$ through the so-called *nugget* effect. In the space-time setting, this can include a purely spatial, purely temporal, and a spatio-temporal term:

$$\mathbb{C}\text{ov}\left[\nu(\boldsymbol{u}, \tau), \nu(\boldsymbol{u} + \boldsymbol{h}, \tau + \delta)\right] = \tau_{st}^2 \mathcal{I}_{(\mathbf{0},0)}(\boldsymbol{h}, \delta) + \tau_s^2 \mathcal{I}_{(\mathbf{0})}(\boldsymbol{h}) + \tau_t^2 \mathcal{I}_0(\delta),$$

with τ_{st}^2, τ_s^2 and τ_t^2 positive constants. Following Equation (2.1.14), this implies the following covariance structure on the process of interest:

$$\mathbb{C}\text{ov}\left(Y(\boldsymbol{u}, \tau), Y(\boldsymbol{u} + \boldsymbol{h}, \tau + \delta)\right) = \begin{cases} \tau_{st}^2 + \tau_s^2 + \tau_t^2 + c_\eta(\mathbf{0}, 0) & \boldsymbol{h} = \mathbf{0},\ \delta = 0 \\ \tau_s^2 + c_\eta(\mathbf{0}, \delta) & \boldsymbol{h} = \mathbf{0},\ \delta \neq 0 \\ \tau_t^2 + c_\eta(\boldsymbol{h}, 0) & \boldsymbol{h} \neq \mathbf{0},\ \delta = 0 \\ c_\eta(\boldsymbol{h}, \delta) & \boldsymbol{h} \neq \mathbf{0},\ \delta \neq 0. \end{cases} \tag{2.4.5}$$

Inference for such a spatio-temporal process in the Bayesian Hierarchical modeling setting is not much different from what we have already seen in Section 2.3.1. The *conditional* and *marginal* models specification also hold in the space-time setting, as long as we can find a valid parametric form for specifying the covariance function of the small scale term $c_{\eta,\boldsymbol{\theta}}(\cdot, \cdot)$. Just as in the spatial case, the chosen parametric form must satisfy positive definiteness, i.e. any corresponding covariance matrix of the form:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathcal{U}) = c_{\eta,\boldsymbol{\theta}}(\mathcal{U}) = \begin{bmatrix} c_{\eta,\boldsymbol{\theta}}(\mathbf{0}, 0) & c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h}_{12}, \delta_{12}) & \cdots & c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h}_{1m}, \delta_{1m}) \\ c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h}_{21}, \delta_{21}) & c_{\eta,\boldsymbol{\theta}}(\mathbf{0}, 0) & \cdots & c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h}_{2m}, \delta_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h}_{m1}, \delta_{m1}) & c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h}_{m2}, \delta_{m2}) & \cdots & c_{\eta,\boldsymbol{\theta}}(\mathbf{0}, 0) \end{bmatrix},$$

must be positive definite for any set of space-time locations $\mathcal{U} = \{(\boldsymbol{u}_i, \tau_i)\}_{i=1}^m \in \mathbb{R}^d \times \mathbb{R}$, where $\boldsymbol{h}_{ij} = \boldsymbol{u}_i - \boldsymbol{u}_j$ and $\delta_{ij} = \tau_i - \tau_j$ for any $i, j = 1, \ldots, m$. This allows performing inference by computing the inverse and getting a valid joint density (and likelihood) for the underlying Gaussian process at any possible set of observed locations.

Mathematically speaking, Bochner's theorem (introduced in Section 2.3, see Equation 2.1.6) holds also at the space-time level when considering time just as an additional coordinate. As a matter of fact, any of the isotropic covariance functions introduced in Section 2.3 is a valid choice on the product space $\mathbb{R}^{d+1}$, and therefore can be used to model spatio-temporal processes. However, we have already discussed in Section 2.4 how this view is conceptually flawed. Time and space are intrinsically different and dependence along these dimensions must be accounted for with different strategies.

There are different techniques to build valid but structurally well-suited spatio-temporal covariance functions. Some derive from physically inspired probability models, and their specification is particular of the stochastic partial differential equation regulating the underlying physical mechanism (Cox and Isham, 1988; Brillinger, 1997; Brown et al., 2000; Kelbert et al., 2005; Prévôt and Röckner, 2007; Gneiting and Guttorp, 2010). Such covariances are not necessarily stationary, since stationarity may not be satisfied in practice, and may even not be compactly supported. For statistical modeling purposes, especially in the *Geostatistical* and

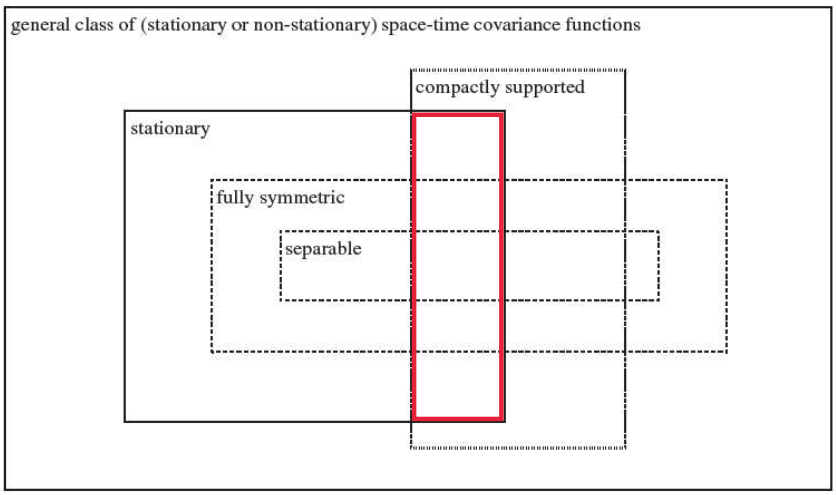


Figure 2.8. Graphical scheme of the classes covariance functions can belong to. In red, the class on which our discussion focuses.

Bayesian Hierarchical framework, the attention is usually restricted to stationary covariance functions with a compact support (see e.g. Cressie and Huang (1999); Kyriakidis and Journel (1999); Gneiting (2002); Allcroft and Glasbey (2003); Stein (2005); Gneiting et al. (2006)). General purposes covariance functions are also often *fully symmetric*, i.e.:

$$c_{\eta,\theta}(\boldsymbol{h},\delta) = c_{\eta,\theta}(-\boldsymbol{h},\delta) = c_{\eta,\theta}(\boldsymbol{h},-\delta) = c_{\eta,\theta}(-\boldsymbol{h},-\delta),$$

for all $(\boldsymbol{h},\delta) \in \mathbb{R}^d \times \mathbb{R}$. Symmetry along the temporal dimension can result particularly unattractive in some practical space-time setting (e.g. time moves only forward, and air or water flows have pre-determined direction in time). We refer to Gneiting et al. (2006) for an extensive reference list of non-symmetric space-time covariance functions.

Figure 2.8 reports a graphical scheme of some key properties of space-time covariance functions. In the sequel, we will discuss the construction and implications of some belonging to the class highlighted in red: stationary, compactly supported, separable and non-separable covariance functions.

Separable covariance functions A straightforward way to construct valid space-time covariance functions stems from the closure of the class of positive definite functions with respect to the operators of sum and product. Therefore, we may construct mathematically valid spatio-temporal covariance functions simply by taking the sum or the product of a purely spatial and a purely temporal covariance function. These two decompositions are:

$$c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h},\delta) = c_{\eta,\boldsymbol{\theta}_s}(\boldsymbol{h}) + c_{\eta,\boldsymbol{\theta}_t}(\delta), \tag{2.4.6}$$

and:

$$c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h},\delta) = c_{\eta,\boldsymbol{\theta}_s}(\boldsymbol{h}) \cdot c_{\eta,\boldsymbol{\theta}_t}(\delta), \tag{2.4.7}$$

for all $(\boldsymbol{h},\delta) \in \mathbb{R}^d \times \mathbb{R}$. Such models arise from two separate processes (one temporal and one spatial) that act independently one from each other (Jones and Zhang, 1997).

The additive decomposition of Equation (2.4.6) corresponds to a *zonal anisotropy model* (e.g. in Bilonick (1985)), and implies that the spatial behavior is exactly the same at all time-instants[10] (Cressie and Huang, 1999). It essentially assumes the underlying spatio-temporal process can be decomposed into the sum of two independent purely spatial and purely temporal processes $\eta_{st}(\cdot,\cdot) = \eta_s(\cdot) + \eta_t(\cdot)$, resulting very prone to identifiability issues of these two components. Indeed, there is no guideline for inferring separately on the two structures, since it would require having observed sets of spatial lags at common temporal lags and viceversa. Furthermore, it can lead to invalid inferences for data collected on specific configurations (e.g. rectangular) when the composing terms are not *strictly* positive definite (Dimitrakopoulos and Luo, 1994). Generally speaking, such an additive structure shall be selected only when there is substantial evidence that the spatial and temporal components do act at different levels.

The multiplicative decomposition of Equation (2.4.7), while sharing similar limitations with the *zonal anisotropy* specifications in terms of identifiability of the two components, does not present general inferential issues. Temporal and spatial terms have a multiplicative interaction, where dependence attenuates across space and time. On the contrary, it is very convenient from a specification point of view since it leaves complete liberty in the choice of suitable spatial and temporal dependence structures. It is is also mathematically congenial, since it admits the representation as a *Kronecker product* of the purely spatial and purely temporal term, in a way analogous to the spatio-temporal Spline Regression model of Section 2.4.1. Indeed, following Rodríguez-Iturbe and Mejía (1974) and then Mardia and Goodall (1993), a frequently used form for Equation (2.4.7) is:

$$c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h},\delta) = \sigma^2 \cdot \rho_{\eta,\boldsymbol{\theta}_s}(\boldsymbol{h}) \cdot \rho_{\eta,\boldsymbol{\theta}_t}(\delta),$$

where σ^2 is a common *sill*, and $\rho_{\eta,\boldsymbol{\theta}_s}(\cdot)$ and $\rho_{\eta,\boldsymbol{\theta}_t}(\cdot)$ are valid spatial and temporal correlation functions. For any set of spatial locations $\mathcal{U}_s = \{\boldsymbol{u}_i\}_{i=1}^{I}$ and time points $\mathcal{U}_t = \{\tau_i\}_{i=1}^{J}$, the full covariance matrix can then be expressed as:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathcal{U}_s \otimes \mathcal{U}_t) = \sigma^2 \cdot \boldsymbol{R}_{\boldsymbol{\theta}_s}(\mathcal{U}_s) \otimes \boldsymbol{R}_{\boldsymbol{\theta}_t}(\mathcal{U}_t),$$

where $\boldsymbol{R}_{\theta_s}(\mathcal{U}_s) = [\rho_{\eta,\boldsymbol{\theta}_s}(\boldsymbol{h}_{ij})]_{ij}$ and $\boldsymbol{R}_{\theta_t}(\mathcal{U}_t)\,[\rho_{\eta,\boldsymbol{\theta}_t}(\delta_{ij})]_{ij}$ are $I \times I$ and $J \times J$ correlation matrices, respectively. Just as in the spatial case, the likelihood evaluation encompasses the computation of determinant and inverse of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathcal{U}_s \otimes \mathcal{U}_t)$ that, by properties of the Kronecker product, can be simplified as:

$$\det\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathcal{U}_s \otimes \mathcal{U}_t)\} = (\sigma^2)^{IJ} \cdot \det\{\boldsymbol{R}_{\boldsymbol{\theta}_s}(\mathcal{U}_s)\}^I \cdot \det\{\boldsymbol{R}_{\boldsymbol{\theta}_t}(\mathcal{U}_t)\}^J,$$

$$\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathcal{U}_s \otimes \mathcal{U}_t)\}^{-1} = \frac{1}{\sigma^2} \cdot \{\boldsymbol{R}_{\boldsymbol{\theta}_s}(\mathcal{U}_s)\}^{-1} \cdot \{\boldsymbol{R}_{\boldsymbol{\theta}_t}(\mathcal{U}_t)\}^{-1},$$

needing only determinant and inverse for an $I \times I$ and $J \times J$ matrix. Kriging (predictions) at unobserved locations can take advantage of the same properties (see Chapter 8.2 of Banerjee et al. (2014) for further details).

Nevertheless, whilst convenient in terms of computation and interpretability, separable models do not allow for arbitrary interaction of the spatial and temporal dimensions (e.g. different spatial dependence at different temporal lags and viceversa) and are necessarily fully symmetric. These yield simplistic dependence patterns that frequently fail to represent the complex spatio-temporal variations of real phenomena. When the practical divide between observed data and the separable model is unbridgeable, resorting to non-separable covariance functions becomes a necessity.

[10] The same holds for the temporal behavior at all spatial locations

Non-Separable covariance functions The naivest approach to build a non-separable space-time covariance function is to consider time just as an additional coordinate and take a valid covariance function on the euclidean space $\mathbb{R}^{d+1}$. However, the fallacy of this king of approach has already been widely discussed in the previous sections.

If the same functional form can sufficiently well describe the marginal variability on both the spatial and temporal dimensions, *geometric anisotropy* is a very convenient strategy to account for different behaviors along the two while maintaining non-separability. Rather than a technique for building a space-time covariance function, it consists of defining a particular metric on $\mathbb{R}^{d+1}$ implying a certain degree of anisotropy:

$$c_{\eta,\theta}\left(\boldsymbol{h},\delta\right)=c^{*}_{\eta,\theta}\left(\sqrt{a_1||\boldsymbol{h}||^2+a_2\delta}\right),$$

where a_1, a_2 are positive constants and $c^{*}_{\eta,\theta}(\cdot)$ is a valid covariance function on $\mathbb{R}^{d+1}$ (Armstrong et al., 1993; Christakos et al., 2000). Whilst representing a very convenient solution, a common functional form is (almost) never able to model appropriately the fundamental differences in the variability along the temporal and the spatial axes.

Another straightforward alternative is *mixing* (Ma, 2002, 2008), as in the *product-sum model* introduced by De Iaco et al. (2001). It expresses the process of interest as the sum of two auxiliary processes, each independent and with its own separable covariance structure (one additive, one multiplicative): $\eta(\boldsymbol{u},\tau)=\eta_1(\boldsymbol{u},\tau)+\eta_2(\boldsymbol{u},\tau)$. The resulting overall spatio-temporal covariance function is:

$$c_{\eta,\theta}\left(\boldsymbol{h},\delta\right)=a_0c^{0}_{\eta,\theta_s}\left(\boldsymbol{h}\right)c^{0}_{\eta,\theta_t}\left(\delta\right)+a_1c^{1}_{\eta,\theta_s}\left(\boldsymbol{h}\right)+c^{2}_{\eta,\theta_t}\left(\delta\right),$$

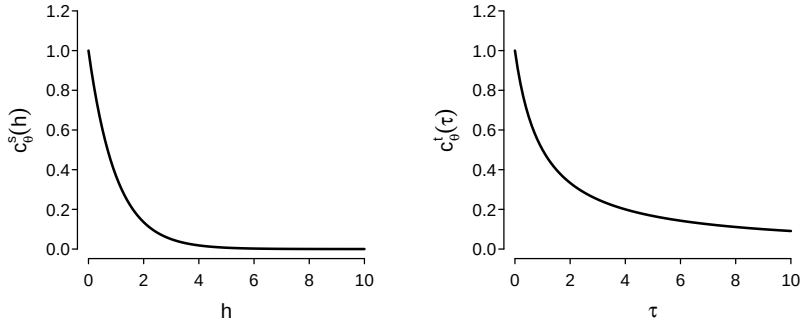
where a_0, a_1 and a_2 are positive constants and $c^{0}_{\eta,\theta_s}(\cdot)$, $c^{0}_{\eta,\theta_t}(\cdot)$, $c^{1}_{\eta,\theta_s}(\cdot)$, $c^{1}_{\eta,\theta_t}(\cdot)$ are purely spatial and purely temporal covariance functions. The result is then clearly non-separable.

More sophisticated methods for building valid non-separable space-time covariance functions are based on the *spectral domain*. Indeed, the only requirements to specify a non-separable second-order structure in the frequency domain are those of *non-negativity* and *integrability* of the spectral density. Cressie and Huang (1999) and Stein (2005) use this strategy to introduce different alternatives of stationary non-separable covariance functions, with the latter extending the well-known *Matérn* family of spatial covariance functions. Unfortunately, suitable specification on the spectral domain is translatable to the original domain only when a closed solution of the Fourier inverse is available, which happens only in very special cases. Gneiting (2002) formulated a criterion based on the same construction technique, but does not depend on the Fourier inversion. He proved that given a *completely monotone*[11] function $g(\cdot):\mathbb{R}^{+}\rightarrow\mathbb{R}$ and a positive function with completely monotone derivative $\psi(\cdot):\mathbb{R}^{+}\rightarrow\mathbb{R}$, then any function of the following form:

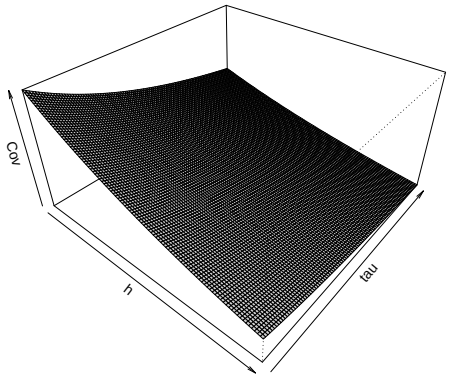
$$c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h},\tau)=\frac{1}{\psi(\tau^2)^{d/2}}g\left(\frac{||\boldsymbol{h}||^2}{\psi(\tau^2)}\right),\quad \boldsymbol{h}\in\mathbb{R}^{d},\ \tau\in\mathbb{R},$$

is a valid space time covariance function over the product space $\mathbb{R}^{d}\times\mathbb{R}$. There are various possible combinations of functions that satisfy the above mentioned

[11] A function $f(\cdot)$ is completely monotone if it possesses derivatives $f^{(n)}(\cdot)$ of all orders and $(-1)^n f^{(n)}(r)>0,\ \forall\, r>0$



(a) Marginals



(b) Complete

Figure 2.9. Reparametrized form of the simplified Gneiting spatio-temporal covariance function of Equation (2.4.9) over $\mathbb{R}^2 \times \mathbb{R}$, for $\tau = 1$, $\beta = 0.8$, $\alpha = 0.5$ and $\gamma = 0.5$

properties, but a very common and useful specification is the following:

$$c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h},\tau) = \frac{\sigma^2}{(1+\tau^{2\alpha})^{\kappa+\beta d/2}} \cdot \exp\left\{-\frac{||\boldsymbol{h}||^{2\gamma}}{(1+\tau^{2\alpha})^{\beta\gamma}}\right\}, \quad \boldsymbol{h} \in \mathbb{R}^d,\ \tau \in \mathbb{R}, \tag{2.4.8}$$

where $\sigma^2, \kappa > 0$ and $\alpha, \gamma \in (0,1]$ are *smoothness* parameters (i.e. the fractal dimension of the spatial and temporal sections) and $\beta \in [0,1]$ is an *interaction* parameter. The spatial margins belong to the *powered exponential* family, and the temporal margin to the *Cauchy* class.

In many instances, Equation (2.4.8) is reparametrized in a simpler form by replacing $\kappa + \beta \cdot d/2$ with a constant $\chi \geq \beta d/2$:

$$c_{\eta,\boldsymbol{\theta}}(\boldsymbol{h},\tau) = \frac{\sigma^2}{(1+\tau^{2\alpha})^{\chi}} \cdot \exp\{-\frac{||\boldsymbol{h}||^{2\gamma}}{(1+\tau^{2\alpha})^{\beta\gamma}}\}, \quad \boldsymbol{h} \in \mathbb{R}^d,\ \tau \in \mathbb{R}. \tag{2.4.9}$$

In this case, the role of the parameter β as *interaction regulator* is even more evident. Indeed, the two marginals are independent of its value and $\beta = 0$ corresponds to a separable model (multiplicative). On the contrary, the larger the value of β and the stronger is the interaction between the two terms. Figure 2.9 shows marginals and joint behavior of this covariance function for a specific choice of the parameters'

values. Both the expression of Equation (2.4.8) and (2.4.9) can be enriched by the inclusion of two positive parameters $a_1, a_2 > 0$ to scale lags along the spatial and temporal dimension.

Once the parametric form of the covariance function has been chosen, and net of eventual computational issues, inference and predictions can be performed in the same exact fashion of the spatial setting (see Section 2.3.1). For further practical details, the reader is referred to Banerjee et al. (2014).

2.5 Big data issues and the Nearest Neighbor Gaussian Process

During its early development, spatial statistics applications were usually designed to model datasets of limited sizes ($n < 10^3$). Considering that an accurate collection of time-stamped and/or geo-referenced data requires technologically advanced and expensive tools and techniques, this is particularly true for data collected on continuous domains. As long as data sizes have been significantly bounded by these technical limitations, the previously discussed estimation methods dis not present any particular computational complication and could be implemented using basic computer programs for matrix manipulations.

However, recent years have seen a rapidly increasing usage and growing capabilities of tracking and monitoring devices, remote sensing technologies, and Geographic Information Systems. Such development provided the scientific community with extraordinary opportunities to understand processes' spatial and temporal complexity at broad scales. Something that earlier was just fantasy developed spawning considerable research to model large *spatial* datasets in diverse disciplines: natural and environmental sciences, economics, biometry, social sciences, etc. Suddenly, in no more than a decade, spatial statistics applications have transitioned from a *data poor* to a *data rich* setting. The scale of such *Big Data* easily extends to the order of millions of spatial and/or temporal points, which can also include a large number of covariates. This is particularly true for spatio-temporal applications, in which the additional temporal dimension can easily make the data-size explode. All these factors insert most of the modern spatial statistics application in the *Big Data* analysis setting, for which traditional methods were not conceived. Indeed, whether one takes a Bayesian or frequentist perspective, fitting of geostatistical models can incur onerous computational costs that severely hinder their implementation for massive datasets. The critical bottleneck stems from storage and computations involving the massive covariance matrices representing the dependence in the data.

For instance, implementing MCMC algorithms to estimate Bayesian hierarchical models requires repeated evaluations of various full-conditional density functions. At the same time, likelihood-based approaches require repeated evaluation of the likelihood inside the optimization routine. This process encompasses the computation of inverse and determinant of the $n \times n$ covariance matrices defining the latent Gaussian process covariance structure (where n is the size of the dataset). These matrices are typically dense and carry no exploitable structure to facilitate computations. Hence the two aforementioned operations require no less than $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ memory to be executed. We will refer to this setting as the *full Gaussian process*. For large n, the computations can then can be very slow, even unfeasible. In practice, even for a modestly large number of observations ($n \approx 10^3$ or greater), computational demands may become prohibitive and preclude inference. Banerjee et al. (2014) informally refers to this situation as *"the big n problem"* (Lasinio et al., 2013).

When this is the case, one must adopt some additional strategies in order to reduce significantly the computational burden of fitting a *full Gaussian process*. A substantial literature in statistics exists on methodologies aimed at achieving that, which is already too vast to be summarized. These methods often exploit low-rank structures in order to approximate the original Gaussian process and/or multi-core and multi-threaded computing environments to hasten computations. We here propose a series of references pointing out to the most successful attempts, in chronological order.

Early solutions attempted at reformulating in a convenient way the original process, rather than work on approximations. This included factoring the joint density (and hence the likelihood and the full-conditionals) into a series of conditional distributions (Vecchia, 1988; Stein et al., 2004), the use of *pseudo-likelihoods* (Varin et al., 2011; Fuentes, 2007) or of tapered covariance functions (Furrer et al., 2006; Kaufman et al., 2008; Stein et al., 2013). However, the reduction in computational burden was not sufficient to keep the pace with the ever-increasing dataset sizes.

Beginning in the late 2000's, several approaches based on low-rank approximations to Gaussian processes were developed (or became popular) including discrete process convolutions (Higdon, 2002; Lemos and Sansó, 2009), fixed rank kriging (Cressie and Johannesson, 2006; Kang et al., 2009; Katzfuss, 2017), predictive processes (Banerjee et al., 2008; Finley et al., 2009), lattice kriging (Nychka et al., 2015), Gaussian Markov random fields with *Integrated Nested Laplace Approximations* (Rue and Held, 2005) and stochastic partial differential equations (Lindgren et al., 2011). Sun et al. (2012), Bradley et al. (2016a) and Liu et al. (2020) provide exceptional reviews of most of these methods, demonstrating their effectiveness for modeling spatial data at large scales. After several years of their use, however, scientists have started to observe shortcomings in many of the above methods for approximating full Gaussian Processes: propensity to over-smooth the data (Simpson et al., 2012; Stein, 2014) and even, for some of these methods, an upper limit on the size of the dataset that can be modeled. The most recent scientific research in this area has focused on the efficient use of modern computing platforms and the development of parallelizable methods. For example, Paciorek et al. (2015) showed how Gaussian Process related computations can be calculated using parallel computing. Katzfuss and Hammerling (2017) and Katzfuss (2017) developed a basis-function approach that lends itself to distributed computing. Alternatively, Barbian and Assunção (2017), and Guhaniyogi and Banerjee (2018) proposed dividing the data into a large number of subsets, draw inference on the subsets in parallel, and then combining the inferences. More recently, Datta et al. (2016a) and Katzfuss and Guinness (2021) build upon Vecchia (1988) by developing novel approaches to factoring the joint model as a series of conditional distributions based only on nearest neighbors. Other recent works that explore this direction are Katzfuss et al. (2020), Katzfuss and Guinness (2021).

Various extensions of each of these methodologies have been proposed in the spatio-temporal settings. For instance, Cressie et al. (2010); Finley et al. (2012); Katzfuss and Cressie (2012) adapt the low-rank approximation, while Xu et al. (2015) opts for the GMRF. However, their solutions are valid only for dynamic models, and hence require the data to be collected at fixed temporal lags. Other similar opportunities for non-Gaussian outcomes are discussed in Bradley et al. (2018, 2020).

Attempts that actually tackle the challenge of continuous space-time modeling of large dataset are instead present in Bai et al. (2012) and Bevilacqua et al. (2012), who used composite likelihoods in order to achieve parameter estimation in such set-up. This approach is limited by the impossibility to provide full inference (i.e. characterize the distribution and uncertainty) at arbitrary locations. Alternatively,

the *Dynamic Nerarest Neighbor Gaussian Process* (DNNGP) proposed in Datta et al. (2016b) as an extension of the NNGP (Datta et al., 2016a) offers a highly scalable spatio-temporal process that provides full inference at all scales for continuous space-time modeling.

The very recent work in Heaton et al. (2019) tries to bring some order to the plethora of available choices to analyze large spatial data-sets. It describes a case-study competition among the *fathers* of some of the most promising techniques mentioned above who agreed and were able to participate. Indeed, the competition took place between the various research groups across the world, each implementing its own method to analyze the same spatial dataset, so that it would be used at its best. Without going too far into the competition's details, all methods have been compared in terms of their prediction ability, uncertainty estimation, and run-time. All performed sufficiently well from various points of view. Nevertheless, the *Nearest Neighbor Gaussian Process* (NNGP) methodology introduced by Datta et al. (2016a) has emerged among the others for its generalized good performances at all tasks combined with an incredibly reduced run-time. It also presents theoretical advantages over many of its competitors. Indeed, the NNGP is not just a mathematical approximation to the Full Gaussian Process it is built upon, but it defines a valid process of its own. For all these reasons, it is the method chosen for the original application in Alaimo Di Loro et al. (2021), described at Chapter 3 of this dissertation.

In the remainder of this Section, a broad introduction of the NNGP methodology is provided, following the seminal and other papers. In particular, Section 2.5.4 introduces the spatio-temporal extension of the method proposed by the same author in Datta et al. (2016b).

2.5.1 The Nearest Neighbor Gaussian Process (NNGP)

The Nearest Neighbor Gaussian process is a *precision sparsity-inducing* method originally introduced in the seminal paper by Datta et al. (2016a). Its introduction is motivated by the computational issues of fitting geostatistical models for massive spatial data-sets. From a practical point of view, these are mostly related to the latent Gaussian process' dense covariance structure defining the underlying dependence structure of the process of interest $Y(\cdot)$.

For example, let us consider the same estimation problem of Section 2.3.3. Under the Full $\mathcal{GP}$ prior on the latent process $\eta(\cdot)$ we generally have that:

$$\boldsymbol{\eta}(\mathcal{S}) \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \boldsymbol{R}_\theta(\mathcal{S})\right).$$

The covariance matrix $\sigma^2 \cdot \boldsymbol{R}_\theta(\mathcal{S})$ is a family of covariance matrices, indexed by an unknown set of parameters θ, that generally presents a full and dense structure. This is also true for the corresponding precision matrix $\left(\sigma^2 \cdot \boldsymbol{R}_\theta(\mathcal{S})\right)^{-1}$, not offering any exploitable structure to facilitate the matrix operations it is involved in. In particular, the computation of the inverse and determinant of the covariance matrix are necessary in order to fit both the *conditional* and the *marginal* models introduced in Section 2.3.3. The former requires these to sample from the latent process full conditional of Equation (2.3.7), while the latter for the exact data likelihood evaluation (in case inference on the latent process is of interest, also for sampling from $\boldsymbol{\eta}(\mathcal{S})$'s posterior distribution). That implies that model fitting typically requires $\mathcal{O}(n^3)$ floating-point operations and $\mathcal{O}(n^2)$ storage, at each MCMC iteration, that rapidly becomes prohibitive for large n.

As presented in the introduction to this section, a large variety of methods to overcome this issue have been proposed in the literature. Loosely speaking, we can

classify them in three categories: *low-rank* models (see e.g. Higdon (2002); Banerjee et al. (2008); Finley et al. (2009) etc.), *covariance sparsity-inducing* methods (Furrer et al., 2006; Kaufman et al., 2008) and *precision sparsity-inducing* methods (Rue and Held, 2005; Vecchia, 1988; Stein et al., 2004; Eidsvik et al., 2014; Katzfuss et al., 2020; Datta et al., 2016a). The NNGP is part of the last class but, differently from the others, defines a valid process that extends to new random variables at arbitrary locations. The latter allows for proper uncertainty quantification of spatial predictions. This is instead the Achille's heel of Gaussian Markov random field approximations (Rue and Held, 2005) and composite likelihood approaches (Eidsvik et al., 2014).

The NNGP methodology's idea is to replace the Gaussian process prior specification on the latent process with a sparsity-inducing spatial process prior derived from the *parent* (original) Gaussian process. Sparsity is achieved by exploiting *neighbors sets* constructed from *Directed Acyclic Graphs* (DAG), starting from the *Vecchia approximation* (Vecchia, 1988). The resulting process is a valid process defined on the whole process domain, whose finite-dimensional distributions are still Normal distributed but have sparse precision matrices available in closed form.

The seminal paper Datta et al. (2016a) introduces the NNGP within the versatile spatially-varying regression framework of Gelfand et al. (2003), considering q-variate Gaussian processes. For the sake of clarity and coherency with the rest of this dissertation, Section 2.3.3 will focus on the univariate case. Generalization to q-variate setting is trivial, and the reader is referred to the seminal paper for insights.

2.5.2 Definition of the NNGP

Let us consider a zero-mean Gaussian process $w(\cdot) : \mathcal{D} \to \mathbb{R}$ defined over $\mathbb{R}^d$:

$$w(\cdot) \sim \mathcal{GP}\left(0, \boldsymbol{c}_\theta(\cdot,\cdot)\right),$$

where $\boldsymbol{c}_\theta(\cdot,\cdot) = \sigma^2 \cdot \rho_\theta(\cdot,\cdot)$ is a valid covariance function that defines uniquely the dependence structure in w. Here, when referring to the covariance function instead of the correlation function, the *sill* σ^2 is included in the set of parameters θ. By basic properties of Gaussian processes, on any finite set of locations $\mathcal{U} = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_l\} \subseteq \mathcal{D}$ the following holds:

$$\boldsymbol{w}(\mathcal{U}) \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{C}_\mathcal{U}\right), \tag{2.5.1}$$

where $\boldsymbol{C}_\mathcal{U} = \boldsymbol{C}_\theta(\mathcal{U})$ is a $l \times l$ positive definite matrix such that $\left[\boldsymbol{C}_\mathcal{U}\right]_{ij} = \boldsymbol{c}_\theta(\boldsymbol{u}_i, \boldsymbol{u}_j),\ i,j = 1,\dots,l$. Let $\boldsymbol{w} = [w_1, \dots, w_l]^\top$ be a realization from $\boldsymbol{w}(\mathcal{U})$. Performing inference through (2.5.1) is computationally burdensome for large l, because the computation of the joint density $f_\theta(\boldsymbol{w})$ entails inverse and determinant of $\boldsymbol{C}_\theta(\mathcal{U})$.

A convenient representation of the joint density is based on the following conditional identity:

$$\begin{aligned} f_\theta(\boldsymbol{w}) &= f_\theta(w_1) \cdot f_\theta(w_2 |\, w_1) \cdots f_\theta(w_l |\, w_1, \dots, w_{l-1}) = \\ &= f_\theta(w_1) \cdot \prod_{i=2}^{l} f_\theta\left(w_i \,|\, w\left(\mathcal{P}_\mathcal{U}(\boldsymbol{u}_i)\right)\right), \end{aligned} \tag{2.5.2}$$

where $\mathcal{P}_\mathcal{U}(\boldsymbol{u}_i) = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{i-1}\}$ is the *full conditioning set* of location i over the locations set $\mathcal{U}$. In practice, Equation (2.5.2) is representing the joint structure of the vector $\boldsymbol{w}$ through a *Directed Acyclic Graph* (DAG) $\mathcal{G} = (\mathcal{U}, \mathcal{P}_\mathcal{U})$ where the corresponding locations $\boldsymbol{u}_i$ are nodes, and edges represent conditional dependence relationships between the process values at the nodes [12].

[12] A Directed Acyiclic Graph is a directed graph with no directed cycles (Bang-Jensen and Gutin,

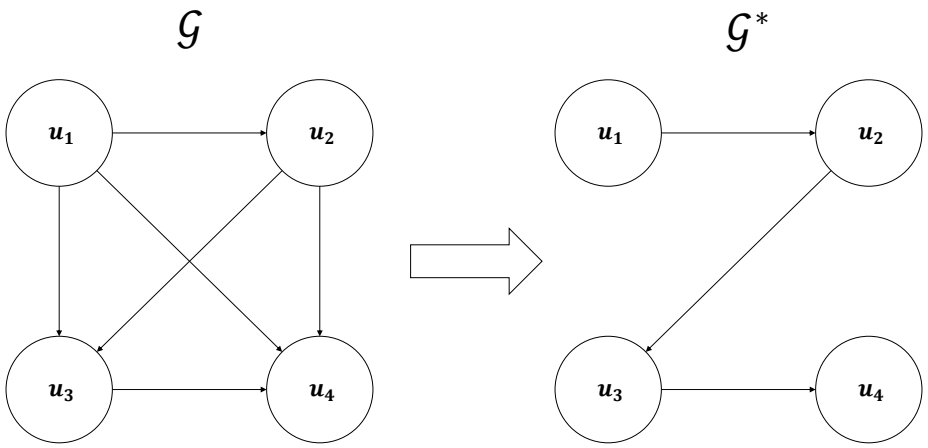


Figure 2.10. Transition from a 4 nodes full DAG $\mathcal{G}$ to the sparse $\mathcal{G}*$, using Vecchia approximation on with $m = 1$ neighbors sets composed of just the previous node.

While convenient for decomposing the dependence pattern in the joint density, (2.5.2) does not provide much help in facilitating computations. Indeed, the conditional densities are again multivariate Normals, and their evaluation encompasses the solution of linear systems involving the covariance matrix of the conditioning set. This may not be an issue for the first terms, but deeper ones present extensive conditioning sets and jeopardize the initial computational advantages.

In light of this observation, as in Vecchia (1988), Stein et al. (2004) and Gramacy and Apley (2015), we may propose to replace the larger conditioning sets in (2.5.2) with smaller, carefully chosen, conditioning sets of size m, where $m << l$. The resulting approximation to the original joint density is:

$$f_\theta(\boldsymbol{w}) \leftarrow \widetilde{f}_\theta(\boldsymbol{w}) = f_\theta(w_1) \cdot \prod_{i=2}^{l} f_\theta\left(w_i | \boldsymbol{w}\left(N_{\mathcal{U}}(\boldsymbol{u}_i)\right)\right), \tag{2.5.3}$$

where $N_{\mathcal{U}}(\boldsymbol{u}_i) \subset \mathcal{U} \setminus \{\boldsymbol{u}_i, \ldots, \boldsymbol{u}_l\}$ is the set of m neighbors of $\boldsymbol{u}_i$ in the conditioning set $\mathcal{P}_{\mathcal{U}}(\boldsymbol{u}_i)$. Therefore, the original DAG $\mathcal{G}$ is modified by cutting edges in such a way that each node has at most m edges pointing at it (i.e. m edges starting from its m neighbors). Let us denote with $\mathcal{G}^* = \{\mathcal{U}, N_{\mathcal{U}}\}$ the new DAG. Figure 2.10 shows the passage from $\mathcal{G}$ to $\mathcal{G}^*$ in an example with 4 nodes ($l = 4$) and neighbor sets of size $m = 1$ composed of the m previous locations.

Being $\mathcal{G}^*$ a DAG, it can be proved that $\widetilde{f}_\theta(\boldsymbol{w})$ is a proper multivariate joint density no matter the initial joint density $f_\theta(\boldsymbol{w})$ (Lauritzen, 1996). This is especially useful in the context of Gaussian processes. Indeed, if $f_\theta(\boldsymbol{w})$ is a multivariate Gaussian (which it is), standard distribution theory reveals that:

$$\widetilde{f}_\theta(\boldsymbol{w}) = \prod_{i=1}^{l} \mathcal{N}\left(w_i \,\middle|\, \boldsymbol{b}_{\boldsymbol{u}_i} \boldsymbol{w}_{N_{\mathcal{U}}(\boldsymbol{u}_i)},\ d_{\boldsymbol{u}_i}\right) \tag{2.5.4}$$

2008)

where:

$$\begin{aligned} \boldsymbol{b}_{\boldsymbol{u}_i} &= \boldsymbol{C}_\theta(\boldsymbol{u}_i, N_{\mathcal{U}}(\boldsymbol{u}_i))\boldsymbol{C}_\theta\left(N_{\mathcal{U}}(\boldsymbol{u}_i)\right)^{-1}, \\ d_{\boldsymbol{u}_i} &= c_\theta(\boldsymbol{u}_i, \boldsymbol{u}_i) - \boldsymbol{C}_\theta(\boldsymbol{u}_i, N_{\mathcal{U}}(\boldsymbol{u}_i))\boldsymbol{C}_\theta\left(N_{\mathcal{U}}(\boldsymbol{u}_i)\right)^{-1}\boldsymbol{C}_\theta(N_{\mathcal{U}}(\boldsymbol{u}_i), \boldsymbol{u}_i), \end{aligned} \tag{2.5.5}$$

and $\boldsymbol{C}_\theta(\cdot,\cdot)$ denotes the cross-covariance between the elements of its two arguments. This product of Gaussian conditional densities yields a multivariate Gaussian as joint distribution:

$$\widetilde{f}_\theta(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w} \;\middle|\; \boldsymbol{0}, \widetilde{\boldsymbol{C}}_{\mathcal{U}}\right), \tag{2.5.6}$$

where $\widetilde{\boldsymbol{C}}_{\mathcal{U}}$ is such that its inverse $\widetilde{\boldsymbol{C}}_{\mathcal{U}}^{-1}$ is sparse with at most $l \cdot m \cdot (m+1)/2$ non-zero entries. Indeed, the explicit expression of the joint density in (2.5.6) is proportional to:

$$\frac{1}{\prod_{i=1}^{l} d_{\boldsymbol{u}_i}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{l}\left(w_i - \boldsymbol{b}_{\boldsymbol{u}_i}\boldsymbol{w}_{N_{\mathcal{U}}(\boldsymbol{u}_i)}\right)^\top \frac{1}{d_{\boldsymbol{u}_i}}\left(w_i - \boldsymbol{b}_{\boldsymbol{u}_i}\boldsymbol{w}_{N_{\mathcal{U}}(\boldsymbol{u}_i)}\right)\right\}, \tag{2.5.7}$$

where some manipulation may be necessary in order to clearly visualize the typical Gaussian kernel. Let us denote with $\boldsymbol{b}_i$ the $1 \times l$ vector defined in the following way:

$$[\boldsymbol{b}_i]_j = \begin{cases} 1 & i = j \\ -[\boldsymbol{b}_{\boldsymbol{u}_i}]_d & \boldsymbol{u}_j = [N_{\mathcal{U}}(\boldsymbol{u}_i)]_d \text{ for some } d \\ 0 & \text{otherwise} \end{cases}$$

so that $\left(w_i - \boldsymbol{b}_{\boldsymbol{u}_i}\boldsymbol{w}_{N_{\mathcal{U}}(\boldsymbol{u}_i)}\right) = \boldsymbol{b}_i\boldsymbol{w}$. Stacking all the $\boldsymbol{b}_i$'s altogether into $\boldsymbol{B}_{\mathcal{U}} = \left[\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_l^\top\right]^\top$, the exponent of (2.5.7) can be expressed in terms of the precision matrix in the following compact way:

$$-\frac{1}{2}\left(\boldsymbol{w}^\top\widetilde{\boldsymbol{C}}_{\mathcal{U}}^{-1}\boldsymbol{w}\right) = -\frac{1}{2}\left(\boldsymbol{w}^\top\boldsymbol{B}_{\mathcal{U}}{}^\top\boldsymbol{D}_{\mathcal{U}}{}^{-1}\boldsymbol{B}_{\mathcal{U}}\boldsymbol{w}\right), \tag{2.5.8}$$

where $\boldsymbol{D}_{\mathcal{U}} = \operatorname{diag}\left(d_{\boldsymbol{u}_1}, \ldots, d_{\boldsymbol{u}_l}\right)$. In particular, given the structure of the $\boldsymbol{b}_i$'s building up $\boldsymbol{B}_{\mathcal{U}}$, the (i,j)-th element of the inverse with $i < j$ is such that:

$$\left[\widetilde{\boldsymbol{C}}_{\mathcal{U}}^{-1}\right]_{ij} = \sum_{k=j}^{l}\left(\frac{1}{d_{\boldsymbol{u}_k}}[\boldsymbol{b}_k]_i \cdot [\boldsymbol{b}_k]_j\right), \tag{2.5.9}$$

where the expression on the right-hand side is different from 0 if and only if there is one $k > j$ such that $\boldsymbol{u}_i \in N_{\mathcal{U}}(k)$ and $\boldsymbol{u}_j$ is either equal to $\boldsymbol{u}_k$ or belongs to $N_{\mathcal{U}}(k)$. Since every neighbor set has at most m elements, there are at most $l \cdot m \cdot (m+1)/2$ such pairs, and hence the sparsity of $\widetilde{\boldsymbol{C}}_{\mathcal{U}}^{-1}$.

All these considerations are valid for any criteria to select neighbors $N_{\mathcal{U}}(\cdot)$ into the conditioning sets $\mathcal{P}_{\mathcal{U}}(\cdot)$, since this always ensures the needed DAG structure. A natural choice may be the one of selecting the closest ones in terms of some distance between the locations (Vecchia, 1988; Stroud et al., 2017), but this is quite arbitrary and any other logic may be applied according to application-specific features (see, e.g., Peruzzi et al. (2020a)). Whatever the choice, the resulting neighboring sets always depend on the ordering of the locations. This is not an issue in temporal applications, but spatial locations are not ordered naturally. Therefore, one shall impose an order (e.g., according to one of the coordinates, random order, etc.). Stein

et al. (2004) and Gramacy and Apley (2015) have shown that this does not have a discernible impact on the accuracy of the approximation. As a matter of fact, the approximation's effectiveness is mostly determined by the size of the neighbor sets and not by the specific ordering. In particular, for $m = (l-1)$, the approximation is perfect and the original joint density of (2.5.2) is obtained. Thus the NNGP can approximate, well at will, the joint density over a finite set of locations of a parent Gaussian process. That is something that a lot of other techniques, even simpler than this one, can achieve. However, the great potentiality of the NNGP resides in $\widetilde{f}_\theta(\cdot)$ defining a legitimate process on the original GP's space.

Let us consider a set $\mathcal{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_g\} \subset \mathcal{D}$ such that $\mathcal{V} = \mathcal{T} \setminus \mathcal{U}$ and let us denote with $\boldsymbol{w}_\mathcal{T}$ and $\boldsymbol{w}_\mathcal{V}$ realizations of $\boldsymbol{w}(\mathcal{T})$ and $\boldsymbol{w}(\mathcal{V})$ respectively. Consistently with the definition of $N_\mathcal{U}$, let $\{N_\mathcal{U}(\boldsymbol{v}_i) : \boldsymbol{v}_i \in \mathcal{V}\}$ be the set of m-nearest neighbors to the elements of $\mathcal{V}$ in $\mathcal{U}$. Akin to (2.5.3), we can define the nearest neighbors density of $\boldsymbol{w}_\mathcal{V}$ conditional on $\boldsymbol{w}$ as:

$$\widetilde{f}_\theta(\boldsymbol{w}_\mathcal{V}|\boldsymbol{w}) = f_\theta(\boldsymbol{w}_{\mathcal{V}i}) \cdot \prod_{i=2}^{r} f_\theta\left(\boldsymbol{w}_{\mathcal{V}i}|\boldsymbol{w}_{N_\mathcal{U}(\boldsymbol{v}_i)}\right).$$

This ensures a proper conditional density on any subset of $\mathcal{D}$ as $\mathcal{T}$, indeed:

$$\widetilde{f}_\theta(\boldsymbol{w}_\mathcal{T}) = \int\int \widetilde{f}_\theta(\boldsymbol{w}_\mathcal{V}|\boldsymbol{w})\widetilde{f}_\theta(\boldsymbol{w}) \prod_{i:\boldsymbol{u}_i \in \mathcal{U}\setminus\mathcal{T}} \partial w_i, \tag{2.5.10}$$

where $\widetilde{f}_\theta(\boldsymbol{w}_\mathcal{V}|\boldsymbol{w}) = 1$ if $\mathcal{V}$ is empty and integration is not necessary if $\mathcal{U} \setminus \mathcal{T} = \emptyset$. The probability densities defined as in (2.5.10), defined on topologies, conform to Kolmogorov's consistency criteria and thus correspond to a valid process over $\mathcal{D}$ (technical details included in the supplementary material of Datta et al. (2016a)). So, given a reference set $\mathcal{U}$ and a parent process, it is possible to construct a new process over $\mathcal{D}$ just exploiting a collection of neighbors in $\mathcal{U}$. This is called the *Nearest Neighbor Process*.

In particular, for a $\mathcal{GP}(0, c_\theta(\cdot,\cdot))$ parent process, the following holds:

$$\widetilde{f}_\theta(\boldsymbol{w}_\mathcal{V}|\boldsymbol{w}) = \prod_{i=1}^{r} \mathcal{N}\left(\boldsymbol{w}_{\mathcal{V}i} \,\middle|\, \boldsymbol{b}_{\boldsymbol{v}_i}\boldsymbol{w}_{N_\mathcal{U}(\boldsymbol{v}_i)},\ d_{\boldsymbol{v}_i})\right) = \mathcal{N}\left(\boldsymbol{w}_\mathcal{V} \,|\, \boldsymbol{B}_\mathcal{V}\boldsymbol{w},\ \boldsymbol{D}_\mathcal{V}\right),$$

where $\boldsymbol{b}_{\boldsymbol{v}_i}$, $d_{\boldsymbol{v}_i}$, $\boldsymbol{B}_\mathcal{V}$ and $\boldsymbol{D}_\mathcal{V}$ are defined analogously to their counterparts in (2.5.5). In particular, $\boldsymbol{D}_\mathcal{V} = \text{diag}\left(d_{\boldsymbol{v}_1}, \ldots, d_{\boldsymbol{v}_r}\right)$ and $\boldsymbol{B}_\mathcal{V}$ is a sparse $r \times l$ matrix with each row having at most m non-zero entries.

So, the result of the integration in (2.5.10) is:

$$\widetilde{f}_\theta(\boldsymbol{w}_\mathcal{T}) = \mathcal{N}\left(\boldsymbol{w}_\mathcal{T} \,\middle|\, \boldsymbol{0},\ \widetilde{\boldsymbol{C}}_\theta(\mathcal{T})\right), \tag{2.5.11}$$

where $\widetilde{\boldsymbol{C}}_\theta(\mathcal{T})$ is the covariance matrix obtained through the application of the following cross-covariance function:

$$\widetilde{c}_\theta(\boldsymbol{t}, \boldsymbol{t}') = \begin{cases} \left[\widetilde{\boldsymbol{C}}_\mathcal{U}\right]_{i,j} & \boldsymbol{t} = \boldsymbol{u}_i,\ \boldsymbol{t}' = \boldsymbol{u}_j \\ \boldsymbol{b}_{\boldsymbol{t}}\left[\widetilde{\boldsymbol{C}}_\mathcal{U}\right]_{N_\mathcal{U}(\boldsymbol{t}'),\boldsymbol{u}_j} & \boldsymbol{t} \notin \mathcal{U},\ \boldsymbol{t}' = \boldsymbol{u}_j \\ \boldsymbol{b}_{\boldsymbol{t}}\left[\widetilde{\boldsymbol{C}}_\mathcal{U}\right]_{N_\mathcal{U}(\boldsymbol{t}),N_\mathcal{U}(\boldsymbol{t}')}\boldsymbol{b}_{\boldsymbol{t}'}{}^\top + I_{\boldsymbol{t}}(\boldsymbol{t}') \cdot d_{\boldsymbol{t}} & \boldsymbol{t}, \boldsymbol{t}' \notin \mathcal{U} \end{cases}, \tag{2.5.12}$$

with $\boldsymbol{t}$ and $\boldsymbol{t}'$ arbitrary locations in $\mathcal{D}$, $\left[\widetilde{\boldsymbol{C}}_{\mathcal{U}}\right]_{A,B}$ denoting submatrices of $\widetilde{\boldsymbol{C}}_{\mathcal{U}}$ indexed by locations in the sets A and B and $I_A(\cdot)$ indicator function over the set A. This completes the construction of a well-defined and non-degenerate process over $\mathcal{D}$, derived from a parent $\mathcal{GP}(0, c_\theta(\cdot,\cdot))$, that is named *Nearest Neighbor Gaussian Process* $NNGP\left(0, \widetilde{c}_\theta(\cdot,\cdot)\right)$. Let us here point out that the reference set $\mathcal{U}$ on which the NNGP is defined can be large at will, even larger than the data-set itself: reduction in computational complexity is still achieved by sparsity and only depends on the number of neighbors m.

In the next section, we will see how we can use the NNGP to facilitate computation in the Bayesian hierarchical geostatistical models' estimation process.

2.5.3 Bayesian implementation of the NNGP

Let us consider the Geostatistical setting of Section 2.3.3, where the observed vector of outcomes $\boldsymbol{y} = \{y_1, \dots, y_n\}$ at locations $\mathcal{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$ is assumed to be a realization from:

$$Y(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta} + \eta(\boldsymbol{s}_i) + \upsilon(\boldsymbol{s}_i), \quad \boldsymbol{s}_i \in \mathcal{S} \subseteq \mathcal{D},$$

where $\boldsymbol{x}(\cdot)$ is a $p \times 1$ vector of spatially varying covariates, $\eta(\cdot)$ is a latent process generally assumed to be a $\mathcal{GP}(0, c_\theta(\cdot,\cdot))$ and $\upsilon(\cdot) \sim \mathcal{GP}(0, \tau^2)$. This generally yields a covariance matrix over the observed set with a full-dense inverse, both at the latent ($\boldsymbol{C}_\theta(\mathcal{S})^{-1} = \boldsymbol{C}_\theta^{-1}$) and at the observables ($\boldsymbol{\Sigma}_\theta(\mathcal{S})^{-1} = \boldsymbol{\Sigma}_\theta^{-1} = (\boldsymbol{C}_\theta + \tau^2 \boldsymbol{I}_n)^{-1}$) levels. The bottlenecks of the estimation for the conditional and marginal models introduced in Section 2.3.3 were the following.

- **Conditional model**. As by Equations (2.3.7) and (2.3.8) the block update of $\boldsymbol{\eta}$ requires the computation of $|\boldsymbol{E}|$ and $\boldsymbol{E}^{-1}$ where:

$$\boldsymbol{E} = \left(\frac{1}{\tau^2}\boldsymbol{I}_n + \boldsymbol{C}_\theta^{-1}\right).$$

- **Marginal model**. Both the marginal likelihood for the block-metropolis update of θ and the Gibbs update of the $\boldsymbol{\beta}$ coefficients require the evaluation of $|\boldsymbol{\Sigma}_\phi|$ and $\boldsymbol{\Sigma}_\phi^{-1}$ with:

$$\boldsymbol{\Sigma}_\phi = \left(\tau^2 \boldsymbol{I}_n + \boldsymbol{C}_\theta\right),$$

  where $\phi = \{\sigma^2, \theta, \tau^2\}$.

For scalability, the $\mathcal{GP}$ prior on the latent process can be replaced with a *Nearest Neighbor Gaussian process* prior:

$$\eta(\cdot) \;\leftarrow\; \widetilde{\eta}(\cdot) \sim NNGP\left(0, \widetilde{c}_\theta(\cdot)\right),$$

where $\widetilde{c}_\theta(\cdot)$ is the sparsity inducing covariance function derived from $c_\theta(\cdot)$ through the *Nearest Neighbor* approximation over a reference set $\mathcal{U} \subseteq \mathcal{D}$. Generally speaking, since it does not have any discernible impact on the estimation accuracy (Datta et al., 2016a), we may use any arbitrary reference set to define the NNGP. Here, we will consider the case in which the reference set for the NNGP coincides with the observation set $\mathcal{U} = \mathcal{S}$.

The NNGP prior on the latent process allows to replace the originally dense inverse $\boldsymbol{C}_\theta^{-1}$ with the sparse $\widetilde{\boldsymbol{C}}_\theta^{-1}$. This has an immediate advantage in the setting

of the conditional model. Indeed, sparsity of $\widetilde{\boldsymbol{C}}_\theta^{-1}$ can be leveraged to compute its sparse Cholesky factorization:

$$\widetilde{\boldsymbol{C}}_\theta^{-1} = (\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{D}^{-1} (\boldsymbol{I} - \boldsymbol{A}),$$

directly from $\boldsymbol{C}_\theta$. This entails solving $n-1$ linear systems of size at most $m \times m$, which can be performed in $\mathcal{O}(n \times m^3)$ operations ($m << n$ implies $n \times m^3 << n^3$). Indeed, elaborating on Equation (2.5.8), the matrix $\boldsymbol{A}$ is lower triangular with at most m elements at each row i (the non-zero values corresponds to columns of $\boldsymbol{s}_i$'s neighbors). These can be computed as:

$$[\boldsymbol{A}]_{i,N_\mathcal{S}(\boldsymbol{s}_i)} = \left([\boldsymbol{C}_\theta]_{N_\mathcal{S}(\boldsymbol{s}_i),N_\mathcal{S}(\boldsymbol{s}_i)}\right)^{-1} [\boldsymbol{C}_\theta]_{N_\mathcal{S}(\boldsymbol{s}_i),i}, \quad i = 1, \ldots, n. \tag{2.5.13}$$

Furthermore, $\boldsymbol{D}$ is a diagonal matrix with elements on the diagonal given by:

$$[\boldsymbol{D}]_{ii} = [\boldsymbol{C}_\theta]_{ii} - [\boldsymbol{C}_\theta]_{N_\mathcal{S}(\boldsymbol{s}_i),i} [\boldsymbol{A}]_{i,N_\mathcal{S}(\boldsymbol{s}_i)}, \quad i = 1, \ldots, n. \tag{2.5.14}$$

Operations in (2.5.13) and (2.5.14) can also be performed in parallel as the computation of each element is independent from the other, allowing for efficient routines that can sensibly improve on the global run-time.

The sparse-Cholesky factorization of $\widetilde{\boldsymbol{C}}_\theta^{-1}$, other than being cheap in its own computation, facilitates the computation of the determinant $\widetilde{\boldsymbol{C}}_\theta^{-1} = \prod_{i=1}^n [\boldsymbol{D}]_{ii}$ and of linear systems or quadratic forms involving $\widetilde{\boldsymbol{C}}_\theta$ or $\widetilde{\boldsymbol{C}}_\theta^{-1}$ (that can be computed in $\mathcal{O}(n \times m)$). In particular, these computational advantages are preserved also for $\widetilde{\boldsymbol{E}} = \left(\frac{1}{\tau^2}\boldsymbol{I}_n + \widetilde{\boldsymbol{C}}_\theta^{-1}\right)$, since $\frac{1}{\tau^2}$ is diagonal and preserves $\widetilde{\boldsymbol{C}}_\theta^{-1}$ sparsity pattern. Exploiting it, the computation time required for sampling from the full-conditional $\boldsymbol{w}$ in its sequential update into the *conditional model* MCMC machinery is incredibly lowered. It also reduces the burden of sampling from the posterior predictive distribution at un-sampled locations $\widetilde{\boldsymbol{\eta}}$ through (2.3.10). The adoption of the NNGP prior in the context of the conditional model yields what is known as the *Sequential NNGP*.

On the other hand, the computational advantages of the NNGP are not so apparent in the case of the *marginal model*. Indeed, even if $\widetilde{\boldsymbol{C}}_\theta^{-1}$ is sparse and has a sparse-Cholesky factorization, the corresponding:

$$\widetilde{\boldsymbol{\Sigma}}_\phi = \left(\tau^2 \boldsymbol{I}_n + \widetilde{\boldsymbol{C}}_\theta\right),$$

does not enjoy such a convenient factorization and, in general, its inverse $\widetilde{\boldsymbol{\Sigma}}_\phi^{-1}$ is not even guaranteed to be sparse. However, using the Sherman-Woodboory-Morrison formula (SWM), one can write:

$$\widetilde{\boldsymbol{\Sigma}}_\phi^{-1} = \frac{1}{\tau^2}\boldsymbol{I}_n - \frac{1}{\tau^4}\widetilde{\boldsymbol{E}}^{-1}.$$

This implies that the determinant, linear systems and quadratic forms involving $\widetilde{\boldsymbol{\Sigma}}_\phi$ and $\widetilde{\boldsymbol{\Sigma}}_\phi^{-1}$ can be computed in terms of $\widetilde{\boldsymbol{E}}$, thus sharing its same computational advantages and settling to the same computational complexity of the *Sequential NNGP* ($\mathcal{O}(n \times m^3)$). This formulation of the NNGP, which works marginally on the latent process, is known as the *Collapsed NNGP*.

The computational complexity in both settings highlights how the choice of the neighbor sets' size regulates the trade-off between the accuracy of the approximation

and the computational advantages of the NNGP. As just shown, the computational complexity is linear in the number of observed locations n, but it is cubic in the number of neighbors m. Therefore, for n very large, a low number of m is critical in reducing the required operations' number. Simulation studies in Datta et al. (2016a) show that values of $m \approx 10$ provide inference almost indistinguishable to full geostatistical models, either when the model is or is not well-specified. Generally, a precautionary recommendation is to pick $m \geq 10$, but $m < 20$ typically suffices, and greater values are usually ineffective in improving estimation and prediction accuracy.

For a more detailed description of the MCMC algorithms implementation of the *Sequential* and *Collapsed NNGP* and for ulterior alternative formulations (e.g. *Response NNGP*, *Conjugate NNGP*), the reader is referred to Finley et al. (2019). Please note that the *Sequential NNGP*, *Response NNGP* and *Conjugate NNGP* are already implemented in a user-friendly R-package named `spNNGP` (Finley et al., 2017a).

2.5.4 The Spatio-temporal NNGP

Datta et al. (2016b) introduces a *Dynamic* version of the NNGP that can account for separable or non-separable dependence structure in spatio-temporal Gaussian processes. It provides a highly-scalable framework for the space-time setting, which is currently scant of effective alternatives able to provide full inference on the underlying process. Indeed, as its only spatial version, it delivers a substantially superior approximation than other state of the art techniques such as *low-rank* approximations (see Stein (2014)).

Theoretical derivations and computational complexity are completely analogous to the ones of the purely spatial NNGP. Let $w(\cdot,\cdot) : \mathcal{D} \times \mathcal{G} \to \mathbb{R}$ defined over $\mathbb{R}^d \times \mathbb{R}$ be:

$$w(\cdot) \sim \mathcal{GP}\left(0, \boldsymbol{c}_\theta(\cdot,\cdot)\right),$$

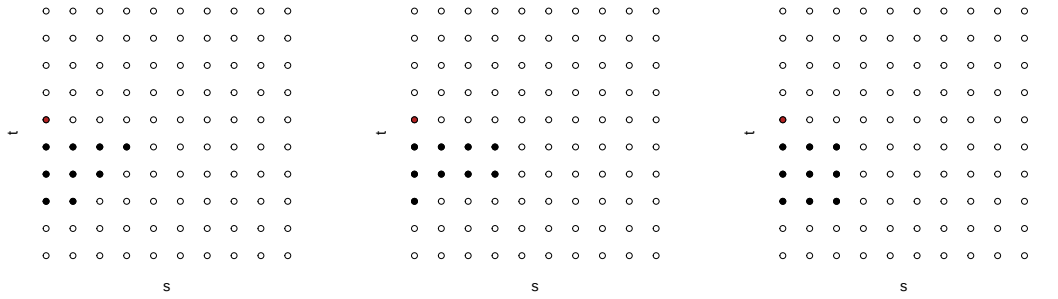
where $\boldsymbol{c}_\theta(\cdot,\cdot)$ is an arbitrary spatio-temporal covariance function. Let us then consider a set of space-time locations $\mathcal{U} = \{(\boldsymbol{u}_i, \tau_i)\}_{i=1}^l \subseteq \mathbb{R}^d \times \mathbb{R}$. As for spatial processes, also spatio-temporal Gaussian processes are such that:

$$w(\mathcal{U}) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{C}_\mathcal{U}\right),$$

where $\boldsymbol{C}_\mathcal{U} = \boldsymbol{C}_\theta(\mathcal{U}) = \left[(\boldsymbol{u}_i, \tau_i\right]_{ij}$ is the $l \times l$ covariance matrix. Let $\boldsymbol{w} = \left[w(\boldsymbol{u}_1, \tau_1), \ldots, w(\boldsymbol{u}_l, \tau_l)\right]^\top$ be a realization from $w(\mathcal{U})$ and construct space-time neighbor sets $\{N(\boldsymbol{u}_i, \tau_i)\}_{i=1}^l$ of size $|N(\boldsymbol{u}_i)| = m << l$ for each $\boldsymbol{u}_i \in \mathcal{U}$. Following the derivations of Section 2.5.2, Equations (2.5.3) and (2.5.4) can be applied in sequence. As a matter of fact, the application of the former is even more natural since the temporal dimension is characterized by a natural ordering among observations (absent in the solely spatial case). The resulting outcome is analogous to the one of Equation (2.5.6):

$$\widetilde{\boldsymbol{w}} \sim \widetilde{f}_\theta(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w} \;\middle|\; \mathbf{0}, \widetilde{\boldsymbol{C}}_\mathcal{U}\right),$$

where $\widetilde{\boldsymbol{C}}_\mathcal{U}$ is such that its inverse $\widetilde{\boldsymbol{C}}_\mathcal{U}^{-1}$ is sparse with at most $l \cdot m \cdot (m+1)/2$ non-zero entries. Storage and computation of the inverse and determinant of this matrix is then reduced to $\mathcal{O}(l \cdot m^3)$. Using Equations (2.5.10) and (2.5.11), the validity of the resulting process extends to arbitrary set of space-time locations in $\mathcal{D} \times \mathcal{G}$,



(a) $\theta = 1$ **(b)** $\theta = 4$ **(c)** Simple neighborhood

Figure 2.11. True and simple neighbor sets on a 20×20 spatio-temporal datasets with one-dimensional spatial domain and covariance function $c_\theta\left((s_i, t_i), (s_j, t_j)\right) = \exp\left\{-|s_i - s_j|^2 - \theta|t_i - t_j|^2\right\}$. Points are ordered from left tow right, from bottom to top. Hence only points below the chosen one (in red) belong to its conditioning set. Black points belong to the corresponding neighboring set, with size fixed to $m = 9$. We see how for increasing values of θ distance along the temporal dimension becomes more relevant.

with a structure of the approximated cross-covariance function $\widetilde{c}_\theta(\cdot, \cdot)$ analogous to Equation (2.5.12). This process is denoted as:

$$\widetilde{w}(\cdot) \sim DNNGP\left((\mathbf{0}, 0),\, \widetilde{c}_\theta(\cdot, \cdot)\right). \tag{2.5.15}$$

So far, the extension of the NNGP methodology to the space-time setting required only some notation adaptation. Nevertheless, we just glossed over the conceptual question about the proper definition of *space-time neighbor sets*. Indeed, spatial correlation functions usually decay with distance and therefore the definition of neighbor sets based on the inter-site distance is completely natural. On the contrary, the definition of a reasonable criteria for establishing neighborhood relationships between space-time locations is not as trivial. Spatial lags are not comparable to temporal lags, hence the naive euclidean distance in $\mathbb{R}^{d+1}$ is not a viable choice. Following the *first law of geography*, closer points shall present higher covariance $c_\theta(\cdot, \cdot)$. Thus covariance values themselves may be used as a general distance metric:

$$d_\theta\left((\boldsymbol{u}_i, \tau_i), (\boldsymbol{u}_j, \tau_j)\right) \propto \left(c_\theta(\boldsymbol{u}_i - \boldsymbol{u}_j, \tau_i - \tau_j)\right)^{-1}, \quad \forall\, (\boldsymbol{u}_i, \tau_i), (\boldsymbol{u}_j, \tau_j) \in \mathcal{D} \times \mathcal{G}.$$

However, isotropic spatio temporal covariance functions (non-separable or separable) depend on both the spatial and temporal lag, and the decay and interaction between these two dimensions is regulated by some set of *unknown* parameters θ. This often precludes the definition of a universal distance metric $d\left((\cdot, \cdot), (\cdot, \cdot)\right) : \left(\mathbb{R}^d \times \mathbb{R}\right)^2 \to \mathbb{R}$ such that $c_\theta(\cdot, \cdot)$ will be monotonic with respect to $d\left((\cdot, \cdot), (\cdot, \cdot)\right)$ for all choices of θ. Loosely speaking, this means that the distance (thus the neighborhood relationships between locations) depend on the unknown value assumed by the parameters regulating the covariance function, impeding an absolute choice of static neighbor sets. This is clear from Figures 2.11a and 2.11b, that show the drastic change of the neighboring structure at variation of the parameter θ in a simplified framework. Datta et al. (2016b) provides two solutions to overcome this issue.

Simple neighbor sets Neighboring structures are defined by treating separately and giving equal importance to the temporal and spatial dimension. Given a neighbor

set of size m which is a perfect square, the $\sqrt{m}$ nearest neighbor in space and the $\sqrt{m}$ nearest neighbors in time are selected to compose the complete neighborhood. Let us take a reference set over a regular space-time grid, where $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^{I}$ and $\mathcal{T} = \{t_j\}_{i=1}^{J}$ are the two sets of spatial and temporal locations, respectively. Let $\mathcal{C}_i = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_{i-1}\}$ be the conditioning sets of each spatial location given a certain order, and $N_s(\boldsymbol{s}, \mathcal{C}, m)$ be the set of m neighbors of $\boldsymbol{s}$ in $\mathcal{C}$. For any point $(\boldsymbol{s}_i, t_j) \in \mathcal{S} \times \mathcal{T}$ the simple neighbor set is defined as:

$$N(\boldsymbol{s}_i, t_j) = \left(\cup_{k=1}^{\sqrt{m}-1} \left\{(\boldsymbol{s}, t_{j-k} | \boldsymbol{s} \in N_s(\boldsymbol{s}_i, \mathcal{S}, \sqrt{m})\right\}\right) \cup \left\{(\boldsymbol{s}, t_j | \boldsymbol{s} \in N_s(\boldsymbol{s}_i, \mathcal{C}_i, \sqrt{m})\right\}. \tag{2.5.16}$$

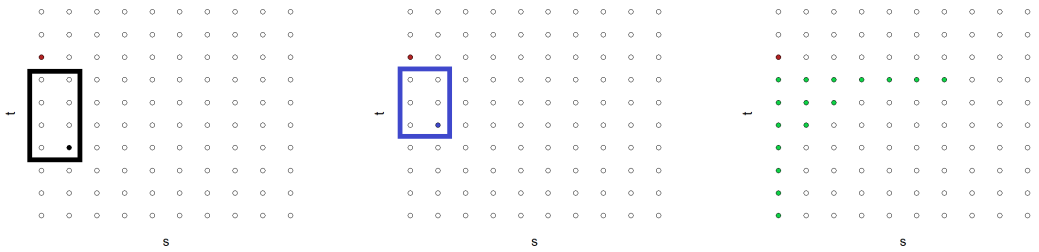
This construction implies that the neighbor set of any point in $\mathcal{S} \times \mathcal{T}$ includes the $\sqrt{m}$ nearest neighbors of the preceding $\sqrt{m}$ time points (including itself at previous times). For any other points $(\boldsymbol{s}, t) \notin \mathcal{S} \times \mathcal{T}$ (e.g. any $(\boldsymbol{u}, \tau) \in \mathcal{U}$), the neighbor set in the reference set is simply defined as the Cartesian product of the $\sqrt{m}$ nearest neighbors in space and time. An example of such neighborhood is reported in Figure 2.11c. A good property of such neighboring structure is that, in case the chosen covariance function satisfies *natural monotonicity*[13], then neighborhoods constructed as in Equation (2.5.16) are guaranteed to contain at least the actual $\sqrt{m} - 1$ nearest neighbors for any value of the parameter θ.

Dynamic neighbor sets The simple neighbor scheme has a clear advantage from a computational point of view, since neighbors sets stay fixed along MCMC iterations. However, it is not exact and in some cases may contain very few of the *true* neighbors (see Figure 2.11). This issue is exacerbated when the correlation decays rapidly in one direction but not in the other. Ideally, if θ was known, one could have fixed neighbor sets $\{N_\theta(\boldsymbol{u}_i, \tau_i)\}_{i=1}^{l}$ by evaluating the pairwise covariances between any point in the reference set $(\boldsymbol{u}_i, \tau_i) \in \mathcal{U}$ and all those in its conditioning set $\mathcal{C}_i$, yielding dynamic neighbor sets $\{N_\theta(\boldsymbol{u}_i, \tau_i)\}_{i=1}^{l}$.

While θ is generally unknown, it does assume a specific value at each iteration of the typical MCMC machinery used for estimation in a Bayesian setting. This permits the consideration of an adaptive neighbor selection scheme where the neighbor sets $N_\theta(\cdot)$ dynamically vary at each iteration of the MCMC algorithm, readjusting to the new values assumed by θ. Whilst attractive from a theoretical point of view, this approach requires a search for neighbors at every iteration and becomes more and more computationally challenging as the conditioning set grows. For instance, given n observed data-points and a l-sized reference set, the flops amount to $\mathcal{O}(l^2 + nl)$ at each iteration. Indeed, the conditioning sets of the l-th point in the reference set and of the n observed locations contain $l - 1$ and l points respectively, which are all possible neighbors. Hence, for large datasets, this kind of adaptive strategy becomes rapidly unfeasible. The inefficiency of this approach is clearly related to the size of the set of *eligible* points for the neighborhood on which the search is performed, that is generally composed of all points in the conditioning set. Jones and Zhang (1997) tries to alleviate this difficulty by permitting only points within a prefixed (small) temporal lag to be neighbors, but this assumption inevitably fails to capture any long term temporal dependence.

Exploiting the same logic, Datta et al. (2016b) instead formulated an algorithm that can efficiently update the neighbor sets at every update of the parameter θ cropping in advance the set of eligible points, while still guaranteeing the presence

[13] $c_\theta(h, \delta)$ is decreasing in h for fixed δ and viceversa. All Matèrn based space-time covariance functions have this property Stein et al. (2013); Omidi and Mohammadzadeh (2016)



(a) Ineligible point **(b)** Eligible point **(c)** Eligible set

Figure 2.12. Construction of eligible sets for finding nearest neighbor sets of size $m = 7$: In figure (a), the black point is ineligible because there are $7 \geq m = 7$ other points closer both in space and time. In figure (b), the blue point is eligible because there are only $6 < m = t$ other points closer both in space and time. Figure (c) shows the final eligible set obtained by applying the same logic to all points in the conditioning set.

of the true m nearest neighbors. This is achieved by specifying carefully constructed small subsets of the conditioning sets. They noticed that if the spatio-temporal covariance function satisfies *natural monotonicity*, not all points in the conditioning set are actually eligible for a neighborhood relationship, but the search can be restricted to a set of $\approx 4m$ points determined in terms of their temporal and spatial distance. The logic behind this scheme starts from the two most extreme scenarios: the case in which temporal covariance is irrelevant, hence the m nearest neighbors would be the m closest in space, and viceversa. As the importance of the irrelevant dimension increases, one point of the previous set will be excluded to make space to a new one which is close on the other dimension. If all points are ordered along two axis representing the corresponding spatial and temporal lags, it is indisputable how all mid-ways possibilities must be composed of points close in both space and time. Thus, points that are too far in both dimensions can be painlessly excluded. Let us assume the interest lies in finding all eligible neighbors of point a, and in a random scan the point b is considered. The rule is that if there are at least m other points closer **both in space and time** to a than b, then b is clearly ineligible as neighbor (will not ever be a neighbor of a for any choice of θ). The visualization of such eligible set is straightforward on a regular grid such as the one introduced for the *simple neighbor set*: a point $(\boldsymbol{u}, \tau)$ is eligible for neighborhood with $(\boldsymbol{s}_i, t_j)$ if the smallest rectangle of points preceding $(\boldsymbol{s}_i, t_j)$ in which it is contained has size $< m$. An example is provided in Figure 2.12. Using this strategy, the eligible sets can be computed only once before the MCMC implementation and the cost of the neighbor search at each iteration reduces to $\mathcal{O}(4m(n+l))$. For further details on this neighbor search algorithm the reader is referred to the Supplementary Material of Datta et al. (2016b).

Whilst important for hasting convergence and improving the quality of the inferential result, we must recall that the NNGP and DNNGP frameworks are valid for any arbitrary specification of neighbor sets. Potentially, these may even be chosen at random. As proved in the seminal papers, these methods proved to be robust with respect to different neighborhood constructions, and the improvement guaranteed by the adaptive strategy is hardly noticeable in standard settings (especially when m grows). Nothing excludes that in particularly imbalanced contexts (the most of the dependence is driven by time or space), the improvement would be more evident and thus a careful choice of the neighboring set is always strongly advised.

These last considerations conclude the last section of Chapter 2. It provided an extensive introduction to the NNGP methodology, exalting its potential and flexibility. As promised, it does offer a highly scalable model (rather than an approximation) for large to massive spatial data-sets that has also proved to be robust with respect to its *tuning* parameters (i.e., ordering of locations, neighbor sets' size etc.). Implementation guidelines in the context of linear and Gaussian geostatistical models have been described, with references to other alternatives. A tentative extension to *Generalized linear geostatistical models* is already present in Datta et al. (2016a), but it is a setting undergoing further developments. Other alternatives based on the same logic are being developed (e.g., Peruzzi et al. (2020a)), and ever-increasing numbers of applications from different fields are making successful use of it (Abdalla et al., 2018; Finley et al., 2017b; Shirota et al., 2019; Grenier and Sansó, 2019; DiFranzo et al., 2020; Segura et al., 2020; Peruzzi et al., 2020b). All things considered, this is a reasonably innovative methodology only recently invented, and it has already drawn an exceptional lot of attention.

Chapter 3

Combining NNGP and Spline regression for modeling physical activity level in a large scale population study

Promoting a healthy lifestyle continues to stoke substantial research activities in public health. The "Physical Activity Guidelines for Americans" (2nd edition) suggests that most individuals, depending on age and body composition, receive 150-300 minutes of moderate to vigorous physical activity (MVPA) weekly (Piercy et al., 2018). In general, the scientific community agrees that regular physical activity can have immediate and long-term health benefits (Reiner et al., 2013; Bull et al., 2020). Despite these well-known benefits, most Americans fail to meet recommended requirements (Piercy et al., 2018). Specifically, only 1 in 5 high-school adolescents and 1 in 4 adults meet recommended levels of physical activity. Given the well-established relationships between lack of physical activity and several leading chronic conditions such as heart disease, type 2 diabetes, and cancer as well as many physical and mental health benefits, an urgent need exists to improve monitoring of physical activity and to establish public health programs that promote more physical activity [1].

Data science technologies for monitoring various physical parameters (e.g. step rates, blood pressure, heartbeat, activity counts, etc.) and promoting physical activity continue to emerge and develop. Many devices also include Global Positioning System (GPS) sensors that allow for recorded parameters to be paired with location tracking. Such data are gathered directly with wearable sensors or indirectly through smart-phone mobile applications, and record repeated measures at a predetermined (potentially very high) frequency. Collected data are easily downloaded and promptly analyzed in order to get insights about their pattern and structure. In general, analyzing such data is sought for several reasons: (i) assessing health effects of different physical activity intensities (Pate et al., 2008; Smuck et al., 2014; Farooq et al., 2020); (ii) improving classification accuracy of activity intensity (e.g. sedentary, light, moderate, vigorous; Degroote et al. (2020); Sagelv et al. (2020)); (iii) assessing the effectiveness of interventions and health promotion techniques (Troped et al., 2010; Dunton et al., 2014; Hartman et al., 2018).

Physical Activity (PA) tracking is especially attractive for research as it allows

[1] More details at `https://www.cdc.gov/chronicdisease/resources/publications/factsheets/physical-activity.htm`

for a better understanding of what behavioral and environmental factors influence population and individual health and, hence, aid in public health recommendation and policy. Research however is often hindered by the challenge of employing a valid, reliable measure for energy expenditure (EE) that also adequately satisfies the research question or design. The current gold standard is known as *Doubly Labeled Water* (DLW), which uses the total EE derived from daily biological measurement as a proxy for the average level of physical activity. While being objective and accurate, and promising as a tool to serve as criterion measure for other instruments (Westerterp, 2009), the DLW remains impractically expensive for large-scale studies. Consequently, physical activity assessment is still often based on (1) self-report questionnaires (Schuch et al., 2017), (2) self-report activity logs or diaries (Bedard et al., 2020), and (3) direct measurement (Maher et al., 2019). While the (1) and (2) are relatively cheap with the second one implying a significant burden on the respondent, (3) is highly demanding in time and energy. More importantly, all three lack an objective measure of EE (Hidding et al., 2018).

New tools for monitoring instant physiological markers related to objective measures of physical activity have been recently developed, enabling comparability with the tools usually adopted for measuring EE in controlled studies such as the VO2 (Miller et al., 2010). Accelerometers, in particular, motion sensors that allow to quantify gross motor activity over time and are increasingly conspicuous because of their affordability, increasing accuracy, and their availability in common devices such as smart-phones, smart-watches, and other wearable devices. More generally, such devices can be administered to individuals in their normal surroundings and allow for individual-specific *Actigraphy*[2] analysis. That's why, in this context, these are usually referred as *Actigraph Units* (Plasqui and Westerterp (2007); Westerterp (2009) Sikka et al. (2019)). These tools have been widely deployed in controlled studies for deriving the EE corresponding to pre-determined physical activities (Abel et al., 2011), but their characteristics makes them superior to surveys for the objective assessment of physical activity also in population studies (Strath et al., 2013; Tudor-Locke et al., 2010) or large studies (Troiano et al., 2014; Berkemeyer et al., 2016; Doherty et al., 2017; Pearce et al., 2020). They however have several well-known limitations. For example, accelerometers usually underestimate energy expenditure of activities such as cycling, swimming, weight lifting, and many household chores. Physical activity estimates can also vary depending on subjective decisions in data reduction such as the choice of cut-points for intensity levels, the minimum number of valid days, the minimum number of valid hours per day, and the definition of non-wear time (Rachele et al., 2012). To mitigate the effects of some of these limitations, it is desirable to use the data at high (or fine) resolutions instead of descriptive analysis or regression modeling on the average level of activity aggregated at the individual level (Mitchell et al., 2017). For an extensive review of physical activity assessment methods see Sylvia et al. (2014) and references within.

We pursue a comprehensive analysis of an original actigraphy data set from the Physical Activity through Sustainable Transport Approaches in Los Angeles (PASTA-LA) study. These data were compiled as part of a longitudinal study conducted by the Fielding School of Public Health of the University of California, Los Angeles (UCLA) that enrolled a cohort of 460 individuals, followed through different tools and devices for approximately 2 years. In this work, we focus on three particular tasks: (i) estimating a subject's physical activity levels along a given trajectory; (ii) identifying trajectories that are more likely to produce higher levels of physical activity for a given subject; and (iii) predicting expected levels of physical activity in

[2]Actigraphy is a non-invasive method of monitoring human rest/activity cycles.

any proposed new trajectory for a given set of health attributes. We propose a fully model-based approach where data points are not aggregated at the individual level, but only at the *epoch* level (set of contiguous instants). Each record is characterized by a time-stamp and (when equipped with a GPS device) by the corresponding geo-referenced location. Hence, it is natural to choose from the rich classes of spatial and spatio-temporal models introduced in Chapter 2. Nevertheless, actigraph data generally present some notable challenges. First, repeated-measures of accelerometry data produce vast amounts of data, even for sample sizes of only a few individuals monitored for short periods of time (e.g. days or weeks). We overcome this limitation by exploiting some inherent data structures and by considering the Nearest Neighbor Gaussian Process (Datta et al., 2016a), or (NNGP), to deal with massive temporal scales (see Section 2.5.1). Second, consecutive recording exhibit dependence along trajectories that must be accounted while predicting PA along arbitrary (unobserved) trajectories (Kestens et al., 2017). We argue against a customary spatiotemporal process over $\mathbb{R}^2$ and disentangle spatial effects from dependence along trajectories.

The balance of this chapter is organized as follows. Section 3.1 introduces the accelerometry data. The model for the temporal correlation is extensively introduced in Section 3.2, while spatial effects are specifically discussed in Section 3.2.4. An extensive simulation study validating our model is proposed in Section 3.2.5. Data analysis, including model assessment and comparisons, are presented in Section 3.3. Finally, the application is concluded with a discussion in Section 3.4. Some promising future developments are finally sketched in Section A.2.

3.1 Data

3.1.1 Data collection

Our data for the current project has been compiled from the **P**hysical **A**ctivity through **S**ustainable **T**ransport **A**pproaches - **L**os **A**ngeles (PASTA-LA) study. The Fielding School of Public Health of UCLA has conducted this study on a cohort of 460 individuals. These individuals were monitored for a period of two years (2017 and 2018) to verify the effect of the introduction of the *Bruin Bike Sharing* system on their physical activity level. Data were collected through different sources: online questionnaires, a smartphone app named *MOVES*, a GPS device (GlobalSat DG-500), and a wearable Actigraph unit (Actigraph GT3X+). While 460 is the sample size for the full study, the GPS and Actigraph devices were deployed only on a nested sample of 163 individuals due to cost considerations. Data collected through the *MOVES* app, whose reliability and comparability across individuals (different devices) must still be verified and discussed, are not considered in the domain of this paper. Here, we rather focus on the values recorded by the Actigraph and GPS devices, worn for only two one-week periods (one in 2017 and one in 2018), because of their constant recording frequency and the wider literature concerning their use and study.

Study protocol for safeguarding participant information received institutional review board (IRB) approval from the UCLA Human Research Protection Program. A final composite directory housed all data, separated by participant. This data was stored on a secure computer, and a redacted version was created for purposes of data sharing and research collaboration. We do not pursue all of the aims of the PASTA-LA study, but we build and test the framework proposed in Section 3.2 for modeling of high-frequency sampled data related to different individuals.

Questionnaires The online questionnaire was supposed to be repeated six times on the whole population, three times before the intervention (Bruin Bike Share Launch) and three times after: this included two baseline surveys and four follow-up surveys. Each survey included a maximum of 184 variable responses per participant pertaining to the user's demographics and transportation habits. Not all participants completed all questionnaires, and the survey available on the largest part of participants is the *First baseline questionnaire*, which contains mostly demographic information such as sex, age, BMI[3], ethnicity and other socioeconomic factors. For this reason, in the domain of this paper, only data resulting from this survey have been included. A user ID has been assigned to each survey response data and a redacted master key was generated using all ID types for joining with other study data.

Actigraph The Actigraph units were provided to a nested sample of 163 individuals for two one-week periods, before and after the Bruin Bike Share launch. As described in Trost (2007), an Actigraph unit is an accelerometer measuring $38 \times 37 \times 18$ mm and weighing 27 g. These devices can be worn in numerous places on the body, including wrist, hip, and thigh and measure the directional acceleration at a pre-specified time frequency (generally 10 Hz to 30 Hz). The most recent models, such as the Actigraph GT3X+ used for the PASTA survey, can detect movements in up to three orthogonal planes (anteroposterior, mediolateral, and vertical). Data are stored in an internal memory and can be downloaded to other hardware for performing analysis through a proprietary software. During download, the proprietary software converts the raw acceleration information to activity counts, step counts, caloric expenditure, and activity levels, aggregated at the level of sample epochs that can be specified by the user. Unfortunately, the software is proprietary and precludes recovering the raw data once they have been processed in the download phase. In our case, the 163 participants were asked to keep the Actigraph unit (on the wrist of their choosing) on them at all times other than during bathing and sleeping (awake time approximately from 7 am to 11 pm). Therefore, observations recorded outside this daily time-window have been deleted. Additionally, the devices were not supposed to be a nuisance for the participants, so they were allowed to take it off if needed. Troiano et al. (2014), showed that such a protocol naturally results in huge amounts of missing data, not random but biased toward an increased general level of physical activity (i.e. people who kept the accelerometer on during these times are likely to be the ones who would be performing physical activity). Here, we focus upon estimating the physical activity level during the active time.

When participants arrived at the research offices to drop off devices, some described issues of efficacy in the ability to keep the device on or charged. Indeed, while the actigraphs were supposed to hold a charge long enough to last the whole week, this was not always the case (possibly due to external conditions affecting the battery life or variations in manufacturing). Hence, we anticipated a large amount of missing data.

During download, the data were aggregated in sample epochs (proprietary unit of time of Actigraph) that span 10 seconds and include the activity counts for the three axles and step counts as these are computed by the proprietary software. Time-stamps of each final measurement (hour, minute, and second) have been obtained as the mid-point between the beginning and the end of the epoch.

Furthermore, the Actigraph GT3X+ automatically recorded a measure of light exposure (*lux*) and inclinometer values on how many of the 10 seconds epoch have

[3]binned according to the World Health organization (WHO) guidance (`https://apps.who.int/bmi/index.jsp?introPage=intro_3.html`)

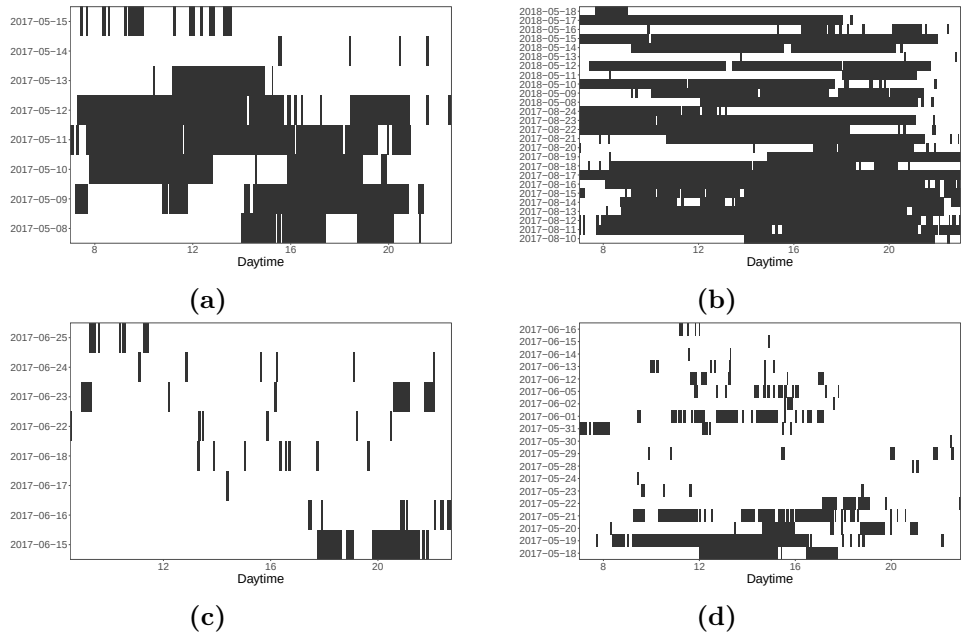


Figure 3.1. Missing data pattern after inclinometer-based dropping for the Actigraph data during daytime for 4 individuals: blank spots are missing data, black spots are observed values

been spent by the participant lying down (*inclinometer.lying*), sitting (*inclinometer.sitting*), standing (*inclinometer.standing*) or without wearing the accelerometer (*inclinometer.off*). These variables were not of primary interest to the PASTA-LA study and were not directly addressed in their data collection protocol for quality assurance. For example, there was no guidance about keeping the device open to the light (e.g. not covered by clothing). Regarding the inclinometer[4], it has reported error rates up to 30% when worn in the most accurate location, and likely less accurate when worn on the wrist (Peterson et al., 2015). We sought to exploit convergence of accelerometry and inclinometer data to derive periods of inactivity in the study data. We checked that large values of *inclinometer.off* corresponded to low (~ 0) values of activity in all the possible endpoints. Therefore, observations with *inclinometer.off* larger than $5s$ (i.e., the accelerometer was inactive for more than half of the epoch) have been dropped as likely to actually represent *non-wear* time. Despite this first pre-processing step, the dataset still presented many records of activity values exactly equal to 0 ($\approx 30\%$ of all the records). A more detailed analysis revealed how such values were mostly associated to participants lying or sitting, thence practically inactive. In a comprehensive analysis, this huge inflation of 0s should be taken properly into account[5]. However, this is out of the scope of the project and will be considered in future work.
Here, as specified above, we sought to analyze activity levels only throughout the "active time"–when the Actigraph device records the individual as being physically active— and excluded all the inactive times (i.e. $MAG = 0$). This pre-processing resulted in $\approx 6.4 \times 10^6$ scattered observations, exhibiting missingness patterns such

[4]see `https://actigraphcorp.com/research-database/validity-of-the-actigraph-inclinometer-algorithm-for-d` for details

[5]For instance, standard approaches include mixtures of distributions or hidden Markov models that allow to model jointly state and activity level (Cappé et al., 2006).

as in Figure 3.1.

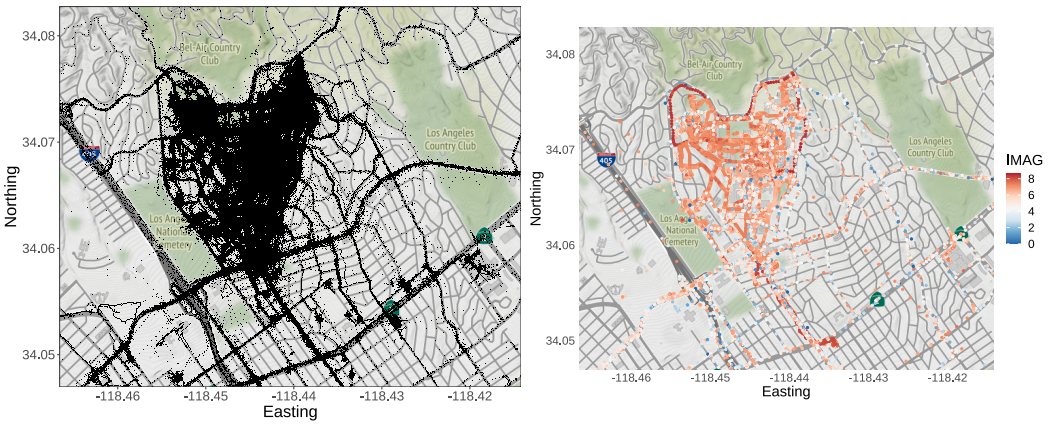
GPS The *Global Positioning System* (GPS) is a satellite-based radio-navigation system. It does not require the user to transmit any data, and it operates independently of any telephonic or internet reception. Any GPS unit can be set to record and store the geo-referenced location at a pre-specified time frequency so that they could be downloaded and subsequently analyzed in a second moment. It is worth mentioning that obstacles such as mountains and buildings can block the relatively weak GPS signals and cause the device to not record or record meaningless locations by time to time (Wing et al., 2005).

The *GlobalSat DG-500* was provided to the same nested sample of PASTA-LA study participants together with the Actigraph unit (163 individuals), with the specific task of taking it with them at any time during the two weeks (e.g. in pants pocket or in the backpack). The GPS unit recorded the location every 15 seconds and the downloaded data contained: date and time of localization (time-stamp), latitude, longitude, altitude in meters (mostly missing), and speed in kilometers per hour (computed as distance over time through linear interpolation). Participants were eligible for recruitment if they worked or lived in the Westwood and UCLA areas. Consequently, Westwood is the zone presenting the highest and most uniform density of observations. Hence, in order avoid a geographical imbalance that could bias and invalidate the model estimates, we decided to restrict the spatial analysis only to GPS records falling inside its boundaries. This allowed, in one shot, to exclude some of the clearly unreasonable GPS values resulting from connection problems or participants that would forget to turn off the tracking during long-range travels (e.g. on a flight). The remaining clear errors (e.g. jumps of > 10 mile in the span of 15 seconds) were detected by verifying coherency rules and dropped before the analysis.

Joining GPS and accelerometer data were all assigned a participant ID, coherent with the questionnaires' masterkey to facilitate joining across all ID types (including email) while redacting and encrypting user data. The first baseline questionnaire, Actigraph and GPS were all available for a group of 134 individuals. Thus, from now on we will refer to this specific group of units. We then build two different sets of data.

- The first dataset, D_1, compres $N \simeq 5 \times 10^6$ records is obtained by joining the first baseline questionnaire with Actigraph data and includes the outcome of interest and all the individual predictors at each timestamp, but no spatial information.

- The second dataset, D_2, contains $N \simeq 5 \times 10^5$ measurements (see Figure 3.2). It is obtained by joining D_1 with GPS data, and the reduction of the sample size is due to the joining process. Indeed, data collected by the Actigraph and the GPS unit are not temporally aligned. We linearly interpolated the GPS locations on the same temporal grid as the Actigraph data, but retained only those interpolated values where the two subsequent GPS measurements were less than 30 seconds away. This decision relied on the assumption that the across such a little time-span, individual's trajectories can be well-approximated by a piece-wise linear curves.



(a) **(b)**

Figure 3.2. (a): Observed locations over the Westwood area. (b): Observed physical activity levels over the Westwood area on a subset of 10 individuals.

3.1.2 Measure of the physical activity

The Actigraph GTX3+ provides three possible endpoints: step counts, z-axis activity counts, y-axis activity counts and x-axis activity counts. Information about step counts and z-axis activity counts is redundant since the former is derived directly from the latter. While z-axis measurements explain most of body movement variation, recent literature has proved how the best prediction of energy expenditure can be achieved by vector summation of the absolute values of all the three axes of orthogonal force measurement. For this reason, our primary endpoint of analysis is the MAG (Magnitude of Acceleration) defined as:

$$MAG_{kt} = \sqrt{x_{kt}^2 + y_{kt}^2 + z_{kt}^2}, \qquad k = 1, \dots, K, \quad t = t_{k1}, \dots, t_{kT}\ , \tag{3.1.1}$$

where t_{kj} is the j-th time point for the k-th individual; and x, y and z are the activity counts of the three axes (Ott et al., 2000). There are substantial investigations into the statistical relationships between accelerometer measurements and *energy expenditure* measures (EE) (Freedson et al., 2012; Taraldsen et al., 2012) and, in particular, the *Metabolic Equivalent of Task* (METs) which is currently the standard measure of *rate of activity intensity* (Crouter et al., 2006; Hall et al., 2013; Lyden et al., 2014). Migueles et al. (2017) offers an extensive review of proposed accelerometer measurement cut-points and transformation into physical activity metrics. More specifically, Sasaki et al. (2011), Santos-Lozano et al. (2013) and Kamada et al. (2016) investigated axis counts and vector magnitudes resulting from the GT3X+ accelerometer in both controlled and free-living environments, while Aguilar-Farias et al. (2019) investigated the accuracy of relationships between MAGs and METs in comparison to those based on the vertical axis counts only and validated the results with the EE and METs as quantified by a portable calorimeter.

The MAG-to-MET relationship expounded in Sasaki et al. (2011) is expressed as a function of the MAG per minute, which we rescale to our 10 seconds aggregated counts as:

$$\text{MET}_{kt} = (0.000863 \times 6) \cdot MAG_{kt} + 0.668876. \tag{3.1.2}$$

Activity intensity	MET range	MAG
Sedentary or light	[0, 3)	[0, 493)
Moderate	[3, 6)	[493, 1029)
Hard	[6, 9)	[1029, 1608)
Very hard	[9, ∞)	[1608, ∞)

Table 3.1. MAG activity count cut-points for different PA intensity levels

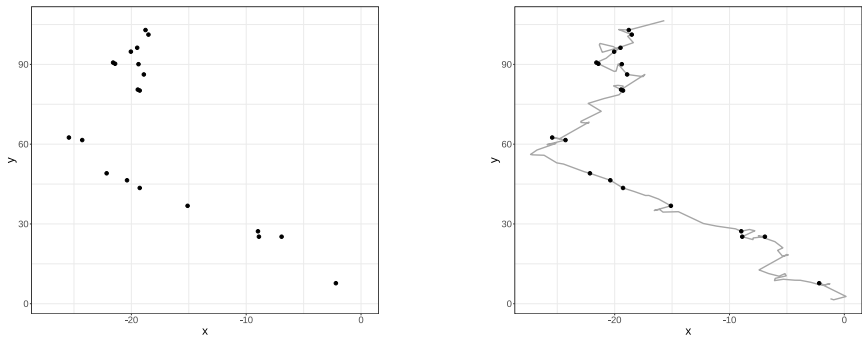
The same rescaling is applied to the corresponding cut points for different PA intensity levels (Table 3.1). Based upon the aforementioned literature, inference for the MAG is transformed into METs through (3.1.2) to interpret results from a physical activity perspective. Nevertheless, equations directly relating accelerometer measurements with physical activity metrics in free-living studies must be interpreted with caution. Relationships between MAG and MET have been posited in controlled studies and validated while patients are performing specific tasks (i.e. walking on a treadmill, gardening etc.). The relationship between the recorded movement (acceleration) and the corresponding energy expenditure, can vary significantly across different tasks affecting the reliability of acceleration-based energy expenditure metrics (Lyden et al., 2011; Freedson et al., 2012; Montoye et al., 2018).

3.2 The model

We devise a class of models to analyze high-frequency temporal observations belonging to different individuals. The proposed model takes into account the dependence structure among different realizations belonging to the same individual. Computations emanating from the massive dataset are addressed exploiting the properties of the uni-dimensional (temporal) setting and efficient approximations for sparse matrix multiplication and inversion.

The outputs corresponding to the K individuals are referenced with respect to the time at which they are recorded and the position in the trajectory. Let us point out that an individual can visit the same location numerous times in his/her trajectory, and the revisits may occur at distant time points. This suggests that the proximity of two spatial locations in a trajectory needs not to result in strongly dependent measurements. Hence, it seems much more reasonable to model dependence between MAGs through a temporal process. In fact, such temporal processes can be motivated by the position vectors defining the trajectories as we describe below.

Let $Z_k(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ be a spatial process corresponding to individual k. The domain of $Z_k(\cdot)$ is restricted to the trajectories $\gamma_k(t) = (\gamma_k^x(t), \gamma_k^y(t))$, where $k = 1, \ldots, K$ and $t \in \mathbb{R}^+$, which defines the movements of the k-th individual along time. As shown in Figure 3.3, the process actually belongs to a one-dimensional space, for which we need to define a proper distance measure $d(t_{ki}, t_{kj}) = \|\gamma_k(t_{kj}) - \gamma_k(t_{ki})\|$, where t_{ki} denotes the i-th recorded time point of individual k. A similar problem has been addressed in Abdalla et al. (2018), where the geographic distance along a coast has been replaced by a piece-wise linear approximation over a coarse grid. Here, we instead approximate distances along trajectories with the elapsed time between the two points $d(t_{ki}, t_{kj}) = |t_{kj} - t_{ki}|$. This results in an exact approximation (up to a proportionality constant) of the spatial distance as long as individual k is moving at constant speed. On the other hand, given two equally distant locations, the faster an individual is moving from one to the other and the shorter the time elapsed, resulting in higher correlation between the two measurements. This is still reasonable in our



(a) **(b)**

Figure 3.3. Example of observed points (a) and trajectory (b): black dots are realizations, grey line is domain of the process.

experimental setting, where we can expect PA variations to be affected more by the temporal than the spatial scale.

In practice, we are assuming the elapsed separation across time will reflect dependence better than the straight spatial distance. Hence, we may decide to parametrize the process just in terms of the time t, considering the composition

$$Z_k \circ \gamma_k(\cdot) \equiv Y_k(\cdot) : \mathbb{R}^+ \to \mathbb{R},$$

which, by construction, is a valid stochastic process (Abdalla et al., 2018).

This will form the edifice of the model in Section 3.2.1, where we are modeling the dependence by solely considering stochastic evolution through time. How should spatial information be introduced in the model? While it may be tempting to work with a spatiotemporal process, dependence introduced by such processes may not be appropriate.

Two individuals at the same spatial coordinate experience the same spatial effect but different temporal effects because their physical activities are a function of their individual trajectory evolution. An added complication is that trajectories intersect and overlap and, in practice, can have multiple observations at the same location. Even more flexible spatiotemporal covariance kernels (e.g., nonseparable or nonstationarity) will struggle to recognize the above features. Hence, the spatial effect must act independently from the intrinsic temporal individual process. Therefore, in order to not over-complicate the dependence structure of the observed process, we decide to unload the spatial effect on the mean term and we do so by *spline regression* in Section 3.2.4.

3.2.1 Temporal model

Let $\mathcal{T} = \cup_{k=1}^{K} \mathcal{T}_k$ where $\mathcal{T}_k = \{t_{ki}\}_{i=1}^{T_k}$ and $t_{ki} \in \mathbb{R}^+$ be the set of the $n = \sum_{k=1}^{K} T_K$ observed time points. In a first instance, we may suppose that $\boldsymbol{Y}(\mathcal{T})$ is the finite realization of a K-variate process $\boldsymbol{Y}(\cdot)$ over $\mathbb{R}^+$:

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t, \boldsymbol{\gamma}(t))\boldsymbol{\beta} + \boldsymbol{w}(t) + \boldsymbol{\varepsilon}(t), \quad t \in \mathbb{R}^+, \tag{3.2.1}$$

where $\boldsymbol{Y}(t) = (Y_1(t), Y_2(t), \ldots, Y_K(t))^\top$ is a $K \times 1$ vector of measurements at time t on the K individuals, $\boldsymbol{X}(t, \boldsymbol{\gamma}(t))$ is a $p \times K$ matrix, each row being the values of a covariate for the K individuals, $\boldsymbol{w}(t) = (w_1(t), w_2(t), \ldots, w_K(t))^\top$ is a $K \times 1$ vector

comprising a temporal process for each individual, and $\boldsymbol{\varepsilon}(t) \sim \mathcal{N}_K(0, \tau^2 \cdot \boldsymbol{I}_K),\ \tau^2 \in \mathbb{R}^+$ is a white noise process for measurement error. $\boldsymbol{X}(t, \boldsymbol{\gamma}(t))\boldsymbol{\beta}$ is also known as the *large scale variation*, while $\boldsymbol{w}(\cdot)$ as the *small scale variation*.

Assuming independence among all the components of the random temporal process $\boldsymbol{w}(t)$, we can use Gaussian Processes (GP) for modeling the stochastic behaviour of each single component $w_k(\cdot)$:

$$w_k(t) \sim \mathcal{GP}\left(0, c_{\theta_k}(\cdot,\cdot)\right), \quad \forall k = 1, \ldots, K, \tag{3.2.2}$$

where $c_{\theta_k}(\cdot,\cdot)$ is a common covariance function depending on a set of parameters $\theta_k \in \Theta$.

Let y_{ki} and $\boldsymbol{x}_{ki}$ denote the outcome and the covariates of individual k at the generic time point t_{ki}, respectively, so:

$$\{(y_{ki}, \boldsymbol{x}_{ki}) :\ k = 1, \ldots, K,\ i = 1, \ldots T_k\}$$

is the observed data. Let $\boldsymbol{y}_k$ be $T_k \times 1$ vectors comprising all measurements on patient k, respectively. We will then refer to the joint $n \times 1$ vector of the outcomes and $n \times p$ matrix of the predictors as:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_{1\cdot}^\top & \boldsymbol{y}_{2\cdot}^\top & \cdots & \boldsymbol{y}_{K\cdot}^\top \end{bmatrix}^\top, \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{1\cdot}^\top & \boldsymbol{X}_{2\cdot}^\top & \vdots & \boldsymbol{X}_{K\cdot}^\top \end{bmatrix}^\top,$$

where $\boldsymbol{X}_{k\cdot}$ is the $T_k \times p$ matrix of predictors corresponding to $\boldsymbol{y}_k$ and values are first ordered by individual and then by time. Let us then denote with $\{\boldsymbol{w}_k\}_{k=1}^K$ the $T_k \times 1$ vectors comprising all the random effects on patient k, forming the $n \times 1$ vector $\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_{1\cdot}^\top & \boldsymbol{w}_{2\cdot}^\top & \cdots & \boldsymbol{w}_{K\cdot}^\top \end{bmatrix}^\top$.

The modeling in (3.2.1) implies the following distribution on our finite realization $(\boldsymbol{y}, \boldsymbol{X})$:

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w} + \boldsymbol{\varepsilon}, \\ &\boldsymbol{w} \sim \mathcal{N}_n\left(0, \boldsymbol{C}_\theta\right), \\ &\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0, \tau^2 \cdot \boldsymbol{I}_n\right), \end{aligned} \tag{3.2.3}$$

where the $n \times n$ covariance matrix $\boldsymbol{C}_\theta$ is such that the i-th row and j-th column entrance is:

$$[\boldsymbol{C}_\theta]_{ij} = c_\theta^*(t_{kp}, t_{lq}) = \begin{cases} c_{\theta_k}(t_{kp}, t_{lq}) & k = l \\ 0 & k \neq l \end{cases}.$$

Therefore, $\boldsymbol{C}_{\boldsymbol{\theta}} = \operatorname{diag}\left(\boldsymbol{C}_{\boldsymbol{\theta}_1,1}, \boldsymbol{C}_{\boldsymbol{\theta}_2,2}, \ldots, \boldsymbol{C}_{\boldsymbol{\theta}_K,K}\right)$ is $n \times n$ block-diagonal with $\boldsymbol{C}_{\boldsymbol{\theta}_k,k} = [c_{\boldsymbol{\theta}}(t_{ki}, t_{kj})]$ as the $T_k \times T_k$ temporal covariance matrix corresponding to individual k. Each individual is allowed its own covariance parameters, θ_k, and $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$ is the collection of all the covariance kernel parameters. In what follows, we will refer to $\boldsymbol{C}_\theta$ as $\boldsymbol{C}$.

Full inference from (3.2.3) is straightforwardly set up by ascribing to the set of parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ a suitable prior $\pi(\boldsymbol{\beta}, \boldsymbol{\theta})$ and deriving the posterior distribution as:

$$\pi(\boldsymbol{\beta}, \theta \,|\, \boldsymbol{y}) \propto \mathcal{N}_n\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \tau^2 \cdot \boldsymbol{I}_n\right) \times \mathcal{N}_n\left(\boldsymbol{w} \,|\, 0, \boldsymbol{C}\right) \times \pi(\boldsymbol{\beta}, \theta).$$

While this solution looks complete and elegant, its application is usually limited by the size of the observed sample. Indeed, applying it involves the computation of the density:

$$\mathcal{N}_n\left(\boldsymbol{w} \,|\, 0, \boldsymbol{C}\right) = \frac{1}{\sqrt{\det(\boldsymbol{C})}} \exp\left\{\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{C}^{-1} \boldsymbol{w}\right\}, \tag{3.2.4}$$

that encompasses the evaluation of $\det(\boldsymbol{C})$ and of $\boldsymbol{C}^{-1}$, which require $\mathcal{O}(n^2)$ storage space and $\mathcal{O}(n^3)$ floating point operations (flops). This makes the estimation of such a model unfeasible for the majority of contemporary applications, and in particular for the data in consideration that include more than 10^6 observations.

Our set of data has characteristics that may ease the troublesome computational burden. Indeed, the covariance between observations belonging to different units is set to 0, and we can exploit the block-diagonal structure of the resulting covariance matrix. The inverse can be easily computed by inverting independently each sub-matrix on the diagonal and the determinant can be obtained by multiplying the determinants of the all the sub-matrices:

$$\begin{aligned} \boldsymbol{C}^{-1} &= \oplus_{k=1}^{K} \boldsymbol{C}_k^{-1} \\ \det(\boldsymbol{C}) &= \prod_{k=1}^{K} \det(\boldsymbol{C}_k) \end{aligned} \tag{3.2.5}$$

This reduces the flop count from $\mathcal{O}(n^3) = \mathcal{O}((\sum_{k=1}^{K} T_k)^3)$ to $\mathcal{O}(K \sum_{k=1}^{K} (T_k)^3)$, with a significant saving of calculations, especially when the T_k's are reasonably small ($< 10^4$). Furthermore, each $\boldsymbol{C}_k$ can be computed in parallel rendering further scalability to the algorithm.[6]

Nevertheless, considering all the data included in the Actigraph dataset (see Section 3.1), some of the individuals present more than 10^5 observations. This means that in order to get full inference on the parameters in a reasonable amount of time, we need to adopt some *ad-hoc* strategy in order to ease the computations also over each single covariance submatrix $\{\boldsymbol{C}_k\}_{k=1}^{K}$.

As discussed in Section 2.5, different strategies have been proposed to overcome the dimensionality issues in the spatial and spatio-temporal literature, e.g.: low-rank approximations (Banerjee et al., 2008; Finley et al., 2009), covariance tapering (Furrer et al., 2010), *Gaussian Markov Random Field* (GMRF) (Rue and Held, 2005), *Nearest Neighbor Gaussian processes* (NNGP) (Datta et al., 2016a,b) and other Vecchia-based approximations (Katzfuss et al., 2020; Katzfuss and Guinness, 2021; Peruzzi et al., 2020a) etc.

We will pursue a DAG-based approximation due to Vecchia (Vecchia, 1988) in the spirit of the NNGP.

3.2.2 Independent DAG models over individuals

Datta et al. (2016a) introduced the NNGP approximation in the general case of observations scattered on a set $\mathcal{D} \subset \mathbb{R}^d$. Here, we give a brief summary of its principles in the context of the process $w_k(\cdot)$ referred to the k-th individual, whose finite realization is $\boldsymbol{w}_k = [w_{ki}]_{i=1}^{T_k}$.

The NNGP stems from the parent GP through the conditioning argument $p(\boldsymbol{w}_k) = p(w_{k1}) \prod_{i=2}^{T_k} p(w_{ki} \,|\, \boldsymbol{w}_{k(1:i-1)})$ and the definition of arbitrary *neighbour sets* $\{N(i)\}_{i=1}^{T_k}$ of fixed size m. Exploiting a variation of the *Vecchia* approximation (Vecchia, 1988), we can define $\widetilde{w}_k(\cdot)$ such that:

$$p(\boldsymbol{w}_k) \approx p(\widetilde{\boldsymbol{w}}_k) = p(w_{k1}) \prod_{i=2}^{T_k} p(w_{ki} \,|\, \boldsymbol{w}_{kN^*(i)}). \tag{3.2.6}$$

[6]Notice that, even if we are assuming the individual latent effects $w_k(\cdot)$ to evolve independently over time, we are still pooling the information coming from all the individuals for the estimation of the common vector of coefficients $\boldsymbol{\beta}$ and (eventually) covariance parameters θ.

where $N^*(i) = N(i) \cap \{1, \dots, i-1\}$ is the set of m neighbours included in the conditioning set. The theory of Directed Acyclic Graphs (DAG) can be used to prove that the resulting multivariate distribution is probabilistically valid and consistent with respect to the parent process, as far as the size of the neighbour sets tends to the full size T_k. While this kind of approximation is valid in general (Lauritzen, 1996; Stein et al., 2004; Murphy, 2012), when dealing with the multivariate Gaussian densities stemming from a parent GP it has the desirable property of preserving Gaussianity. The connection between sparsity and conditional independence follows by writing (3.2.6) as a linear model:

$$\begin{aligned} \boldsymbol{w}_k &= \boldsymbol{A}_k \boldsymbol{w}_k + \boldsymbol{\eta}_k, \\ \boldsymbol{\eta}_k &\sim \mathcal{N}_{T_k}(0, \boldsymbol{D}_k) \end{aligned} \tag{3.2.7}$$

where $\boldsymbol{A}_k$ is a $T_k \times T_k$ strictly lower triangular matrix with all positive elements and $\boldsymbol{D}_k$ is a $T_k \times T_k$ diagonal matrix such that $[\boldsymbol{D}_k]_{ii} = d_{ii} = \text{Var}\left(w_{ki} \,|\, \{w_{kj}, j < i\}\right), i = 1, \dots, T_k$. On the one hand, there is no apparent gain in using the Cholesky decomposition to compute the inverse of $\boldsymbol{C}_k$, since the calculation of $\boldsymbol{A}_k$ and $\boldsymbol{D}_k$ is still of the order of $\mathcal{O}(n^3)$. On the other hand, elements of $\boldsymbol{A}_k$ can be set to 0 in order to obtain a sparse version (i.e. sparse Cholesky) $\widetilde{\boldsymbol{C}}_k^{-1}$ of $\boldsymbol{C}_k^{-1}$[7].

This DAG representation imposes a lower-triangular structure on $\boldsymbol{A}_k$, but additional sparsity is achieved by allowing its (i, j)-th $(i < j)$ entry to be nonzero only for $j \in N^*(i)$. Therefore, each row of $\boldsymbol{A}_k$ has at most m nonzero entries and:

$$\widetilde{\boldsymbol{C}}_k^{-1} = (\boldsymbol{I}_{T_k} - \boldsymbol{A}_k)^\top \boldsymbol{D}_k^{-1} (\boldsymbol{I}_{T_k} - \boldsymbol{A}_k),$$

is sparse, where we recall that $\widetilde{\boldsymbol{C}}_k^{-1}$ is the precision matrix of $\widetilde{\boldsymbol{w}_k}$. This corresponds to the *sparse Cholesky* decomposition of $\boldsymbol{C}_k^{-1}$. The key observation is that the nonzero elements of the i-th row of $\boldsymbol{A}_k$ is the solution $\boldsymbol{a}_k$ of the $m \times m$ linear system $\boldsymbol{C}_{\boldsymbol{\theta},k}[N(i), N(i)]\boldsymbol{a}_k = \boldsymbol{C}_{\boldsymbol{\theta},k}[N(i), i]$, where $[\cdot, \cdot]$ indicates submatrices defined by the given row and column index sets. Obtaining the nonzero elements of $\boldsymbol{A}_k$ and $\boldsymbol{D}_k$ costs $\mathcal{O}(T_k m^3)$ (scales linearly with T_k) instead of $\mathcal{O}(T_k^3)$ as would have been without sparsity. This cheaply delivers the quadratic form $\boldsymbol{w}_k^\top \widetilde{\boldsymbol{C}}_k^{-1} \boldsymbol{w}_k$ in terms of $\boldsymbol{A}_k$ and $\boldsymbol{D}_k$ and the determinant $\det(\widetilde{\boldsymbol{C}}_k) = \prod_{i=1}^{T_k} d_{ii}$ at almost no additional cost. Algorithm 1 shows how it is possible to compute the sparse versions of $\boldsymbol{L} = (\boldsymbol{I} - \boldsymbol{A})^\top$ and $\boldsymbol{R} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{A})$, where $\boldsymbol{A} = \oplus_{k=1}^K \boldsymbol{A}_k$ and $\boldsymbol{D} = \oplus_{k=1}^K \boldsymbol{D}_k$.

Hence, we can approximate the parent Gaussian latent process $\boldsymbol{w}_k$ with its NNGP version $\widetilde{\boldsymbol{w}_k}$ and replace the original density in Eq. (3.2.4) with $\mathcal{N}_{T_k}(\widetilde{\boldsymbol{w}_k} \,|\, \boldsymbol{0}, \widetilde{\boldsymbol{C}_k})$, whose computation requires $\simeq \mathcal{O}(T_k)$ flops.

Although Datta et al. (2016a) demonstrated to have no discernible impact on the final approximation, one of the biggest critical points of the NNGP process is that the results in Eq. (3.2.6) and (3.2.7) depend upon an ordering of the observations. Unlike spatial locations, temporal observations possess a natural order. Indeed, observations along time can be ordered from the least to the most recent $t_{k1} < t_{k2} < \dots < t_{kT_k}$, with the additional property to be arranged according to their mutual distance. More precisely, the neighbour set of each time-point t_{ki} is always composed by its m preceding values, if they exist:

$$N^*(t_{k1}) = \emptyset, \qquad N^*(t_{ki}) = \{t_{k\,\max(i-m,1)}, \dots, t_{k(i-1)}\}, i = 1, \dots, T_k.$$

As a result, the lower triangular resulting matrix $\boldsymbol{A}_k$ is not just sparse but is banded, with a lower bandwidth equal to m. Consequently, also the inverse of the covariance

[7] the structure of $\boldsymbol{A}_k$ suggests this solution immediately

Algorithm 1: Sparsity inducing computation of $\boldsymbol{L} = (\boldsymbol{I}_n - \boldsymbol{A})^\top$, $\boldsymbol{d}$ and $\boldsymbol{R} = \boldsymbol{D}^{-1}(\boldsymbol{I}_n - \boldsymbol{A})$

Input: $\{\boldsymbol{C}_k\}_{k=1}^K$
Output: $\boldsymbol{L}$, $\boldsymbol{R}$, $\boldsymbol{d}$
for k *in* $1 : K$ **do**
  $\boldsymbol{L}_k[1, 1] = 1$
  $\boldsymbol{d}_k[1] = \boldsymbol{C}_k[1, 1]$
  $\boldsymbol{R}_k[1, 1] = 1/\boldsymbol{d}_k[1]$
  for i *in* $1 : (T_k - 1)$ **do**
    $\boldsymbol{L}_k[i+1, i+1] = 1$
    $\boldsymbol{L}_k[i+1, N(i+1)] = -\boldsymbol{C}_k[N(i+1), N(i+1)]^{-1} \cdot \boldsymbol{C}_k[N(i+1), i+1]$
    $\boldsymbol{d}_k[i+1] = C[i+1, i+1] - \boldsymbol{C}_k[i+1, N(i+1)] \cdot \boldsymbol{L}_k[i+1, N(i+1)]^\top$
    $\boldsymbol{R}_k[i+1, i+1] = 1/\boldsymbol{d}_k[i+1]$
    $\boldsymbol{R}_k[N(i+1), i+1] = \boldsymbol{L}_k[i+1, N(i+1)]^\top / \boldsymbol{d}_k[i+1]$
  $\boldsymbol{L}[(T_{k-1}+1 : T_k), (T_{k-1}+1 : T_k)] = \boldsymbol{L}_k$
  $\boldsymbol{R}[(T_{k-1}+1 : T_k), (T_{k-1}+1 : T_k)] = \boldsymbol{R}_k$
  $\boldsymbol{d}[T_{k-1}+1 : T_k] = \boldsymbol{d}_k$

matrix $\widetilde{\boldsymbol{C}_k}^{-1}$ is banded with lower and upper bandwidth equal to m. This specific sparsity pattern enables even more efficient matrix manipulation routines and can sensibly boost the computations in Eq. (3.2.4) and (3.2.7).

Finally, assigning an independent NNGP prior to the latent random effect $\boldsymbol{w}_k$ of each individual $k = 1, \ldots, K$, we get the following density:

$$\boldsymbol{w} \sim \mathcal{N}_n\left(\boldsymbol{0}, \widetilde{\boldsymbol{C}}\right), \tag{3.2.8}$$

where $\widetilde{\boldsymbol{C}} = \oplus_{k=1}^K \widetilde{\boldsymbol{C}}_k$. Exploiting the block diagonal structure, inverse and determinant can then be obtained by operating independently on the single individuals with a overall computational cost of $\mathcal{O}(\sum_{k=1}^K T_k m^3) = \mathcal{O}(nm^3)$.

3.2.3 Implementation using collapsed models

The Bayesian hierarchical model considered in (3.2.3), both in its classical and NNGP version, allows for the useful interpretation of the latent temporal component. Its estimation can be crucial in many applications, since it can help in the understanding of the phenomenon behaviour over its domain. The standard representation of the NNGP methodology, known as *Sequential NNGP*, can be estimated using a *sequential sampler* that envisions a direct Gibbs sampling with random walk Metropolis steps. It exploits the full conditional distributions in closed form for $\{\boldsymbol{\beta}, \boldsymbol{w}\}$ and also for τ^2 with an $\mathcal{IG}(a_\tau, b_\tau)$ prior. However, this convenient representation is often nullified in practice by strong autocorrelation and poor mixing of the chains (Liu et al., 1994).

Nevertheless, the flexibility of the Bayesian approach allows for the definition of alternative valid estimation procedures for both the vector of regression coefficients $\boldsymbol{\beta}$ and covariance parameters θ. Several specific samplers for spatial DAG-based models have been devised, explored and compared in Finley et al. (2019), and *collapsed samplers* (marginalized over the latent component $\widetilde{\boldsymbol{w}}$) have been seen to considerably improve convergence. In particular, the author compared three alternatives to the original sequential sampler that try to improve on its performances through the

exploitation of high-performance computing libraries to obviate expensive numerical linear algebra. These have been named as the *Collapsed NNGP*, the *NNGP for the response* and the *Conjugate NNGP*. As a matter of fact, the *Collapsed NNGP* is the only valid fully Bayesian alternative to the original sequential version. Indeed, it allows for the full recovery of the latent process while offering improved chain convergence. The other alternatives are supposed to outperform the first one in terms of computational speed, but do not represent an appropriate choice if the objective is to provide full inference on the latent component. In the sequel, we describe the implementation of the collapsed NNGP in the specific context of temporal processes. In particular, we describe some computational shortcuts linked to convenient patterns arising from the temporal structure.

Starting from the two-stage hierarchical specification of the model in Eq. (3.2.3), the *collapsed model* is obtained by integrating out the latent process $w(\cdot)$, thereby "collapsing" the parameter space to a much smaller domain without $\boldsymbol{w}$. The resulting complete likelihood is:

$$\mathcal{L}\left(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \theta\right) = \mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta},\, \widetilde{\boldsymbol{\Lambda}}),$$

where $\widetilde{\boldsymbol{\Lambda}} = \widetilde{\boldsymbol{C}} + \tau^2 \cdot \boldsymbol{I}_n$. Hence, instead of (3.2.3), we sample from:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2 \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\theta}, \tau^2) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \widetilde{\boldsymbol{C}}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I}_n). \tag{3.2.9}$$

In this framework, it is common practice to assign a flat or conjugate prior $\mathcal{N}_p\left(\boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta\right)$ to the coefficient vector $\boldsymbol{\beta}$ and an inverse-gamma prior $\mathcal{IG}(a, b)$ to the variance of the error term τ^2. On the other hand, the prior for θ varies according to the form of the covariance function. Without loss of generality, we will denote the prior of θ as $\pi(\theta)$. A Gibbs-sampler can be used to update $\boldsymbol{\beta}$, but a Metropolis step is needed in order to update θ and τ. The latter requires the computation of the likelihood in Eq. (3.2.3) which, as in the full model, cannot skip the computation of the inverse of $\widetilde{\boldsymbol{\Lambda}}$. Unfortunately, $\widetilde{\boldsymbol{\Lambda}}^{-1}$ does not share the same convenient factorization of $\widetilde{\boldsymbol{C}}^{-1}$ and is also not guaranteed to be sparse. Nevertheless, as pointed out in Finley et al. (2019), the *Sherman-Woodbury-Morrison* formula allows the expression:

$$\widetilde{\boldsymbol{\Lambda}}^{-1} = \tau^{-2}\boldsymbol{I} - \tau^{-4}\boldsymbol{\Omega}^{-1}, \quad \text{with} \quad \boldsymbol{\Omega} = \left(\widetilde{\boldsymbol{C}}^{-1} + \tau^{-2}\boldsymbol{I}\right), \tag{3.2.10}$$

where $\boldsymbol{\Omega}$ enjoys the same sparsity of $\boldsymbol{C}^{-1}$ and, from basic linear algebra properties, $\det(\widetilde{\boldsymbol{\Lambda}}) = \tau^{2n}\det(\widetilde{\boldsymbol{C}})\det(\boldsymbol{\Omega})$. The core of the algorithm is therefore to compute $\boldsymbol{\Lambda}^{-1}$ through $\boldsymbol{\Omega}$. In our application, the random effect is assumed to be the realization of K independent NNGP. As discussed in Section 3.2.2, this implies a block-diagonal structure for $\widetilde{\boldsymbol{C}}$ that can be shown to be shared also by $\boldsymbol{\Omega}$ (see Eq. (3.2.10)). Each block $\boldsymbol{\Omega}_k$ of $\boldsymbol{\Omega}$ can be computed independently for each individual and, by the same properties used in Eq. (3.2.5), the same holds for its inverse and its determinant. This means that the body of the algorithm will consist of a loop over all the individuals, which allows for straightforward parallelization. Unlike in spatial DAGs (Datta et al., 2016a; Finley et al., 2019), we do not need fill-reducing permutation methods since neighbors sets for temporal processes consist of contiguous observations and $\{\boldsymbol{\Omega}_k\}_{k=1}^K$ are banded matrices with no gaps. We devised a Gibbs sampler with Metropolis random walk updates for (3.2.9), where $\boldsymbol{\beta}$ is updated from its full conditional distribution, while $\{\theta, \tau^2\}$ are updated using an adaptive Metropolis step based on Haario et al. (2001). Here, after the first few iterations, a new proposal covariance matrix is regularly computed on the run according to the empirical covariance of the current chain. Subsequently, a mixture of the original and adaptive proposal is used

Algorithm 2: Sampling from the posterior of the collapsed Temporal NNGP

0: Initialization
begin
  for $k = 1, \dots, K$ **do**
    a: Compute $d_{ij}^k = |t_j - t_i|, \quad \forall t_j, t_i \in \mathcal{T}_k$
    b: Find the neighbour sets $\{N_k(i)\}_{i=1}^{T_k}$

1: Metropolis-Hastings update for $\{\theta, \tau^2\}$
$p(\theta, \tau^2|\cdot) \propto p(\theta, \tau^2) \times \frac{1}{\sqrt{\det \widetilde{\boldsymbol{\Lambda}}}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \widetilde{\boldsymbol{\Lambda}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)$
begin
  for $k = 1, \dots, K$ **do**
    a: Compute $\boldsymbol{L}_k = (\boldsymbol{I}_{T_k} - \boldsymbol{A}_k)^\top$, $\boldsymbol{d}_k = \text{diag}(\boldsymbol{D}_k)$ and $\boldsymbol{R}_k = \boldsymbol{D}_k^{-1} (\boldsymbol{I}_{T_k} - \boldsymbol{A}_k)$ using $\boldsymbol{C}_k$ and $\{N_k(i)\}_{i=1}^{T_k}$
    b: Compute $\boldsymbol{\Omega}_k = \boldsymbol{L}_k \cdot \boldsymbol{R}_k + \tau^{-2} \boldsymbol{I}_{T_k}$ exploiting sparsity
    c: Compute $\boldsymbol{r}_k = \boldsymbol{y}_{k\cdot} - \boldsymbol{X}_{k\cdot}\boldsymbol{\beta}$ and $\delta_{\boldsymbol{D}_k} = \prod_{i=1}^{T_k} d_{k,i}$
    d: Compute $\boldsymbol{v}_k = \boldsymbol{\Omega}_k^{-1} \boldsymbol{r}_k$, $\boldsymbol{u}_k = \boldsymbol{\Omega}_k^{-1} \boldsymbol{X}_{k\cdot}$ and $\delta_{\boldsymbol{\Omega}_k} = \det(\boldsymbol{\Omega}_k)$ exploiting the sparse Cholesky decomposition of $\boldsymbol{\Omega}_k$
    e: Collect $\boldsymbol{r}_k$, $\boldsymbol{v}_k$ and $\boldsymbol{u}_k$ into $\boldsymbol{r}$, $\boldsymbol{v}$ and $\boldsymbol{u}$, respectively.
  f: Compute $q_1 = \tau^{2n} \cdot \prod_{k=1}^K \delta_{\boldsymbol{D}_k} \cdot \prod_{k=1}^K \delta_{\boldsymbol{\Omega}_k}$ and $q_2 = \boldsymbol{r}^\top \boldsymbol{r}/\tau^2 - \boldsymbol{r}^\top \boldsymbol{v}/\tau^4$
  g: Get $p(\theta, \tau^2|\cdot) \propto \frac{\exp(-q_2/2)}{\sqrt{q_1}} \cdot (\theta, \tau^2)$

2: Gibbs' sampler update for β
$\boldsymbol{\beta}|\cdot \sim \mathcal{N}_p(\boldsymbol{B}^{-1}\boldsymbol{b}, \boldsymbol{B}^{-1})$, where $\boldsymbol{B} = \boldsymbol{X}^\top \widetilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{X} + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ and $\boldsymbol{b} = \boldsymbol{X}^\top \widetilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{y} + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}}$
begin
  a: Compute $\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ and $= \mathbf{V}_{\boldsymbol{\beta}}^{-\mathbf{1}} \boldsymbol{\mu}_{\boldsymbol{\beta}}$
  b: Compute $\boldsymbol{b} = \boldsymbol{y}^\top \boldsymbol{X}/\tau^2 - \boldsymbol{y}^\top \boldsymbol{v}/\tau^4 +$ and $\boldsymbol{B} = \boldsymbol{X}^\top \boldsymbol{X}/\tau^2 - \boldsymbol{X}^\top \boldsymbol{v}/\tau^4 + \boldsymbol{F}$
  c: Generate $\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{B}^{-1}\boldsymbol{b}, \boldsymbol{B}^{-1})$
Repeat steps **1** and **2** to obtain M MCMC samples for $\{\boldsymbol{\beta}, \theta, \tau^2\}$

as the new proposal. Convergence toward the desired acceptance rate is assured for an appropriate choice of the variance terms and of the adaptation rule (Neal et al. (2006); Roberts and Rosenthal (2009)). The algorithm has been coded using `R 4.0.1` statistical environment and `C++`, exploiting the interface provided by the `Rcpp` packageEddelbuettel et al. (2011). All expensive computations are managed by `Eigen` library (version 3.3.7, Guennebaud et al. (2010)), which provides efficient routines for linear algebra, matrix and vector operations and numerical solvers, with an emphasis on sparse matrices. Our implementation of (3.2.9) outperforms the algorithms that update $\boldsymbol{w}$ in terms of computational speed as it is implemented in the `spNNGP` package (Finley et al., 2017a). We present these comparisons in the Appendix A.

3.2.4 Including spatial effects

The inclusion of spatial covariates in such an experimental setting is often cumbersome from a practical perspective. Indeed, the (potentially retrievable) spatially-referenced features may be recorded at different resolutions, either among them or with respect to the observed process. A thorough preliminary data analysis would be required to build a coherent dataset and to not invalidate inference.

We circumvent the aforementioned issues by accounting for *spatial heterogeneity* through an unobserved (latent) predictable surface, as in the methods described in Chapter 2. In the considered setting, we build upon the temporal framework of Section 3.2.1 and rather than complicating the already convoluted dependence structure, we unload the spatial effect upon the mean. The underlying assumption is

that the contribution of environmental factors on the process mean varies smoothly over the domain of interest, and this ground truth can be expressed through a smooth function $f_S(\cdot) : \mathbb{R}^2 \to \mathbb{R}$. We can then approximate it using spline basis representation, as exposed in Section 2.2.1 (see, e.g., Ramsay and Silverman, 2007; Nürnberger and Zeilfelder, 2000; Goodman and Hardin, 2006).

Let J_x and J_y be the dimensions of independently defined B-spline basis expansions on the x and y coordinates, respectively. Then $f_S((x,y)) \approx \widetilde{f}_S((x,y)) = \sum_{j_X=1}^{J_X}\sum_{j_Y=1}^{J_Y}\beta_{S,(j_X,j_Y)}B_{x,j_x}(x)B_{y,j_y}(y)$, where $B_{x,j_x} = [\boldsymbol{B}_x]_{j_x}$ and $B_{y,j_y} = [\boldsymbol{B}_y]_{j_y}$ are the j_x-th and j_y-th element of the B-spline basis along the two axis. For any location $(x,y) \in \mathbb{R}^2$ the elements of the previous sum can be more compactly expressed through the tensor product basis $\boldsymbol{B}_S(x,y) = (\boldsymbol{B}_x \otimes \boldsymbol{B}_y)(x,y)$. The size of this basis is $J_S = J_x \cdot J_y$ and depends on the size of the two original spline basis, which in turn depends on the chosen number of knots $knots_x, knots_y$ and degree deg_x, deg_y (namely $J_c = knots_c + deg_c$ for $c = x, y$). We can now modify (3.2.1) to include the spline regression term as:

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{B}_S(\boldsymbol{\gamma}(t))\boldsymbol{\beta}_S + \boldsymbol{w}(t) + \boldsymbol{\varepsilon}(t), \qquad t \in \mathbb{R}^+, \tag{3.2.11}$$

where $\boldsymbol{\gamma}(t) = (\boldsymbol{\gamma}_1(t), \boldsymbol{\gamma}_2(t), \ldots, \boldsymbol{\gamma}_K(t))^\top$, $\boldsymbol{\gamma}_k(\cdot) = (\gamma_{k,x}(t), \gamma_{k,y}(t)) : \mathbb{R}^+ \to \mathbb{R}^2$ is the trajectory function mapping time t for individual k to its position and $\boldsymbol{B}_S(\boldsymbol{\gamma}(t))$ is the $K \times J_S$ matrix with row k corresponding to the J_S basis elements for the coordinates at time point t for individual k. Tuning of J_S (i.e. knots and degree) is required to fit a spline surface flexible enough to describe the spatial variations at the scale of interest without incurring in over-fitting. We must consider that our Actigraph data includes millions of observations in a limited study area, of which some assume different values in the same location (or in its immediate vicinity). Unless an incredibly fine grid of knots is chosen, over-fitting will not be an issue in most of the considered domain.

Let us denote with $\boldsymbol{B} = \boldsymbol{B}_S(\boldsymbol{\gamma}(\mathcal{T}))$ the $n \times J_S$ matrix containing the B-spline basis elements evaluated at the observed location of each individual $\boldsymbol{\gamma}(\mathcal{T}) = [\gamma_1(t_{11}), \gamma_1(t_{12}), \ldots, \gamma_K(t_{KT_K})]^\top$. Following Equation (3.2.9), posterior inference can be attained by sampling from:

$$p(\boldsymbol{\beta}, \boldsymbol{\beta}_S, \theta, \tau^2 \,|\, \boldsymbol{y}) \propto p(\theta, \tau^2) \times p_S(\boldsymbol{\beta}_S) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\beta}_S, \widetilde{\boldsymbol{C}}_\theta + \tau^2\boldsymbol{I}_n), \tag{3.2.12}$$

where $p_S(\cdot)$ is the prior over the spline regression coefficients.

We must consider that the collected data are sparsely distributed across the Westwood area, with some areas presenting only few observed points (trajectories are not uniformly distributed, see Figure 3.2). This may cause coefficients corresponding to those regions to be weakly identified, making the function locally prone to over-fitting and inducing large variance in the estimates. To control for the balance of these two components, we may assign ad-hoc priors to the spatial spline regression coefficients (Eilers and Marx, 1996) for penalizing deviation from a certain degree of smoothness and favoring identifiability. This yields the Bayesian P-Spline (Carter and Kohn, 1994; Hastie et al., 2000; Lang and Brezger, 2004) previously discussed in Section 2.2.3. While retaining a multivariate Gaussian as a prior, we effectuate shrinkage by choosing a suitable precision matrix $\boldsymbol{P}$ and introducing a shrinkage parameter λ at a deeper level of the hierarchy. To be precise:

$$\boldsymbol{\beta}_S \,|\, \lambda \propto \exp\left\{-\frac{\lambda}{2} \cdot \boldsymbol{\beta}_S \boldsymbol{P} \boldsymbol{\beta}_S^\top\right\}, \qquad \lambda \sim Ga(\alpha_\lambda, \beta_\lambda).$$

We consider two possible forms for $\boldsymbol{P}$, which imply different penalization on the values of the coefficients:

- **Ridge-like** prior, which is to say $\boldsymbol{P} = \boldsymbol{P}_{RL} = \boldsymbol{I}_{J_S}$;
- **First-order random walk** prior, which is to say:

$$\boldsymbol{P} = \boldsymbol{P}_{RW} : \; [\boldsymbol{P}_{RW}]_{ij} = \begin{cases} n_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

  where n_i is the number of neighbors of knot i and $i \sim j$ denotes a neighboring relationship between the knots.

Both precision matrices provide a multivariate Gaussian prior distribution on the coefficients. However, the latter is improper since $\text{rank}\,(\boldsymbol{P}_{RW}) < J_S$. Nevertheless, if the B-Spline basis elements are collected together with the other covariates $\boldsymbol{X}^* = [\boldsymbol{X}, \boldsymbol{B}]$ and the corresponding coefficients are stacked into the joint vector $\boldsymbol{\psi} = [\boldsymbol{\beta}, \boldsymbol{\beta}_S]$, then the posterior distribution of the latter is a proper multivariate Gaussian with full conditional distribution

$$\boldsymbol{\psi} \,|\, \cdot \propto \mathcal{N}_J \left(\boldsymbol{\psi} \,\middle|\, \boldsymbol{G}^{-1}\boldsymbol{g},\, \boldsymbol{G}^{-1} \right)$$

$$\boldsymbol{G} = \boldsymbol{X}^{*\top} \widetilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{X}^* + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$$

$$\boldsymbol{g} = \boldsymbol{X}^{*\top} \widetilde{\boldsymbol{\Lambda}}^{-1} \boldsymbol{y} + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\psi}}$$

where:

$$\boldsymbol{V}_{\boldsymbol{\psi}}^{-1} = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \lambda \cdot \boldsymbol{P} \end{bmatrix}, \qquad \boldsymbol{g} = \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{\mu}_{\boldsymbol{\beta}_S} \end{bmatrix} = \boldsymbol{0}.$$

Moreover, the Gamma prior on λ implies a Gamma full-conditional with updated parameters:

$$\lambda \,|\, \cdot \propto Ga \left(\lambda \,\middle|\, \alpha_\lambda + 1/2,\, \beta_\lambda + \boldsymbol{\beta}_S^\top \boldsymbol{P} \boldsymbol{\beta}_S \right).$$

Estimating the model in (3.2.11) is achieved through a straightforward extension of Algorithm 2, where $\boldsymbol{\psi}$ and λ can be jointly updated from their full conditional distributions. Algorithm 3 shows how the Gibbs' sampling step of Algorithm 2 can be modified in order to get full inference also on the spline coefficients $\boldsymbol{\beta}_S$ and the shrinkage parameter λ. In practical terms, this requires J_S additional linear coefficients to be estimated, whose size $p^* = p + J_S$ may undermine the efficiency of the algorithm. For example, calculations in Step 1b are quadratic with respect to $p^* \to \mathcal{O}(np^{*2})$. Fortunately, some tweaks can be contemplated in order to limit the additional computational cost. For instance, Steps 1a and 1b (i.e. the most expensive in p^*) must be executed in the first iteration and, subsequently, only in those iterations where new values of θ are accepted. When θ is rejected, the previously computed value would stay unchanged and can be retained in memory with few storage burden. Thus, if attain an optimal acceptance rate of $\approx 20\% - 30\%$ is attained in the Metropolis Hastings step on θ, the computation is avoided in the majority of cases and this yields a sensible improvement in computation time and speed.

Algorithm 3: $\boldsymbol{\psi}$ and λ Gibbs' update in the collapsed algorithm with shrinkage

1: Gibbs' sampler update for $\boldsymbol{\psi}$
$\boldsymbol{\psi}|\cdot \sim \mathcal{N}_J(\boldsymbol{G}^{-1}\boldsymbol{g}, \boldsymbol{G}^{-1})$, where $\boldsymbol{G} = \boldsymbol{X}^{*\top}\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^* + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$ and $\boldsymbol{g} = \boldsymbol{X}^{*\top}\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{y} + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\psi}}$
begin
- **a:** Compute $\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$ and $= \mathbf{V}_{\boldsymbol{\psi}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{\psi}}$
- **b:** Compute $\boldsymbol{g} = \boldsymbol{y}^\top\boldsymbol{X}^*/\tau^2 - \boldsymbol{y}^\top\boldsymbol{v}/\tau^4 +$ and $\boldsymbol{G} = \boldsymbol{X}^{*\top}\boldsymbol{X}^*/\tau^2 - \boldsymbol{X}^{*\top}\boldsymbol{v}/\tau^4 + \boldsymbol{F}$
- **c:** Generate $\boldsymbol{\psi} \sim \mathcal{N}_{p^*}(\boldsymbol{G}^{-1}\boldsymbol{g}, \boldsymbol{G}^{-1})$

2: Gibbs' sampler update for λ
$\lambda|\cdot \sim Ga\,(\alpha^*_\lambda,\, \beta^*_\lambda)$, where $\alpha^*_\lambda = \alpha_\lambda + 1/2$ and $\beta^*_\lambda = \beta_\lambda + \boldsymbol{\beta}_S^\top\boldsymbol{P}\boldsymbol{\beta}_S$
begin
- **a:** Compute $h = \boldsymbol{\beta}_S^\top\boldsymbol{P}\boldsymbol{\beta}_S$ and get: $\alpha^*_\lambda = \alpha_\lambda + 1/2$ and $\beta^*_\lambda = \beta_\lambda + h$
- **b:** Generate $\lambda \sim \mathcal{G}\,(\alpha^*_\lambda,\, \beta^*_\lambda)$

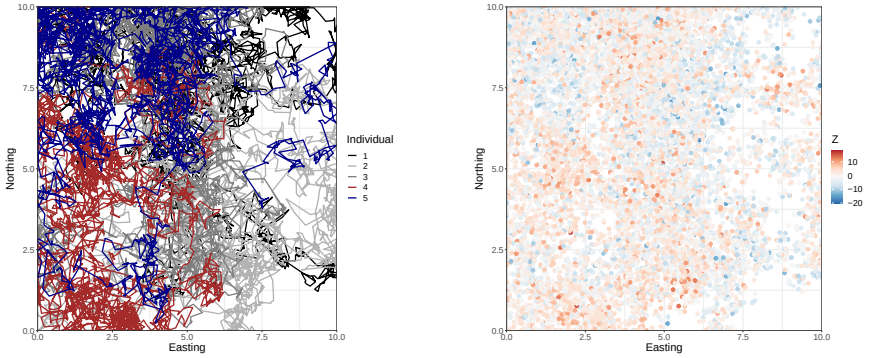
3.2.5 Simulations

We conduct simulation experiments to evaluate the model described in Section 3.2.4 and compare the performances of our algorithm in terms of fitting, prediction error and computational speed with other routines already available in the `spNNGP` package. Additional comparative experiments are provided in the supplementary material.

All the computations were performed using Cluster Terastat, an HPC infrastructure developed by the Department of Statistical Sciences (DSS) in the University of Rome *"La Sapienza"*, in collaboration with CINECA, for the resolution of mathematical and statistical models on big data. It is currently equipped with 12 modern computational nodes with 16 cores each (bringing the overall number of cores to 192), roughly equivalent to 3 TeraFlop/sec, and 64Gb of RAM (`https://www.dss.uniroma1.it/en/node/5870/technical-specifications`). Each of the following jobs, including the real data applications in Section 3.3, have been executed on a single node exploiting, whenever possible, all 16 cores.

In this simulation experiment we demonstrate our model's ability to disentangle the true temporal and spatial components, recovering the parameters of the temporal covariance and the spatial effect over the region of interest, when observing points belonging to random trajectories evolving in space. $T_k = 2 \times 10^4$ time points have been randomly generated for $K = 5$ individuals, where each time point t_{ki} has been obtained assuming exponential waiting times between observations, i.e. $t_{ki} = \sum_{h=1}^{i-1}\delta_h$, and $\delta_h \sim \text{Exp}(5)$, $\forall\, h$. Given the time points, fictitious spatial trajectories $\boldsymbol{\gamma}_k(\cdot)$ have been obtained by simulating a vector of locations $\boldsymbol{s}_k = [\gamma_k(t_{k1}), \ldots, \gamma_k(t_{kT_k})]^\top$, where subsequent components are independent Gaussian Random walks over the square $\mathcal{S} = (1, 10) \times (1, 10)$, with the variance of each step along the horizontal and vertical axis proportional to the elapsed time between two subsequent observations. Should the trajectory leave the squared domain, it would be projected onto the border and the following step would resume from there. The simulated trajectories are shown in Figure 3.4a. Conditionally on the time points and positions, the latent temporal Gaussian processes $\{\boldsymbol{w}_k\}_{k=1}^5$ have been generated (independently for each individual). Common covariance parameters $\theta_k = \theta \forall k$ have been chosen, with the covariance function belonging to the exponential family:

$$\mathbf{Cov}_\theta\,[w(t),\, w(t')] = c_\theta(t,\, t') = \sigma^2 e^{-\phi\cdot|t-t'|}, \quad \sigma^2, \phi \in \mathbb{R}^+. \tag{3.2.13}$$



(a) Trajectories **(b)** Points

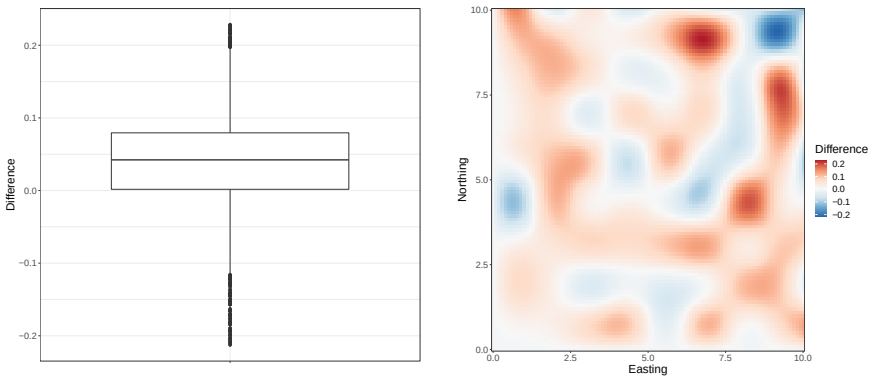
Figure 3.4. Example of observed trajectories (a) and observed points (b) for the simulated dataset

We here recall that $\sigma^2 > 0$ represents the variance of the process (sill), ϕ is the decay in temporal correlation (range) and τ^2 the residual variance (nugget). The spatial effects have been then introduced through spline function $f_S(\cdot)$ built as the tensor product of independent spline bases of degree 2, with 9 knots uniformly spread over the squared domain (including boundary knots). The spline coefficients $\boldsymbol{\beta}_S$ have been fixed to randomly generated values from $\mathcal{N}_{81}(\mathbf{0}, \lambda \boldsymbol{I}_{81})$ with $\lambda = 0.5$. The model also included an individual intercept for each individual $\{\beta_{0k}\}_{k=1}^{5}$ and the effect of 3 covariates with random values drawn independently at each location from a $\mathcal{N}(0,\, 1)$ distribution, leading to covariate vectors $\{\boldsymbol{x}_{ki}\}_{i=1}^{T_k}$, $k = 1, \ldots, K$. The effect of the covariates is assumed to be common across individuals, as it is determined by three random slopes $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]$. Finally, the outcome values for individual k at time t_{ki} and location $\boldsymbol{s}_{ki} = \gamma_k(t_{ki})$ have been generated according to the generative process of Equation (3.2.11) with parameters fixed as above.

This whole simulation setting produced a simulated dataset $D_{sim} = \left\{(\text{Ind}_j, t_j, \boldsymbol{s}_j, y_j, \boldsymbol{x}_j^\top)\right\}_{j=1}^{n}$ of $n = 10^5$ observations, where Ind_j denotes the individual corresponding to row j. The spatio-temporal model proposed in Section 3.2.4 is trained on the 70% of the total observations belonging to D_{sim}, while the the remaining 30% have been excluded to assess the out-of-sample predictive performances in terms *Relative Mean Squared Prediction Error* (RMSPE), *Root Mean Squared Prediction Error* (rMSPE), *Coverage*, *Predictive Interval Width* (PIW). Chain convergence has been verified through visual inspection of the chain behavior (traceplots, autocorrelations etc.) and automatic diagnostic tools (Geweke, Raftery, Heidelberger from the `coda` package (Plummer et al., 2006)). Intercept and slope regression parameters have been given a flat normal prior distribution $\mathcal{N}(0,\, 10^6)$; the variance components, σ^2 and τ^2, were both assigned inverse Gamma $\mathcal{IG}(2, 2)$ priors; the decay parameter ϕ received a Gamma prior $\mathcal{G}(1, 1)$. For the spline coefficients, we considered both the penalized versions proposed in Section 3.2.4; the first one will be denoted as S-Spline (shrinking splines), the second as P-Spline (penalized splines).

Param. (True)	S-Spline		P-Spline	
	Point	Interval	Point	Interval
β_{01} (−3.76)	-3.799	(-3.846,-3.752)	-3.797	(-3.844,-3.75)
β_{02} (0.65)	0.572	(0.523,0.62)	0.575	(0.526,0.623)
β_{03} (−0.60)	-0.649	(-0.697,-0.6)	-0.646	(-0.693,-0.598)
β_{04} (2.36)	2.326	(2.277,2.374)	2.328	(2.28,2.376)
β_{05} (−0.33)	-0.359	(-0.408,-0.31)	-0.356	(-0.404,-0.308)
β_1 (2.59)	2.599	(2.59,2.608)	2.599	(2.59,2.608)
β_2 (2.70)	2.691	(2.683,2.7)	2.691	(2.683,2.7)
β_3 (−0.58)	-0.586	(-0.595,-0.577)	-0.586	(-0.595,-0.577)
σ^2 (1)	1.001	(0.973,1.032)	0.993	(0.965,1.023)
ϕ (1)	0.994	(0.948,1.04)	1.01	(0.964,1.063)
τ^2 (1)	1.001	(0.984,1.018)	1.001	(0.984,1.018)
Metric	Out-of-sample	In-sample	Out-of-sample	In-sample
Coverage	0.95	0.99	0.95	0.99
RMSPE (r)	0.07 (1.18)	0.03 (0.84)	0.07 (1.19)	0.03 (0.84)
PIW	4.66	4.44	4.66	4.44
DIC	115’543		115’556	
Fitting time (h)	2.18		2.2	

Table 3.2. Parameter estimates, predictive validation and fitting times (hours) on the simulated dataset for all the considered models.

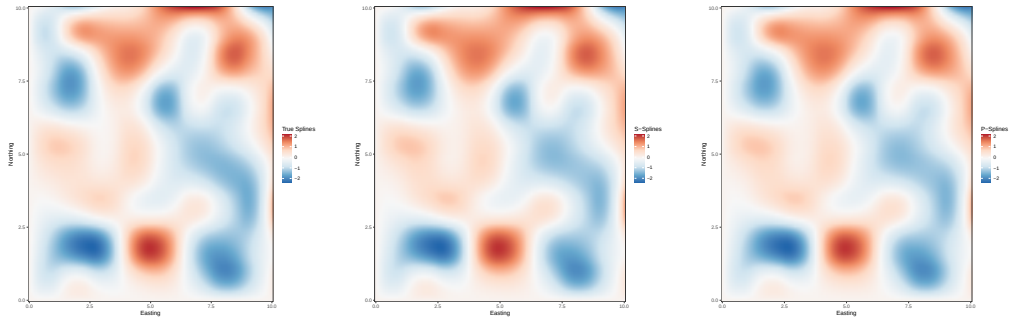


(a) Difference distribution **(b)** Spatial distribution of error

Figure 3.6

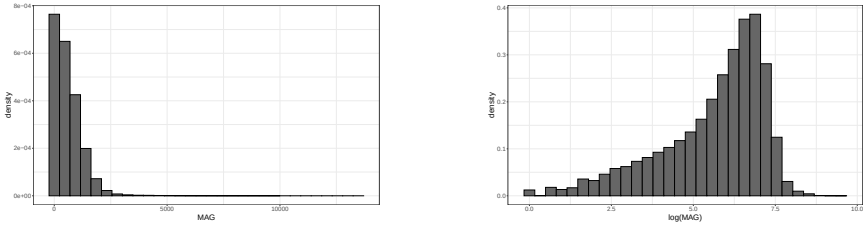
Table 3.2 presents the posterior estimates, performance metrics, and the fitting time for both models [8]. Performances under the two priors appear to be almost identical, but the DIC values suggest a slight better fit of the S-Spline model. This is not surprising, since the data have been generated using an analogous shrinkage prior for the $\boldsymbol{\beta}_S$'s. Figure 3.5 presents the posterior estimate of the spatial surface. The true latent surface is compared with the two (practically identical) estimates. The differences between the true spline surface and the estimates are visualized in Figure 3.6 and depicts a negligible upward bias.

[8] For the sake of brevity and readability, we did not include the spline coefficients in the table.



(a) True splines surface **(b)** Estimated S-Splines **(c)** Estimated P-Splines

Figure 3.5. True (top left) and estimated spline surfaces (bottom left and right), including the point-wise difference between the true one and the S-Spline estimated (top right).



(a) MAG **(b)** log(MAG)

Figure 3.7. Observed MAG (a) and $lMAG$ (b) in the whole sample.

3.3 Application

The data processing and merging of Actigraph data with GPS locations resulted in two final datasets (Section 3.1). These are treated separately, assuming the exponential covariance function on the underlying temporal process as in Eq. (3.2.13). Common covariance parameters $\theta_k = \theta \, \forall \, k$ are assumed across individuals. In both applications, the 70% of the total observations are used as training set, while the the remaining 30% have been excluded to assess the out-of-sample predictive performances in terms *Relative Mean Squared Prediction Error* (RMSPE), *Root Mean Squared Prediction Error* (rMSPE), *Coverage*, *Predictive Interval Width* (PIW). Performance of our model are always compared to standard linear regression, which is the current standard in the field.

3.3.1 Temporal model

We first analyze D_1. All numerical variables have been centered for improving the efficiency of the MCMC sampling (Gilks and Roberts, 1996). A dummy variable indicating if the measures are referred to the period before the *Bruin Bike Share* (BBS) launch (i.e. year 2017) or after (i.e. year 2018) has been included, so that the model could detect any effect of this new specific policy which aims at improving the physical activity level of the participants.

We account for the daily periodic behavior that characterizes most human activities by modeling the impact of the hour of the day on the physical activity level as a non-linear function $f_H(\cdot) : [7, 23) \to \mathbb{R}$, which is approximated by a finite linear combination of J_H spline basis functions $\phi_j(\cdot)$ with unknown coefficients $\beta_{H,j}$'s,

$$f_H(h) \approx \widetilde{f}_H(h) = \sum_{j=1}^{J_H} \beta_{H,j} B_{H,j}(h) = \boldsymbol{B}_H(h)\boldsymbol{\beta}_H.$$

The full process specification yields:

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{B}_H\left(h(t)\right)\boldsymbol{\beta}_H + \boldsymbol{w}(t) + \boldsymbol{\varepsilon}(t), \qquad t \in \mathbb{R}^+, \tag{3.3.1}$$

where $h(\cdot) : \mathbb{R}^+ \to [7, 23)$ links each time point to the corresponding hour of the day and $\boldsymbol{B}_H(\cdot) : [7, 23) \to \mathbb{R}^{J_H}$ links each hour of the day to the values of the spline at that point. The *hour of the day* spline is approximated using 4 internal knots, spread uniformly over the day times. Collecting the basis elements in the design matrix and stacking the coefficients as for the spatial model of Section 3.2.4, this boils down to simply introducing 6 additional columns in the design matrix (i.e. the spline basis functions evaluated at the observed time-points). Hence only 6 additional parameters need to be estimated in $\boldsymbol{\beta}$ and this is beneficial in that, given the large number of data points in each section of time and the reduced number of knots considered, we do not need to concern ourselves with over-fitting and robustness of inference.

Figure 3.7a shows the MAG distribution and highlights how it is heavily skewed. In order to correct for the skewness and to have the response variable belonging to a domain coherent with the Gaussian assumption, we consider its logarithmic transformation (see Figure 3.7b):

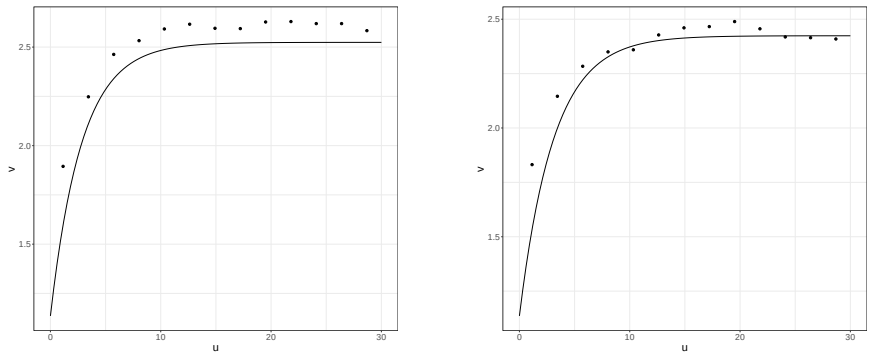
$$lMAG_k(t) = \log(MAG_k(t)), \qquad k = 1, \dots, K, \quad t = t_{k1}, \dots, t_{kT_k}. \tag{3.3.2}$$

Let *varname* be a short form (first three letters) of any variable name. We denote the parameter associated to each variable as β_{varname} and the levels of each categorical covariate as $\text{varname}_{(j)}$, with $j = 1, \dots, J_{\text{varname}}$. Hence, the final model specification can be written as:

$$\begin{aligned}
lMAG_k(t) &= \beta_0 + \mu_k + \mu_k(t) + w_k(t) + \epsilon_k(t) \\
\mu_k &= \sum_{j=2}^{J_{\text{Eth}}} \beta_{\text{Eth},j} \cdot \mathbb{I}\left(\text{Ethnicity}_k = \text{Eth}_{(j)}\right) + \sum_{j=2}^{J_{\text{Age}}} \beta_{\text{Age},j} \cdot \mathbb{I}\left(\text{AgeClass}_k = \text{Age}_{(j)}\right) + \\
&\quad + \beta_{\text{BMI}} \cdot \text{BMI}_k + \sum_{j=2}^{J_{\text{Sex}}} \beta_{\text{Sex},j} \cdot \mathbb{I}\left(\text{Sex}_k = \text{Sex}_{(j)}\right) \\
\mu_k(t) &= \beta_{\text{BBS}} \cdot \text{BruinBikeShare}(t) + \sum_{j=1}^{J_H} \beta_{H,j} B_{H,j}\left(h(t)\right) \\
w_k(t) &\overset{\text{ind}}{\sim} \mathcal{NNGP}\left(0, c_\theta(\cdot,\cdot)\right) \\
\epsilon_k(t) &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \tau^2\right)
\end{aligned} \tag{3.3.3}$$

where $\mathbb{I}(\cdot)$ is the indicator function of the statement between brackets and summation over categorical variables' levels starts from 2 because the *corner constraint*[9] has been adopted in order to guarantee the coefficients identifiability. Let us point out that, the baseline individual has been chosen (at random) as an Asian female with age in $[20, 25)$ years. Furthermore, since all the numerical variables have been centered about their mean, the baseline individual is the one with the mean value

[9] For each categorical variable one level is taken as the baseline and the coefficients referred to the remaining levels represent the variations with respect to the baseline



(a) Individual 1 **(b)** Individual 2

Figure 3.8. Variograms of the standard linear regression residuals on individuals 109 and 139

in all the numerical features. Other socioeconomic factors (e.g. education and income level) have been excluded from the analysis since highly associated with the already included ethnicity and age-class, while the *Lux* variable (detecting light exposition) has been dropped away after a preliminary run that highlighted its low predictive power[10]. The original $w_k(\cdot)$ is assigned a $\mathcal{NNGP}(0, \widetilde{c}_\theta(\cdot,\cdot))$ prior, with $\widetilde{c}_\theta(t,t')$ being the DAG-based approximation (Section 3.2.2) stemming from an *exponential* covariance function. The choice of the covariance family has been driven by an initial residual analysis of individual-specific ordinary least squares linear regressions. Figure 3.8 shows two examples of residual variograms and does reveals a stationary temporal structure, seemingly well-model by an exponential variogram.

We have assigned priors such as:

$$\begin{aligned}&\boldsymbol{\beta} \sim \mathcal{N}_J\left(\mathbf{0}, 10^6 \cdot \boldsymbol{I}_J\right)\\ &\sigma^2 \sim IG(2,2), \qquad \phi \sim \Gamma(1,1), \qquad \tau^2 \sim IG(2,2),\end{aligned}$$

with J being the total number of $\boldsymbol{\beta}$ coefficients.

Inference from (3.3.3) has been based on 10000 posterior samples after discarding an initial 5000 iterations as initial burn-in for the MCMC algorithm in Algorithm 2, where point estimates from the standard linear model were used as starting values for the regression coefficients. The run-time of the collapsed sampler (Section 3.2.3) on D_1 was ≈ 15 hours, achieving a desirable acceptance rate of $\approx 28\%$ at convergence.

3.3.2 Results from temporal analysis

Posterior estimates and performance metrics are presented in Table 3.3. The regression coefficients were slightly different from the ordinary least squares model, but presented very similar inference. African Americans, Latinos and Whites revealed higher values of $lMAG$ than Asian-Americans as did males over females. As expected, higher age-groups revealed lower $lMAG$. Unsurprisingly, the introduction of the temporal process effectuates slightly wider credible intervals for the regression parameters. This results in BMI being marginally less credible from the temporal process model than from linear regression. What is more surprising in both models

[10]The PASTA study did not contemplate a rigorous protocol for the light exposition sensor, and hence this variable is likely to not have been recorded accurately

Parameter	Collapsed NNGP		Linear regression	
	Point	Interval	Point	Interval
Intercept	5.514	(5.507, 5.520)	5.872	(5.854, 5.888)
Eth. Latin-American	0.166	(0.149, 0.183)	0.136	(0.131, 0.142)
Eth. White	0.073	(0.005, 0.095)	0.081	(0.076, 0.086)
Eth. Black or other	0.203	(0.184, 0.221)	0.164	(0.158, 0.170)
Sex Male	0.017	(0.003, 0.033)	0.023	(0.019, 0.027)
BMI	0.004	(-0.002, 0.01)	0.003	(0.002, 0.004)
Age [25-35)	-0.106	(-0.121, -0.091)	-0.124	(-0.129, -0.119)
Age [35-50)	-0.110	(-0.131, -0.09)	-0.123	(-0.129, -0.117)
Age [50-70]	-0.092	(-0.121, -0.065)	-0.144	(-0.152, -0.137)
BBS	-0.051	(-0.066, -0.037)	-0.067	(-0.071, -0.064)
σ^2	1.537	(1.528, 1.546)		
ϕ	0.315	(0.312, 0.319)		
τ^2	1.138	(1.135, 1.141)		
Metric	Out-of-sample	In-sample	Out-of-sample	In-sample
Coverage	0.94	0.97	0.94	0.94
RMSPE (r)	0.60 (1.24)	0.34 (0.93)	1 (1.59)	1 (1.59)
PIW	4.80	4.62	6.24	6.24

Table 3.3. Parameter credible intervals, $95\%(2.5\%, 97.5\%)$ and predictive validation for 15×10^3 MCMC iterations on D_1.

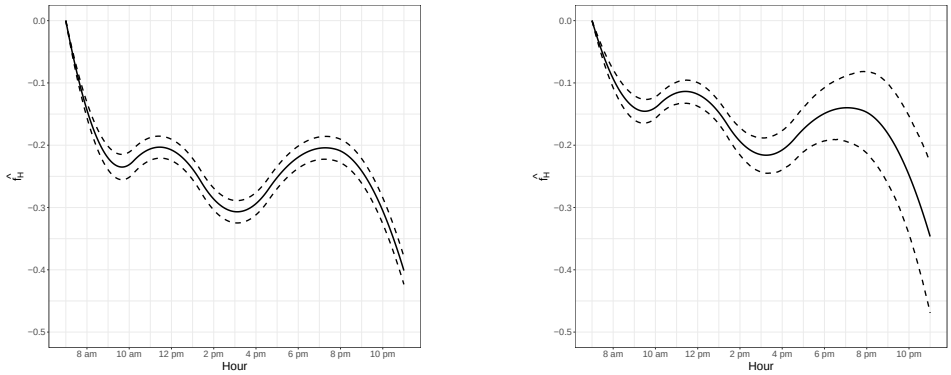
is the slightly negative effect of BBS on physical activity. This, however, is likely a consequence of the fact that at least $2 \cdot 10^6$ data points in D_1 corresponded to individuals outside of Westwood without access to the program. Any conclusion on its effect shall be further investigated.

The average of the $lMAG$ for the reference individual is represented by the common intercept, which is estimated ≈ 5.514 by our model. This implies a MAG per minute count of $1,488$, which corresponds to *hard* physical activity and an average MET of ≈ 7 according to the Table 3.1 and (3.1.2). This value, while large, is not surprising as we are modeling the epochs corresponding to active time.

Turning to the temporal decay, ϕ, we find a fairly sharp decline in temporal association with it dropping to below 0.05 after ≈ 1 minute. This time is derived from the *practical range* that, for the exponential covariance function, is computed as $\frac{1}{3\phi}$. While the estimated variance of the temporal process, σ^2, is slightly larger than τ^2, the latter's estimate indicates substantial residual variation beyond the temporal process—motivating our analysis in Section 3.3.3.

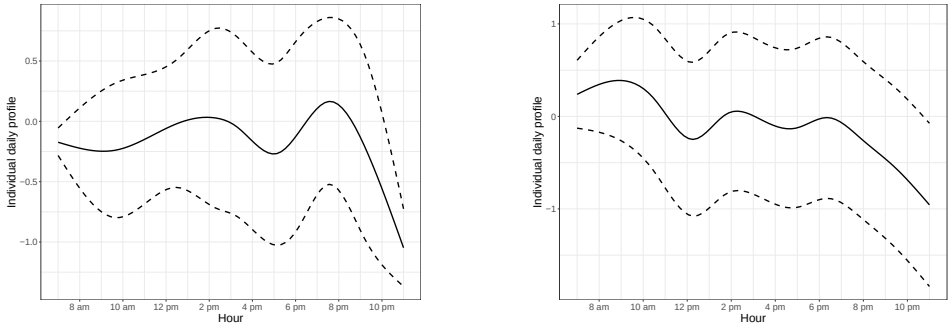
Comparison between the estimates of the *hour of the day* spline term is shown in Figure 3.9. The two models provide coherent patterns, but with slightly different magnitudes. It is way more pronounced in the linear model than in the temporal process model, where some of the temporal effect is likely to be absorbed by the temporal latent component. Combination of the *hour of the day* spline and of the temporal process can capture subject-specific diurnal variation in physical activity. This implies that our model can delivers statistical estimates (with uncertainty quantification) of personalized daily PA profiles for any individual for any day. For instance, Figure 3.10 presents the posterior estimates of daily MAGs (log) of two such individuals (number 204 (a) and 188 (b)) throughout the day to evince the inter-subject variation. This figure illustrates the need to accommodate variations among subjects when predicting their daily physical activities.

Eventually, the dependence structure of the temporal process sensibly improves



(a) Linear model **(b)** Collapsed NNGP

Figure 3.9. $\hat{f}_H$ for the linear model (left) and the NNGP model (right), with 95% credible intervals in dashed lines.



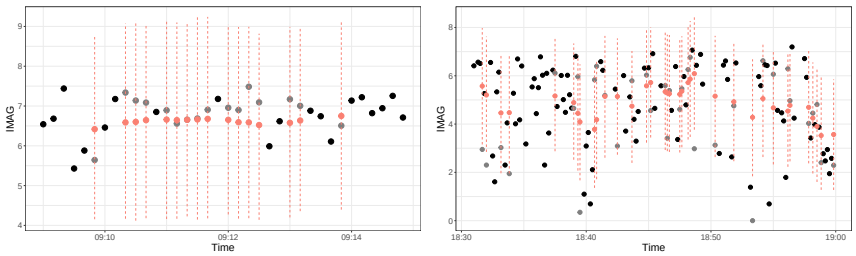
(a) **(b)**

Figure 3.10. Personalized PA profiles for two individuals estimated with 95% credible intervals (dashed lines) using the spline daily effect and the temporal process in (3.3.3).

the predictive performances either in terms of MSPE or PIW. While both modeling alternatives provide satisfactory coverage, the temporal process largely outperforms the standard linear model in all the other indices both in the training and in the testing set. An example of the out-of-sample predictions of the temporal process for a set of 100 subsequent points from one specific subject (number 77) is shown in Figure 3.11, which demonstrates the proposed model's ability to interpolate the $lMAG$ values at unobserved time-points or intervals. The interpolated values (red dots) provide a slightly over-smoothed but accurate reconstruction of the held out $lMAG$ (grey dots), which is always included in the corresponding 95% predictive bounds. This smoother behavior characterizes both in-sample and out-of-sample predictions when compared to the true values and is not necessarily a limitation of our model. Indeed, accelerations recorded by accelerometers are generally noisy, and the predicted values may be interpreted as a denoised version of the raw signal.

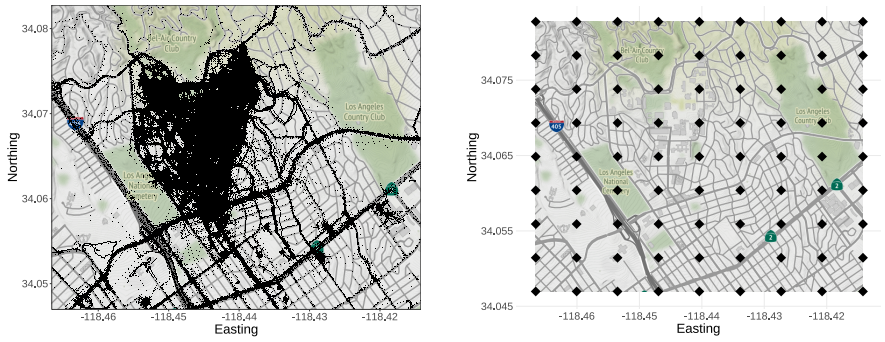
3.3.3 Spatial-temporal model

For the fit of the spatio-temporal model of Section 3.2.4 we consider D_2, where we can exploit the GPS information. We here recall that D_2 is restricted to only those observations recorded in the Westwood neighborhood of Los Angeles, which is one of the most popular districts on the Westside of the city and is the home of



(a) **(b)**

Figure 3.11. Out-of-sample predictions for two random individuals: black dots (observed values), grey dots (test set), pink dots (oos predictions), dashed line (95% confidence intervals)



(a) **(b)**

Figure 3.12. Observed locations (a) and knots (b) over the Westwood area.

UCLA.

The model accounts for *unobserved spatial heterogeneity* through a Bayesian Spline Regression, where the bi-variate spline basis has been obtained through the tensor product of two analogous uni-variate B-spline basis on longitude and latitude. The use of such approximation implies a rectangular domain, and hence the analysis is pursued on the square enveloping the Westwood area. There exist more sophisticated (yet computationally expensive) methods based on *finite element basis* theory and *thin-plate splines* that allow to build polynomial basis functions also over non rectangular domains (Sangalli et al. (2013)). Implementation and optimization of such methods will not be discussed here but is currently under consideration for future developments.

Two bases of degree 2 with 9 equally spaced knots over a square encompassing Westwood have been chosen. This sums up to $J_S = (7+2) \times (7+2) = 81$ terms for the complete spline basis, including the boundary knots. The position of the knots over the considered study area is shown in Figure 3.12b.

In practice, since locations are functions of time through the trajectory function $\gamma_k(\cdot)$, $k = 1, \dots, K$ of each individual, the time dependent component of the process

mean can be rewritten as:

$$\mu_k(t) = \beta_{\text{BBS}} \cdot \text{BruinBikeShare}(t) + \sum_{j=1}^{J_H} \beta_{H,j} B_{H,j}\left(h(t)\right) + \sum_{j=1}^{J_S} \beta_{S,j} B_{S,j}\left(\gamma_k(t)\right),$$

where $B_S = B_X \otimes B_Y$ is the tensor product bivariate spline. Given the reduced number of knots and the high spatial density of observations in several areas of the map (see Figure 3.12a), over-fitting is not a concern. However, there are also areas of Westwood which present sparsely observed data-points and the model can struggle to identify the spline coefficients referred to those areas, jeopardizing convergence of the MCMC algorithm. Therefore, the S-Spline (Ridge-like prior) has been ascribed to the spatial splines coefficients. The shrinkage parameter λ has been assigned a $\mathcal{G}(1,1)$ prior. Other parameters have been assigned the same priors of the temporal application in Section 3.3.1.

We ran 10000 MCMC iterations here. Adequate convergence was diagnosed after $5,000$ iterations and the last $5,000$ iterations were retained for posterior inference. Fitting the model to D_2 required ≈ 30 hours on Cluster Terastat. The acceptance rate obtained is $\approx 28\%$, supporting the consistency of our adaptive strategy.

3.3.4 Results from spatial-temporal analysis

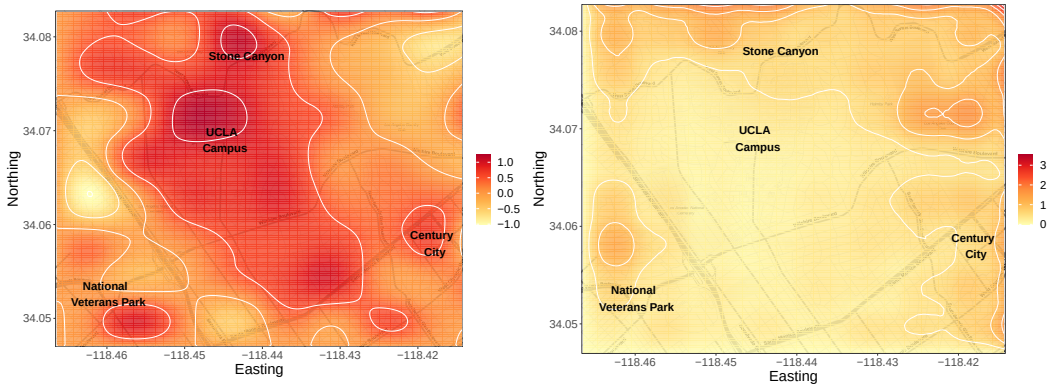
Table 3.4 presents parameter estimates and predictive performances of the spatio-temporal model, comparing it with the same model without the spatial spline component (i.e. the same of Section 3.3.1) and with a standard linear regression including the spatial spline terms (but neglecting any temporal dependence structure).

The coefficients estimated by the temporal model mimic those already obtained in Section 3.3.1, manifesting a good degree of consistency between the two datasets. Conclusions on the obtained regression coefficients, while apparently similar across the three different models, manifests some important differences. The general intercept is estimated to be ≈ 5.31 by the S-Spline model, implying MAG per minute count of 1214 (slightly lower than in the temporal application). This would correspond to *vigorous* physical activity and a MET of ≈ 6.5. In particular, this is a way lower value than the one attained by solely temporal model; this is probably because the former is able to fit area presenting particularly high PA intensities through the spline surface, without implying overall larger levels of PA. Furthermore, it is possible to notice a magnitude switch between coefficients associated to some relevant individual covariates (e.g. the ethnicity and the age class) after the spatial effect is included. There is then some confounding between individual characteristics and the areas where different people spend most of their time. The spatial model is discounting for this previously confounded factor and, for instance, provides a more reasonable interpretation of the age-group regression coefficients: the older the person the lower is the expected physical activity level. The BBS coefficient is estimated once again (by all alternatives) to have a negative impact on the physical activity level. This somewhat surprising finding, since individuals in D2 were all exposed to the program. Howeer, there are still some few factors that may negatively affect the estimation of the BBS variable. First, the observations after the BBS launch are mostly from the winter season (February to April, the coldest months in L.A. together with December), while the others include summer and autumn (June to November, the warmest months). Given that physical activity levels tend to be lower in the colder months, there is indication of some possible confounding between the BBS effect and seasonality. Second, there is no guarantee that a wrist-worn

Param.	Linear regression	Temporal	S-Spline
Intercept	5.613 (5.536, 5.689)	6.13 (6.102, 6.146)	5.31 (5.29, 5.33)
Eth. Latin-American	0.093 (0.079, 0.108)	0.105 (0.063, 0.149)	0.114 (0.069, 0.159)
Eth. White	0.053 (0.040, 0.067)	0.042 (-0.009, 0.095)	0.053 (0.004, 0.102)
Eth. Black or other	0.066 (0.052, 0.080)	0.111 (0.063, 0.157)	0.095 (0.054, 0.135)
Sex Male	0.019 (0.008, 0.029)	0.017 (-0.019, 0.053)	0.021 (-0.014, 0.055)
BMI	0.005 (0.003, 0.006)	0 (-0.014, 0.014)	0.006 (-0.007, 0.02)
Age [25-35)	-0.170 (-0.183, -0.156)	-0.373 (-0.408, -0.333)	-0.191 (-0.227, -0.155)
Age [35-50)	-0.217 (-0.233, -0.201)	-0.609 (-0.683, -0.531)	-0.249 (-0.298, -0.199)
Age [50-70]	-0.381 (-0.404, -0.359)	-0.429 (-0.48, -0.378)	-0.456 (-0.528, -0.384)
BBS	-0.008 (-0.091, -0.071)	-0.098 (-0.133, -0.062)	-0.107 (-0.140, -0.073)
σ^2		1.556 (1.531, 1.580)	1.489 (1.461, 1.517)
ϕ		0.336 (0.327, 0.346)	0.364 (0.351, 0.376)
τ^2		0.783 (0.776, 0.791)	0.777 (0.768, 0.786)
Metric (In-sample)			
Coverage	0.94 (0.95)	0.93 (0.97)	0.95 (0.99)
RMSPE	0.95 (0.94)	0.53 (0.26)	0.52 (0.25)
PIW	5.62 (5.60)	4.3 (4.1)	4.78 (4.74)

Table 3.4. Parameter estimates and predictive validation of the three models on D2.

actigraph may be able to detect properly the PA exerted while biking as compared to walking/running.



(a) Spline effect **(b)** Width

Figure 3.13. (a) Spatially smoothed estimates from a shrinkage spline over Westwood, Los Angeles; (b) width of 95% posterior predictive intervals for the shrinkage spline.

The estimate of ϕ implies that the dependence drops to ≈ 0 in less than a minute for both models including temporal dependence. Unsurprisingly, the temporal dependence provides the larger improvement in terms of performances, but the inclusion of the spatial effects is able to further improve the predictive performances in terms of RMSPE. In particular, the spatial-temporal model also delivers very satisfactory coverage, outperforming its competitors, but at the cost of a larger predictive interval width with respect to the only temporal model. This happening in correspondence to an improved coverage (both on training and test set) implies

that the model is able to better quantify the uncertainty in the observed data.

Figure 3.13a shows the estimated spatial surface, while Figure 3.13b presents the width of the posterior predictive intervals. The map clearly evinces zones (darker shades of red highlighted with white contours) that tend to depict high levels of physical activity. For example, the largest dark red blob in the north center-left almost perfectly tracks the UCLA campus boundary reflecting a campus environment with active mobility (walking, running, biking). Other zones of high activity identify with locations where more participants in the study live, including those residing in student dorms (northwest corner) and residential areas immediately around and in the predefined Westwood/UCLA study area (such as the south central zone) or Century City shopping center (to the east). Lighter shades (orange) correspond to areas that are less developed (open space), such as the areas in the north east; or they are areas with a high degree of transportation infrastructure and traffic (e.g., toward the western boundary). These correspond to highways (such as the Interstate-405 highway or other vehicular transportation corridors) that often have lower levels of activity because they inhibit outdoor physical activities due to noise, pollution, safety, etc. Our analysis reveals three additional high activity areas that are not gleaned from non-spatial models: the *Los Angeles National Veteran Park*; the *Century City shopping center* and the *Stone Canyon Park*. The color gradient closely follows the spatial characteristics of the Westwood neighborhood and reveal how spatial patterns can impact physical activity behavior after accounting for variation attributable to known explanatory variables.

Figure 3.14 shows two examples of observed (left) and reconstructed (right) MAGs along trajectories carved out by two subjects. We find a good degree of agreement between the two plots, and the ability of our model to recover the $lMAG$ in locations where it has not been observed. The reliability of the predictions can be proved through different metrics and, unsurprisingly, including the spatial effect and the temporal process improves predictive performances either in terms of MSPE or PIW. We deliver these personalized trajectory plots for every subject in the study and also predict personalized MAGs for each subject along any new trajectory. This enables personalized recommendations based upon an individual's health attributes including suggestions for more effective paths to follow for optimal physical activities, while also informing community level interventions in the built environment.

3.4 Conclusions and further developments

We have developed and executed a Bayesian modeling framework to conduct fully model-based inference for high-resolution accelerometer data. Our key data analytic developments included (i) modeling dependence over trajectories; (ii) accounting for subject-specific spatial-temporal variation for daily mobility; (iii) predicting or interpolating PA levels across trajectories; and (iv) identify zones of high physical activity in Westwood, Los Angeles. Our spatiotemporal analysis offers richer inference and evinces relationships between physical activity levels and a variety of factors, both at the subject level (e.g., personal attributes) and as a function of space and time. The temporal process was able to effectively glean the features of the data at finer resolutions, while the spatial splines accounted for residual spatial heterogeneity. Accommodating both temporal dependence and spatial heterogeneity demonstrably improved predictive ability and enabled us to effectively delineate zones of high physical activity. Furthermore, the ability of the model to pool information across individuals at all time points allows us to infer about those who present sparsely observed space-time points (due to technical issues or protocol violation).

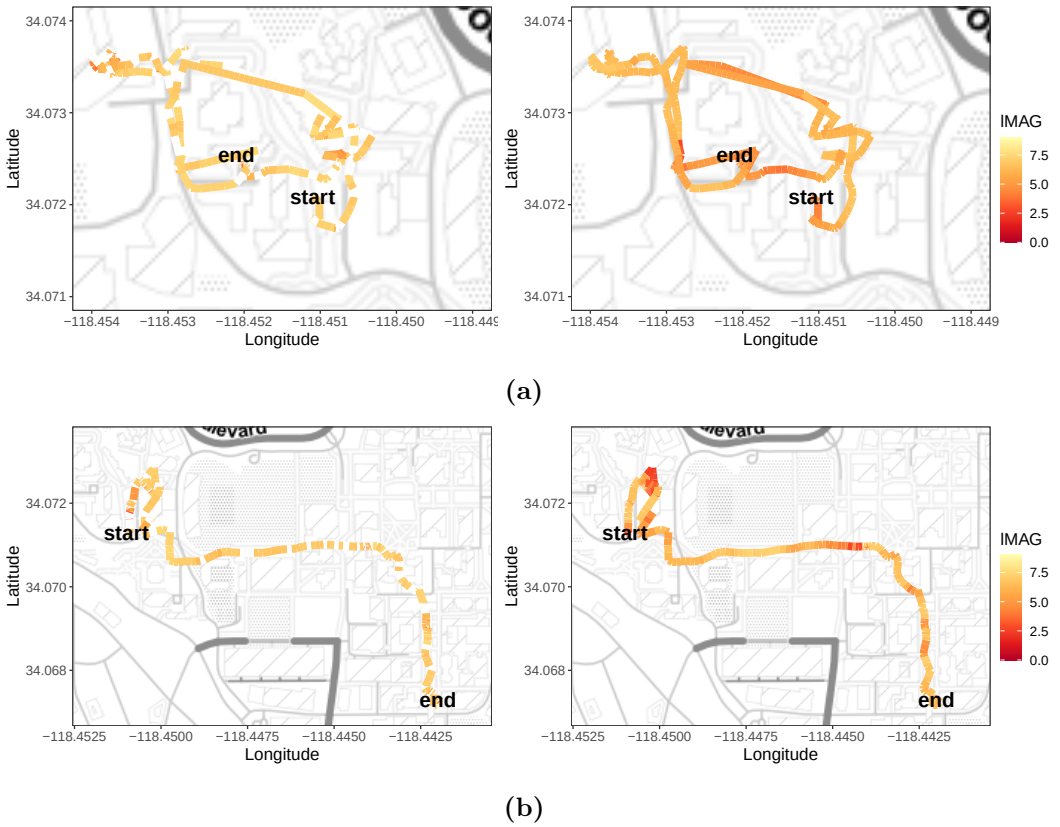


Figure 3.14. Two randomly chosen $lMAG$ trajectories over Westwood from individual 204 (a) and individual 566 (b). Observed trajectories with gaps are seen on the left panels; spatially reconstructed (predicted) trajectories are seen on the right panels.

In particular, given our improved predictive power, we can fill gaps and infer about PA levels with good accuracy and ensure the desired coverage by our prediction intervals.

Our analysis also resolves practical difficulties in using actigraph data. It is not cost-effective to deploy research-grade GlobalSat GPS and Actigraph units as they are very expensive and continued usage requires heavy staff involvement. Our methods can be applied to analogous, but less complete, data derived from smart phones and smart watches, then such devices could be deployed in much larger studies with much larger sample sizes at a fraction of the cost. Given the spatiotemporal nature of outdoor PA research, our ability to predict in areas of data missingness drastically improves inference related to the impacts of the built and natural environments on physical activity and active mobility.

While our approach offers trajectory-based inference for actigraph data, we recognize that there are several avenues for further research. Our DAG-based approach for scalable temporal processes can be further enriched with recent developments (Katzfuss and Guinness, 2021; Peruzzi et al., 2020a), although any of the methods reviewed and evaluated by Heaton et al. (2019) can be incorporated into our framework. Finally, there is possible merit in modeling the activity counts in each axis jointly and relaxing the assumptions of Gaussianity using recent developments in multivariate spatiotemporal count models and for non-Gaussian outcomes (see, e.g. Bradley et al., 2018, 2020).

Recent public health reviews call for interdisciplinary technological advances to more effectively measure spatiotemporal energetics of activity spaces in obesity and chronic disease research (James et al., 2016; Kestens et al., 2017; Drewnowski et al., 2020). Individual-level data, at aggregate, can be used to identify anchor points for physical activity and reveal causal pathways between built environment exposures and health. Our work is a novel contribution demonstrating methodologies for how these pressing research questions may be answered.

Funding

The work of the authors has been supported in part by National Science Foundation (NSF) under grants NSF/DMS 1916349 and NSF/IIS 1562303, and by the National Institute of Environmental Health Sciences (NIEHS) under grants R01ES030210 and 5R01ES027027.

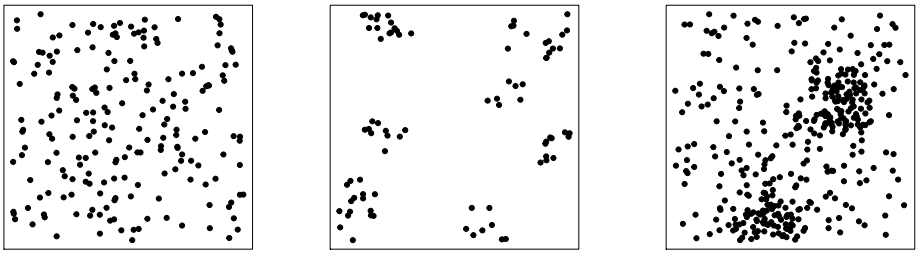
Chapter 4

Point patterns

Chapter 2 discussed the analysis of continuous spatial variations, namely the study of dependence patterns between measurements recorded at different times and/or locations on continuous domains. The time and location of sampling occasions are seen as deterministic, known values. It is conditionally upon them that inference on the distribution of the outcome of interest is performed.

This approach is entirely reasonable when observation points are pre-determined by a specific sampling design, defined *a-priori* of the data-collection process. However, in many applications, the researcher has no control over the sampling locations' positioning. This can happen either because they are determined by nature, or he is just an observer of events happening out of his control, e.g.: locations of trees in a forest (Gerrard et al., 1969), crimes committed in a city (Mohler et al., 2011), nuclei in a microscopic section of tissue (Diggle, 1986), etc. In such cases, the set of irregularly scattered observed locations $\mathcal{S} \subset \mathbb{R}^d$ is a random object by itself. If the interest still lies in the sole study of spatial variations, then this additional source of randomness may be ignored (as long as sampling bias can be reasonably excluded).

Nevertheless, there are many contexts in which the spatial distribution of locations is relevant from a topological perspective. For example, in species sampling in ecological sciences (Cormack, 1979; Baddeley and Turner, 2000; Perry et al., 2006), where the sightings of individuals belonging to different species are used to infer their mutual distribution in a region. In epidemiological applications (Breslow et al., 1980; Besag and Newell, 1991), where points are reference locations for a disease's number of cases in a geographical region. In finance (Björk et al., 1997; Bauwens and Hautsch, 2009; Hawkes, 2018), where the events may be market shocks and identifying regularities can drive future investments. We generally refer to this kind of data as *point pattern data*, and refer to the locations as *events* (to distinguish these from arbitrary points in the region). The main inferential objective of their study lies in distinguishing areas (sub-regions) of the underlying domain by the level of presence/risk of an event. In practice, the researcher seeks to find interpretable and predictable deviations of the observed pattern from the so-called *Complete Spatial Randomness* (CSR), that corresponds to events dropped *uniformly* over the domain. Loosely speaking, these deviations may be due to: *clustering*, in which events are more likely to appear close in space; *inhibition*, in which events are likely to appear at some distance from each other; *regularity*, when the events' density changes regularly over the domain. Three illustrative examples are shown in Figure 4.1. The one we proposed is an over-simplification of the essential characteristics a point pattern may exhibit, which should be opportunely qualified in more comprehensive summary statistics or by identifying the underlying generative mechanism.



(a) CSR **(b)** Clustering **(c)** Uneven distribution

Figure 4.1. Examples of random point patterns on the unit square.

As an extension, one may also consider assigning labels $Y(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{S}$ to the events; the latter yields a *marked point pattern*, where marks may be features of any kind (e.g., qualitative, discrete or continuous measures). The introduction of marks enlarges the space of interest's dimension, bridging the gap with spatial variations. Indeed, one may think of modeling the point pattern first, and the distribution of the marks given the pattern of points after.

This chapter will first give an introduction to the basics of standard *point patterns* analysis (Sections 4.1, 4.2). The literature is very vast and embraces many different perspective. Here, we take the *generative mechanism* angle and provide a brief description of stochastic processes (and models) that have a pattern of events as their realization: *point processes*. Basic references are Daley and Vere-Jones (2007) for the general theory of point processes, Last and Brandt (1995) for temporal and Moller and Waagepetersen (2003) and Diggle (2013) for spatial analysis of point patterns. Later on, Section 4.3 focuses on the *Hawkes process* and its spatio-temporal version, of which a more comprehensive account is provided (Laub et al., 2015; Reinhart et al., 2018). Such focus on the Hawkes process is because it is the key feature of the model adopted in the two applications of Chapter 5 on *road accidents* occurrences. Let me point out that the second application is the result of a collaboration with colleagues from the University of Warwick, recently published in Kalair et al. (2020).

4.1 Analysis of spatial point patterns

A point pattern over a specified region $\mathcal{D} \in \mathbb{R}^d$ is a set of locations $\mathcal{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$ where the locations are viewed as *random*. Randomness concerns both the *number* of locations and their *configuration* over the region $\mathcal{D}$. Proper modeling considerations about the realized point patterns shall consider different aspects of the mechanism that generated it, among which the study region and sampling from it are the most important.

Standard literature usually puts a division between *sparse sampling*, where limited information is recorded from a large number of smaller regions (Du Rietz, 1929; Cottam and Curtis, 1949; Cox, 1955; Diggle and Cox, 1983), and *intensive mapping*, where the events in a region are recorded at the point-scale level (Baddeley and Turner, 2000; Diggle, 2013; Illian et al., 2008; Moller and Waagepetersen, 2003). The former alternative was more common in the original ecological and epidemiological literature, mostly because of the technological limitations that did not allow accurate mapping at the exact location level. These can generally be tackled within the context of *discrete spatial variations* (Besag, 1974; Diggle et al., 1976)(lattice models,

Markov random fields, etc.) and will not be discussed in the sequel.

On the other hand, the last decades have seen a surge in the application of point patterns presenting the exact location information that was, indeed, not often available in the past. Historically, their study has been developed within a very formal probabilistic framework without proper attention to inference and applications. This has substantially changed over the last 20 years.

Another interesting aspect is whether the realized point pattern extends over the boundaries of the region $\mathcal{D}$ and belongs to a larger (potentially infinite) region, or exists only over $\mathcal{D}$. In the first case, the *edge-effects* and the shape of the region might matter, and appropriate tweaks must be contemplated not to invalidate inference (buffering regions, on-average adjustments, wrapping the region onto itself, etc.).

The typical modeling of a point pattern starts from considering it is as the realization of a *point process* $\mathcal{P}$ over $\mathcal{D}$. The next section provides a brief introduction to the theory of finite point processes, highlighting some essential theoretical properties in terms of summary functions regulating their behavior. Section 4.1.2 mirrors the previous one, introducing the empirical counterparts of these theoretical summaries.

4.1.1 A brief introduction to finite point processes

Most point patterns encountered in practice are observed in bounded regions, either because it is the whole domain where the phenomenon occurs, either because data have been mapped in a smaller window for convenience. In any case, the resulting map always contains a finite number of points and hence the focus is usually restricted to processes satisfying the finiteness condition. Loosely speaking, it is possible to define a (simple) *finite point process* X over a bounded and connected $\mathcal{D} \subset \mathbb{R}^d$ as a stochastic process whose generic realization $\boldsymbol{X}$ is a finite set of locations in $\mathcal{D}$. In general spaces these realizations are *unordered* sets of distinct points:

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\} \in \mathcal{X} \subset \mathcal{D}^l,$$

where $l \in \{0, 1, \ldots\}$ is not fixed but a feature of the random realization[1]. In particular, the finite-ness implies that:

$$N(\mathcal{D}) = \sum_{\boldsymbol{x}_1 \in \boldsymbol{X}} I_{\mathcal{D}}(\boldsymbol{x}_i) < \infty \quad \forall\, \boldsymbol{X} \in \mathcal{X},$$

where $N(\cdot)$ is the counting measure of X defined for any subset $A \subseteq \mathcal{D}$ and $I_A(\boldsymbol{x}_i)$ is the indicator function over the set A.

A probabilistic model for X must place a distribution over all its possible realizations. Let us suppose that $\mathcal{D}$ is equipped with a measure $0 < \nu(\mathcal{D}) < \infty$ on its Borel sets $\mathcal{B}(\mathcal{D})$. Then, it is sufficient to specify:

- a discrete probability distribution $p(\cdot) : \mathbb{N}_0 \to [0, 1]$ for the number of points in the whole domain $N(\mathcal{D}) = l$;
- a family of multivariate symmetric probability location densities $f_l(\cdot) : \mathcal{D}^l \to \mathbb{R}^+$, $l \in \mathbb{N}$ with respect to the l-fold product of $\nu(\cdot)$ for the points themselves.

The symmetry requirement is necessary in order to make sure that patterns generated by $f_l(\cdot)$ are permutation invariant, and therefore do not depend on the order in which the points are listed. It is then possible to formally define the point process

[1] If one dimension is *time*, then it induces an order among locations and some peculiarities of this specific case are discussed in Section 4.3.1.

distribution through a marginal-conditional form known as the *Janossy density* (Jánossy, 1950):

$$j_l(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l) = p(l) \cdot f_l(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l) \cdot l!, \quad l \in \mathbb{N}_0, \tag{4.1.1}$$

where the $l!$ appears because the $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$ points can be assigned to the l locations in $l!$ ways. In an infinitesimal sense $j_l(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l) \prod_{i=1}^{l} \delta(\boldsymbol{x}_i)$ is the probability of finding *exactly* l points, one at each infinitesimal region centered at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$. Stationarity of the process is assessed in terms of the location distribution, and is verified if and only if:

$$f_l(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l) = f_l(\boldsymbol{x}_1 + \boldsymbol{h}, \ldots, \boldsymbol{x}_l + \boldsymbol{h})$$

for all l, $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\}$ and $\boldsymbol{h}$. While looking intuitive, finding the right ingredients to define a valid point process using the Janossy distribution is challenging and presents a viable solution only in rare cases. Nevertheless, any point pattern realization can equivalently be expressed in terms of the corresponding counting measure $N(\cdot) : \mathcal{B}(\mathcal{D}) \to \mathbb{N}^0$ over *Borel* sets of its domain, including *void sets*. There is indeed a *one-to-one* relationship between point processes on general spaces (even not bounded) and this measure. Informally, it is evident how a point pattern must determine the counting measure for any $A \in \mathcal{B}(\mathcal{D})$; by converse, it is trivial how isolating all points through arbitrary union and intersection of sets in $\mathcal{B}(\mathcal{D})$ the point pattern can be uniquely determined from the counting measure. Therefore, mathematically speaking, the family of all joint distributions $(N(A_1), \ldots, N(A_m))$ over collections of disjoint sets $A_i \in \mathcal{B}(\mathcal{D})$ defines uniquely the point process X. This perspective represents a more attractive (and often more flexible) alternative to define point processes on general spaces. For a more detailed and technical illustration, the reader is referred to Chapter 16 of Gelfand et al. (2010) and Daley and Vere-Jones (2007).

The most relevant properties of point processes can be defined in terms of moments of the counting measure (Campbell measure, Palm distributions, etc.). Here, without delving into the most formal aspects, we discuss some common *first* and *second-order* properties and ignore higher-order ones.

First-order properties consider the points individually (i.e. do not consider interactions). In particular, the first moment measure is the set function $M(A) = \mathbb{E}[N(A)]$ with $A \in \mathcal{B}(\mathcal{D})$, which returns the expected number of points falling into any subset of $\mathcal{D}$. If $M(\cdot)$ is absolutely continuous with respect to the Lebesgue measure, then it can be computed as:

$$M(A) = \int_A \lambda(\boldsymbol{x}) d\boldsymbol{x},$$

where $\lambda(\cdot)$ is referred to as the intensity function of the process:

$$\lambda(\boldsymbol{x}) = \lim_{|d\boldsymbol{x}| \to 0} \frac{\mathbb{E}[N(d\boldsymbol{x})]}{|d\boldsymbol{x}|}. \tag{4.1.2}$$

For a simple point process (no multiple points can fall into the same infinitesimal region) we have that $\mathbb{E}[N(d\boldsymbol{x})] \approx P(N(d\boldsymbol{x}) > 0) \approx \lambda(\boldsymbol{x}) d\boldsymbol{x}$. Thus, under mild assumptions, it is interpretable as the density of points over $\mathcal{D}$. The intensity function is the counterpart of the mean-function for Gaussian processes (see Section 1.1) and indeed, for a stationary point process over $\mathcal{D}$, the following holds: $\lambda(\boldsymbol{x}) = \lambda \; \forall \, \boldsymbol{x} \in \mathcal{D}$.

Second-order properties refer to properties of the model that consider the points in pairs. The second moment measure $M(A_1 \times A_2) = \mathbb{E}\left[\sum_{\boldsymbol{x},\boldsymbol{x}' \in X} I_{A_1}(\boldsymbol{x}) I_{A_2}(\boldsymbol{x}')\right]$ counts the expected number of pairs of points with one falling in A_1 and the other in A_2. Under measurability it can be expressed as:

$$M(A_1 \times A_2) = \int_{A_1} \int_{A_2} \lambda_2(\boldsymbol{x}, \boldsymbol{x}') d\boldsymbol{x} d\boldsymbol{x}',$$

with $M(A_1 \times A_2) = \mathbb{E}\left[N(A_1) \cdot N(A_2)\right]$ for disjoint sets $A_1 \cap A_2 = \emptyset$. Hence, second order intensity function can be defined as:

$$\lambda_2(\boldsymbol{x}, \boldsymbol{x}') = \lim_{|d\boldsymbol{x}| \to 0, |d\boldsymbol{x}'| \to 0} \frac{\mathbb{E}\left[N(d\boldsymbol{x}) \cdot N(d\boldsymbol{x}')\right]}{|d\boldsymbol{x}||d\boldsymbol{x}'|}. \tag{4.1.3}$$

Under second-order stationarity $\lambda_2(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) = \lambda_2(\boldsymbol{h})$ and, if isotropy holds, $\lambda_2(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) = \lambda_2(h)$, with $h = ||\boldsymbol{h}||$. Considering that:

$$\mathbb{E}\left[N(d\boldsymbol{x}) N(d\boldsymbol{x}')\right] \approx P\left(N(d\boldsymbol{x}) > 0,\, N(d\boldsymbol{x}') > 0\right) \approx \lambda_2(\boldsymbol{x}, \boldsymbol{x}') d\boldsymbol{x} d\boldsymbol{x}',$$

a limiting interpretation of the second order intensity is as the probability of observing two points in the infinitesimal regions $d\boldsymbol{x}$ and $d\boldsymbol{x}'$, respectively. In practice, it can be interpreted as the density of inter-point distances.

From the second moment measure it is possible to derive the *pair correlation function*, namely:

$$g(\boldsymbol{x}, \boldsymbol{x}') = \frac{\lambda_2(\boldsymbol{x}, \boldsymbol{x}')}{\lambda(\boldsymbol{x})\lambda(\boldsymbol{x}')}, \tag{4.1.4}$$

that simplifies to $\frac{\lambda_2(\boldsymbol{x}, \boldsymbol{x}')}{\lambda^2}$ for a stationary process. If the correlation function depends only on the spatial separation between the two points $g(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) = g(\boldsymbol{h})$ the process is said to be *intensity re-weighted second order stationary* (Moller and Waagepetersen, 2003). It simplifies to $g(h)$ under isotropy.

An alternative characterization of the second-order properties for a stationary and isotropic process is provided by the so-called K-function, one definition of which is:

$$K(h) = \lambda^{-1} \mathbb{E}\left[N_{\boldsymbol{o}}(h)\right],$$

where $N_{\boldsymbol{o}}(h)$ is the number of events within distance h from an *arbitrary event* $\boldsymbol{o}$. The formal definition of an arbitrary event is quite complicated, and the reader is referred to Daley and Vere-Jones (2007) for a more technical discussion. Intuitively, it is possible to define it as an event selected at random from the population, hence a point distributed uniformly over $\mathcal{D}$. For instance, under stationarity, any point is a *typical* point and is equivalent to each other. The link between this function and the second-order properties is contained in the so-called *Papangelou conditional intensity* of a process:

$$\lambda(\boldsymbol{x} \,|\, \boldsymbol{X}) = \begin{cases} \lambda(\boldsymbol{x} \,|\, \boldsymbol{X} \setminus \boldsymbol{x}) & \boldsymbol{x} \in \boldsymbol{X} \\ \lambda(\boldsymbol{x} \,|\, \boldsymbol{X}) & \boldsymbol{x} \notin \boldsymbol{X}. \end{cases} \tag{4.1.5}$$

It is the intensity at a location $\boldsymbol{x}$ given the realization of a configuration $\boldsymbol{X}$ where, roughly speaking, $\lambda(d\boldsymbol{x}|\boldsymbol{X})$ represents the probability of observing a point in $d\boldsymbol{x}$ given that the remaining points are outside of $d\boldsymbol{x}$. Under mild assumptions, it is always possible to write:

$$\lambda(\boldsymbol{x}|\boldsymbol{X}) = \frac{f(\{\boldsymbol{x}, \boldsymbol{X}\})}{f(\boldsymbol{X})},$$

where $f(\cdot)$ is the probability measure over alternative configurations of the process, and $\{\boldsymbol{x}, \boldsymbol{X}\}$ is the configuration obtained by adding point $\boldsymbol{x}$ to configuration $\boldsymbol{X}$. Whilst apparently puzzling, the *Papangelou conditional intensity* is a very powerful tool through which a lot of interesting processes can be easily defined through with respect to other more simplistic processes .

Considering points in pairs is practically equivalent to take $\boldsymbol{X}$ as the single arbitrary point $\boldsymbol{o}$ and then study the conditional density of an additional point $\boldsymbol{x}$:

$$\lambda(\boldsymbol{x} \mid \boldsymbol{o}) = \frac{\lambda_2(\boldsymbol{x}, \boldsymbol{o})}{\lambda(\boldsymbol{o})} = \frac{\lambda_2(h)}{\lambda},$$

where $h = ||\boldsymbol{x} - \boldsymbol{o}||$ and the last equality holds under isotropy and stationarity. Hence, the expected number of events within distance h of an arbitrary event $\boldsymbol{o}$ (in a circle of radius h centered in $\boldsymbol{o}$) does not depend on $\boldsymbol{o}$ and can be computed as:

$$\begin{aligned} K(h) &= \lambda^{-1} \cdot \int_0^{2\pi} \int_0^h h \cdot \lambda(h \mid \boldsymbol{o}) \, dh \, d\theta = \frac{2\pi}{\lambda} \int_0^h h \cdot \frac{\lambda_2(h)}{\lambda} \, dh \\ &= \frac{2\pi}{\lambda^2} \int_0^h h \cdot \lambda_2(h) \, dh. \end{aligned} \tag{4.1.6}$$

4.1.2 Empirical estimation of summary properties

All the theoretical quantities introduced in the previous section have empirical counterparts, computable on any observed point pattern. The typical flow of the analysis starts by comparing the empirical estimates obtained on the observed patterns with the theoretical forms arising from the probabilistic properties of known point processes. Indeed, differently from the *mean* and *covariance* function approach of Chapter 2, there is no practical way to specify theoretical parametric forms for these summary functions as models on their own right while ensuring the validity of the resulting point process. The theoretical shapes must instead be derived directly from the underlying stochastic model, as the result of the specific generative process. The baseline comparison terms are usually represented by *CSR* and by the concepts of clustering, inhibition, and stationarity.

Let $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be an observed point pattern over $\mathcal{D}$. Under the assumption of stationarity, the study of its properties starts from the analysis of its empirical second order structure. It is usually investigated in terms of the so-called *nearest neighbor* $G(\cdot)$ and *empty space* $F(\cdot)$ distributions:

$$G(h) = P(\text{nearest event} \leq h), \qquad F(h) = P(\text{nearest event} \leq h),$$

where the first consider *event to event* distances, while the latter any *point to event* distance. Theoretical benchmarks for $G(\cdot)$ and $F(\cdot)$ can be computed via Monte-Carlo integration, simulating point patterns according to the chosen random mechanism and computing the corresponding *typical* distributions. For instance, under CSR, we have that $G(h) = F(h) = 1 - \exp(-\lambda h)$ for points in $\mathbb{R}$ and $G(h) = F(h) = 1 - \exp(-\lambda \pi h^2)$ for points in $\mathbb{R}^2$. This results arise from the correspondence of CSR with the Poisson process, which is defined and described in Section 4.2. Empirical estimates of $G(\cdot)$ arise from the n observed nearest neighbor distances:

$$\widetilde{h}_i = \arg \min_{j=1,\ldots,n} ||\boldsymbol{s}_i - \boldsymbol{s}_j||, \quad i = 1, \ldots, n,$$

where the estimation shall also take proper account of edge effects, when performed on bounded regions. Indeed, if $\boldsymbol{s}_i$ is at distance $\widetilde{h}_i^{max}$ from the edge, then events

further than $\widetilde{h}_i^{max}$ cannot be observed. An example of edge-corrected empirical estimate is obtained as:

$$\hat{G}(h) = \sum_{i=1}^{n} \frac{I_{(\widetilde{h}_i, \widetilde{h}_i^{max})}(h)}{I_{(0, \widetilde{h}_i^{max})}(h)}.$$

The empty space distribution $F(\cdot)$ can be estimated analogously, where a grid of arbitrary points $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\}$ is super-imposed on $\mathcal{D}$ and the l nearest neighbor distances from these are considered to build-up $\hat{F}(\cdot)$. The arbitrariness of the grid choice slightly diminishes the value of this diagnostic tool. However, the fact that under CSR $G(h) = F(h)$ can be exploited. Indeed, it is possible to define a new function as the ratio of the two:

$$J(h) = \frac{1 - G(h)}{1 - F(h)},$$

which is 1 in the case of CSR, while indicates clustering for $J(h) < 1$ and inhibition for $J(h) > 1$. The corresponding customary estimate is:

$$\hat{J}(h) = \frac{1 - \hat{G}(h)}{1 - \hat{F}(h)},$$

and its value informs about the observed direction and magnitude of the (eventual) deviation from CSR.

Next, we may consider the K-function introduced in Equation (4.1.6). Also here the theoretical forms, when not available analytically, can be approximated through Monte-Carlo simulations. It can be proved that under CSR $K(h) = h$ for point processes over $\mathbb{R}$ and $K(h) = \pi h^2$ for processes over $\mathbb{R}^2$. An empirical estimate of $K(\cdot)$ can be obtained as:

$$\hat{K}(h) = \hat{\lambda}^{-1} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{I_{(0,h_{ij})}(h)}{n} = (n \cdot \hat{\lambda})^{-1} \sum_{i=1}^{n} r_i^h,$$

where $\hat{\lambda} = \frac{n}{|\mathcal{D}|}$ is an estimate of the overall intensity, $h_{ij} = ||\boldsymbol{s}_j - \boldsymbol{s}_i|| \; \forall \, i, j$ and r_i^h is the number of events in the observed configuration $\mathcal{S}$ at distance lower than h from $\boldsymbol{s}_i$. Actually, edge correction through weights w_{ij} is necessary in order to get a proper contribute from $\boldsymbol{s}_i$'s close to the boundaries. *On average* this can be done by rescaling every term by the portion of space that lies at distance h from it *within* $\mathcal{D}$. As with $\hat{J}(\cdot)$, also $\hat{K}(\cdot)$ can be compared to some numerical benchmarks, that are nonetheless different according to the dimensionality of the space. For instance, if the point process is over $\mathbb{R}^2$ then: $\hat{K}(h) \approx \pi h^2$ implies CSR; $\hat{K}(h) < \pi h^2$ implies inhibition; $\hat{K}(h) > \pi h^2$ implies clustering.

Given an estimate of $\hat{K}(\cdot)$, there is a simple way to estimate the pair correlation function. Equation (4.1.6) establishes a indirect relationship between $K(\cdot)$ and $g(\cdot) = \lambda_2(\cdot)$, which becomes apparent after applying the derivative at both sides. The empirical counterpart can then be obtained by substituting $\hat{K}'(\cdot)$ for $K'(\cdot)$. Strictly speaking, $\hat{K}(\cdot)$ is a step-function, and therefore non-differentiable. However, it is usually evaluated at a discrete set of equally spaced values and interpolated linearly. Consequently, its derivative $\hat{K}'(\cdot)$ is piece-wise constant and yields a histogram-like estimator for $\lambda_2(\cdot)$:

$$\hat{\lambda}_2(h) = (2\pi h)^{-1} \hat{K}'(h), \quad h > 0. \tag{4.1.7}$$

All the previous summary statistics and benchmarks hold for point processes that deviate from CSR but are stationary in the sense that, marginally, $\lambda(\cdot) = \lambda$. For non-stationary processes moving away from CSR, it is incredibly useful to get also estimates $\hat{\lambda}(\cdot)$ of their first-order intensity $\lambda(\cdot) : \mathcal{D} \to \mathbb{R}^+$. Either because it is of primary interest, or just because it can be used to re-scale (thin) the original process into a stationary version (on which second-order properties can be investigated). Given the general absence of a sufficiently comprehensive parametric expression for it, its estimation is usually attained through non-parametric methods based on the theory of density estimation. After all, as argued in Section 4.1.1, the concept of intensity is representing (under mild assumptions) nothing else but the density of points over the region $\mathcal{D}$.

The most naive estimate is equivalent to an histogram and starts from defining a refined grid (or partition) $\mathcal{A} = \{A_1, \ldots, A_m\}$ of $\mathcal{D}$, where $\cup_{i=1}^m A_i = \mathcal{D}$ and $A_i \cap A_j = \emptyset \; \forall \, i, j$. Then, for each element of the grid, this is built up as:

$$\lambda(\boldsymbol{x}) = \frac{\hat{N}(A_i)}{|A_i|}, \quad \boldsymbol{x} \in A_i, \tag{4.1.8}$$

where $\hat{N}(A_i)$ is the empirical counting measure of the observed configuration. This yields a discontinuous function (step function in $\mathbb{R}$ and step surface in $\mathbb{R}^2$) resembling an histogram, whose volume is proportional to the number of points in the pattern.

The continuous (and slightly more sophisticated) alternative attains a smooth version of (4.1.8) using a kernel function: *Kernel density estimation* (Silverman, 1986). With very similar spirit, this estimate takes the following form:

$$\hat{\lambda}_b(\boldsymbol{x}) = \sum_{i=1}^{n} k_b(||\boldsymbol{x} - \boldsymbol{s}_i||/b), \quad \boldsymbol{x} \in \mathcal{D},$$

where $k_b(\cdot)$ is the so-called *kernel function* and b is a scaling factor named *bandwidth*. The latter regulates the freedom/smoothness of the resulting approximation. On bounded domains, edge effects can be corrected (*on average*) by re-scaling the kernel so that it integrates to 1 on it:

$$\widetilde{k}_b(h_i/b) = \frac{k_b(h_i/b)}{\int_{\mathcal{H}_\mathcal{D}} k_b(h_i/b) dh_i}, \tag{4.1.9}$$

where $h_i = ||\boldsymbol{x} - \boldsymbol{s}_i||$. Let us specify that the *Kernel function* generally is a decreasing and symmetric function, which in this context is used to quantifies the influence of point $\boldsymbol{s}_i$ on the intensity at location $\boldsymbol{x}$. A lot of alternatives have been proposed in the literature. We here list some very common in 1-dimensional settings: *rectangular* (or Parzen), *triangular*, *quadratic* (or Epanechnkikov) and *Gaussian*. The last one is probably the most common, and presents the following form:

$$k_b(h) = \frac{1}{2\pi b} \exp\left\{-\frac{1}{2b} h^2\right\}. \tag{4.1.10}$$

Higher-dimension kernels can be constructed from lower-dimensional ones by composition, namely taking the product of the two. In the case of $d-$dimensional composition of a Gaussian kernel, the result is a Gaussian multivariate kernel with independent components:

$$k_b(\boldsymbol{h}) = \frac{1}{2\pi \prod_{i=1}^{d} b_i} \exp\left\{-\frac{1}{2} \boldsymbol{h}^\top \boldsymbol{B} \boldsymbol{h}\right\}, \tag{4.1.11}$$

where b_i is the bandwidth of the i-th dimension and $\boldsymbol{B} = \text{diag}(b_1, \dots, b_d)$. Multi-dimensional kernels obtained through *composition* are called *separable*.

These last considerations conclude the brief introduction to some of most widely known tools to analyze point patterns. Other important metrics are the *Average Nearest Neighbor* (ANN), the *L-function* derived from $K(\cdot)$, the *Scan Statistics* based on the Kulldorf's distance (Kulldorff, 1999), and others. For a lengthier exposition the reader is referred to Last and Brandt (1995) and Diggle (2013) for temporal and spatial point patterns respectively.

4.2 The Poisson process

The list of notable point processes originating from different probabilistic assumptions on their stochastic behavior is incredibly vast. Providing a comprehensive description of all of them and their properties would be an interminable and complicated task which, by the way, is out of this work main scope. Instead, this section provides a brief account of some of the most famous examples of point processes, some of which present properties relevant to the analysis of the Hawkes processes (more extensively discussed in Section 4.3).

The *Poisson process* is the cornerstone upon which many other processes are built upon. Its most basic version is known as the *Homogeneous Poisson* (**HP**) process with constant intensity λ. In applications it is used as an idealized standard for CSR. It is conveniently defined through the following three postulates.

HP1 For some $\lambda > 0$ and any bounded region $A \subset \mathcal{D}$, the counting measure satisfies:

$$N(A) \sim \mathcal{P}oi(\lambda |A|),$$

where $|\cdot|$ denotes the d-dimensional size (area, volume,...) of the argument.

HP2 Given $N(A) = l$, the l events are uniformly distributed over A.

HP3 For any two disjoint regions $A, B : A \cap B = \emptyset$, we have that the random variables $N(A)$ and $N(B)$ are independent.

It can be easily proved that these three properties are self-consistent with each other, and the reader is referred to Diggle (2013) or to Daley and Vere-Jones (2007) for the formal proof. Derivation of the components of the Janossy density (hence of the likelihood) over $\mathcal{D}$ for the homogeneous Poisson process is trivial. Indeed, from HP1 and HP2 follows that:

$$p(l|\mathcal{D}) = \frac{(\lambda \cdot |\mathcal{D}|)^l}{l!} \cdot e^{-\lambda|\mathcal{D}|} \quad \text{and} \quad f_l(\boldsymbol{x}_1, \dots, \boldsymbol{x}_l|\mathcal{D}) = |\mathcal{D}|^{-l},$$

and then:

$$j_l(\boldsymbol{x}_1, \dots, \boldsymbol{x}_l\,|\mathcal{D}) = \mathcal{L}_\lambda(\boldsymbol{x}_1, \dots, \boldsymbol{x}_l|\mathcal{D}) = \lambda^l \cdot e^{-\lambda|\mathcal{D}|}, \tag{4.2.1}$$

where $\mathcal{L}_\lambda(\cdot|\mathcal{D})$ denotes the likelihood of λ given the observed pattern on $\mathcal{D}$.

As a matter of fact, while rarely useful to describe any interesting process happening in practice, the **HP** process is very useful as a CSR benchmark. Furthermore, one may construct a wide range of point process models by specifying their probability density with respect to a homogeneous Poisson process. This can be done by *conditioning*, *superimposition*, *thinning*, etc.

The most trivial example is the *Inhomogeneous Poisson* (**IP**) process, for which the stationarity assumption is dropped and the constant intensity term is replaced with an arbitrary intensity function $\lambda(\cdot) : \mathcal{D} \to \mathbb{R}^+$. It retains the most important properties of the Poisson process, with some slight modification.

IP1 For any bounded region $A \subset \mathcal{D}$ the counting measure satisfies:

$$N(A) \sim \mathcal{P}oi\left(\lambda(A)\right), \tag{4.2.2}$$

where $\lambda(A) = \int_A \lambda(\boldsymbol{x})d\boldsymbol{x}$

IP2 Given $N(A) = l$, the l events are distributed over A according to density induced by $\lambda(\cdot)$.

IP3 For any two disjoint regions $A, B : A \cap B = \emptyset$, we have that the random variables $N(A)$ and $N(B)$ are independent.

Following *IP1* and *IP2*, the likelihood can be obtained according to similar considerations to the homogeneous case. First, the number of events over $\mathcal{D}$ can be assigned probabilities as:

$$p(l) = \lambda(\mathcal{D})^l \cdot e^{-\lambda(\mathcal{D})}.$$

Then the realized locations are modeled conditionally on $N(\mathcal{D}) = l$ by rescaling the local intensity by the intensity of whole area:

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l | N(\mathcal{D}) = l) = \prod_{i=1}^{l} \frac{\lambda(\boldsymbol{x}_i)}{\lambda(\mathcal{D})}.$$

This yields the following joint density:

$$\mathcal{L}_{\lambda(\cdot)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l) = \lambda(\mathcal{D})^l \cdot e^{-\lambda(\mathcal{D})} \cdot \prod_{i=1}^{l} \frac{\lambda(\boldsymbol{x}_i)}{\lambda(\mathcal{D})} = e^{-\lambda(\mathcal{D})} \cdot \prod_{i=1}^{l} \lambda(\boldsymbol{x}_i)^l, \tag{4.2.3}$$

which is a function of the whole surface and not just of the observed point pattern. Loosely speaking, the first term explains the observed pattern, while the second accounts for the chance of observing other points over the whole region $\mathcal{D}$. The computation of the likelihood can be useful to compare the inhomogeneous model with other alternatives or, if a parametric model for $\lambda(\boldsymbol{x}) = \lambda_\theta(\boldsymbol{x})$ is proposed, to perform inference on the set of parameters θ.

4.2.1 General Cox Processes

The *Inhomogeneous Poisson* process denies the first-order stationarity assumption of the process. It can produce apparent clusters in regions with relatively high intensity, but still retaining independence at all levels. A trivial extension that allows to include dependence to some extent is the *Cox process* (**CP**, Cox (1955)). Technically, this defines a class of *doubly stochastic* processes, in which the intensity function varies stochastically over the region $\mathcal{D}$. Its validity can be proved through a conditioning argument, for which the reader is referred to Chapter 6 of Daley and Vere-Jones (2007).

It is defined by the following properties.

CP1 $\Lambda(x) : \mathcal{D} \to \mathbb{R}^+$ is a non-negative valued stochastic process.

CP2 Conditionally on $\Lambda(x) = \lambda(x)$, the corresponding point process is an inhomogeneous Poisson process with intensity function $\lambda(x)$.

The likelihood can be computed analogously to the (4.2.3), with the relevant complication that $\Lambda(\mathcal{D})$ is not anymore a standard integral but a stochastic integral (McKean, 1969; Kuo, 2006). These are never available explicitly and require integral approximation, which must be ensured to converge (as a random variable) to the actual integral.

For doubly stochastic processes first and second-order properties are determined in terms of expectations with respect to the random measure governing $\Lambda(\cdot)$. For instance, the process is stationary if:

$$\mathbb{E}\left[\Lambda(\boldsymbol{x})\right] = \lambda,$$

while isotropy is verified for:

$$\lambda_2(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) = \mathbb{E}\left[\Lambda(\boldsymbol{x}) \cdot \Lambda(\boldsymbol{x}+\boldsymbol{h})\right] = \lambda_2(h),$$

with $h = ||\boldsymbol{h}||$.

A great variety of Cox processes can be defined by alternative specifications of the driving process $\Lambda(\cdot)$. The most common and widely used is the *Log-Gaussian Cox process*, where $\log\left(\Lambda(\cdot)\right) \sim \mathcal{GP}\left(\mu, c(\cdot,\cdot)\right)$: the log-scale is necessary in order to guarantee the almost certain non-negativeness of the intensity at all locations.

4.2.2 Poisson cluster processes

Poisson cluster processes were initially introduced by Neyman and Scott (1958). These incorporate an explicit form of spatial clustering and can be adopted to model *aggregate* spatial point patterns. The typical clustering process is known as the *Neyman-Scott* (**NS**) process, obtained as the superimposition of a *parent process* with multiple *offspring processes* stemming from each of the parent events. It can be defined through the following three postulates:

NS1 Parent events form a Poisson process with intensity $\lambda(\cdot)$ (originally $\lambda(\cdot) = \lambda$).

NS2 Each parent produces a random number K of offspring, realized independently and identically for each parent according to a probability distribution $p_k = P(K = k)$, $k = 0, 1, \ldots$.

NS3 The positions of the offspring relative to their parents are independently and identically distributed according to a bivariate density $g_s(\cdot)$.

In practice, this process is obtained by initially generating *parent* events from a standard Poisson process, and afterward, producing a random number of offspring in the parents' neighborhoods. Conventionally, the final pattern consists of the offspring only but this is not strictly necessary for the general process properties to hold. We refer to the most common specification that considers offsprings only in the sequel.

The resulting point process is stationary, with intensity depending on the parent and offspring processes intensities as $\lambda \cdot \mu$, where $\mu = \mathbb{E}\left[K\right]$. Moreover, it is isotropic whenever **NS3** specifies a radially symmetric density $g_s(\cdot)$. Usually, bivariate Gaussian densities or uniform distributions over circles are considered, leading to particular cases known in the literature as the *Modified Thomas* process and the *Matèrn* process (Illian et al., 2008).

The corresponding second-order properties can be identified by deriving the density of the vector difference between the positions of two offsprings from the same parent:

$$g_s^{(2)}(x) = \int g_s(x) g_s(x-y) dx,$$

and the corresponding cumulative distribution function $G_s^{(2)}(\cdot)$. Taking into account the probability distribution of the size of each cluster to which an arbitrary event belongs (i.e. $p_k^*(k) = k \cdot p_k(\cdot)/\mu$), the expected number of related events within distance t of an arbitrary event is $\mathbb{E}\left[K(K-1)\right] G_s^{(2)}(\cdot)/\mu$. Instead, being unrelated events located independently of the original event, their expected number is just $\lambda\pi t^2$. Therefore, the resulting K-function is:

$$K(t) = \pi t^2 + \mathbb{E}\left[K(K-1)\right] \frac{G_s^{(2)}(\cdot)}{\lambda\mu}$$

and, differentiating, the resulting pair correlation function is:

$$\lambda_2(t) = \lambda^2 + \rho\mathbb{E}\left[K(K-1)\right] g_s^{(2)}(\cdot).$$

Many other properties (e.g. quadrat counts, empty space distance and nearest neighbor distributions) can be derived starting from the **NS** postulates. Here, we will not mingle in further details but refer again the reader to Daley and Vere-Jones (2003) and Diggle (2013) for a deeper probabilistic and statistical investigation of its properties. Let us just mention that the Poisson cluster process construction can be easily extended to *multi-generation* processes, in which the offspring is the parent of a next generations, and so on. Generally speaking, this type of construction tends to generate mathematically intractable summaries and it is very complicated to detect simple properties starting from its (apparently simple) defining postulates. However, when the reproduction occurs (also or only) along time according to some well-defined distribution, its study can be tackled in the old and well-researched domain of *branching processes* (renewal processes, immigration-birth processes) (Daley and Vere-Jones, 2003; Inés et al., 2016). In particular, the Hawkes process introduced in the next section is a peculiar kind of temporal (or spatio-temporal) point process that exhibits an events inter-dependence with dynamics analogous to branching (in time) and clustering (in space) process. Being it the primary building block of the application in Chapter 5, its properties and analysis are extensively discussed in the next section.

4.3 Self-excitation and the Hawkes process

Hawkes processes are a very interesting class of point processes named after its creator, Alan G. Hawkes, first published in the seminal paper Hawkes (1971b). Their defining characteristic is the peculiar kind of inter-independence induced by a *self-exciting* mechanism. Each event occurrence increases the probability of observing another event in its *spatial proximity* and *immediate future*. In practice, events have the ability to *trigger* others just after they happened, naturally originating patterns that present a *clustered* behavior. This is particularly useful to model phenomena that tend to exhibit such pattern under the effect of similar physical mechanics, e.g.: earthquakes and after-shocks (Ogata, 1988), crime and retaliations/imitations (Mohler et al., 2011; Zhuang and Mateu, 2019), financial shocks and market reactions (Azizpour et al., 2018; Hawkes, 2018), primary and secondary road accidents (Li et al., 2018; Kalair et al., 2020) etc.

Let us emphasize that the triggering/excitation effect can spread along both time and space, but only events that occurred in the past can affect the process's future, and not vice-versa. The existence of a *past* and a *future* is necessary to define a Hawkes process, and hence it cannot dispense of the temporal dimension.

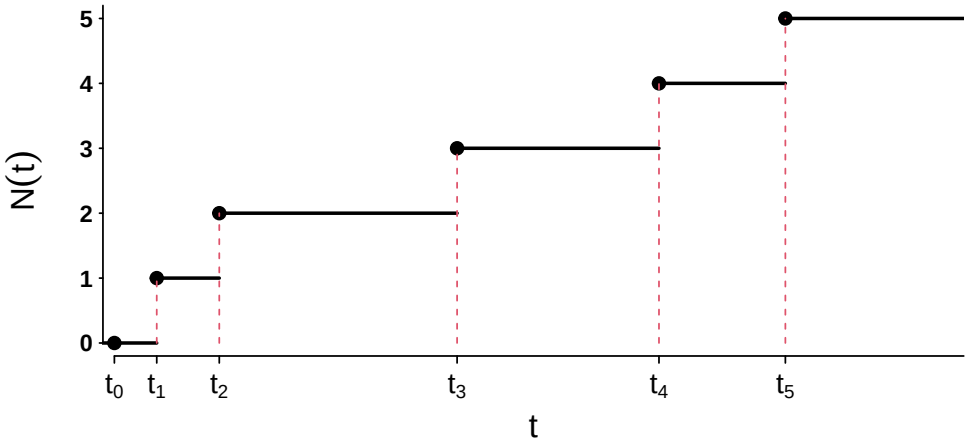


Figure 4.2. Example of temporal point process realization over a bounded window and corresponding counting measure N(t)

In Section 4.3.1 we introduce a building method for temporal point processes based on the so-called *conditional intensity function*[2]. We then define the Hawkes process through the proper specification of this one and provide an interesting analogy as a *branching process* in 4.3.2. Finally, Section 4.3.3 concerns the spatio-temporal extension of the classic Hawkes process and the following two sections (4.3.4, 4.3.5) briefly introduce some of the most common estimation methods.

4.3.1 Temporal point processes and conditional intensity functions

A (simple) temporal point process X is a sequence of random variables $X = \{T_1, T_2, \dots\}$ taking values in $[0, \infty)$ such that $0 < T_k < T_{k+1}\ \forall\, k$. Each T_k is the time of the k-th arrival (event) in the process. Let us recall that any point process can also be represented in terms of the corresponding counting measure. In the temporal case, it is a random-valued function $N(\cdot) : \mathbb{R}^+ \to \mathbb{N}_0$ defined on the half-lines $[0, t]$, $t \in \mathbb{R}^+$ and such that: $N(0) = 0$, $N(t)$ is almost surely finite $\forall\, t$, $N(\cdot)$ is a right-continuous step function with increments of size $+1$. In practice, each $N(t)$ is a random variable counting the number of arrivals T_k happened before t, for all t, and has a one-to-one relationship with the point process itself (see Figure 4.2).

Section 4.1.1 argued how providing a general characterization of the process distribution is a very challenging task that (theoretically) can be tackled from either the point of view of the arrival times T_k or of the counting measure $N(\cdot)$. The latter usually provides a viable option to define some relevant characteristics of the process. When independence between disjoint sets holds (e.g., Poisson process), it can also be used to define all its finite-dimensional-distributions. However, when complex dependence relationships are at play, its specification on general spaces is no longer obvious. Fortunately, when the process takes place along time, an alternative approach can be considered to specify a (finite) point processes distribution uniquely.

Let the domain of the process be a bounded region (interval) $\mathcal{D} = [t_0, T] \subset \mathbb{R}^+$ where, without any loss of generality, the initial point is usually set to $t_0 = 0$. That does not necessarily mean that the process started in $t_0 = 0$, but this implies that any

[2]To not be confused with the *Papangelou conditional intensity*

dependence on events before t_0 shall be null or already incorporated in the density specification. Under these hypotheses and assuming mild regularity of the process (simple and finite), suitable specification of the Janossy densities $j_l(\cdot \,|\, \mathcal{D})$ for any l uniquely determines all finite dimensional distributions on $\mathcal{D}$. Valid specification of such densities is not obvious at all and, in general cases, is a dead-end. However, it is possible exploit that realizations of a temporal point process over $\mathcal{D}$ are *ordered* sets of points $\mathcal{T}_l = \{t_1, \ldots, t_l\} \subset [0, T]$, such that $0 < t_1 < \cdots < t_l < T$, to make the problem more tractable. Indeed, the ordering induced by the temporal dimension allows for an *evolutionary* interpretation of the process, for which *survival functions* $S_k(t|t_1, \ldots, t_{k-1})$ of the k-th event T_k given the previous $k-1$ events can be easily defined:

$$S_k(t \,|\, t_1, \ldots, t_{k-1}) = P(T_k > t \,|\, t_1, \ldots, t_{k-1}),$$

together with the corresponding conditional densities:

$$\begin{aligned} p_k(t \,|\, t_1, \ldots, t_{k-1}) &= -\frac{d}{dt} S_k(t \,|\, t_1, \ldots, t_{k-1}) = \\ &= P\left(T_k \in dt \,|\, t_1, \ldots, t_{k-1}\right). \end{aligned}$$

Using a straightforward conditioning argument, it is possible to express any Janossy density in $[0, T]$ through the following expressions:

$$\begin{aligned} &j_0(\emptyset \,|\, T) = S_1(T), \\ &j_l(t_1, \ldots, t_l \,|\, T) = p_1(t_1) p_2(t_2|t_1) \cdots p_l(t_l|t_1, \ldots t_{l-1}) \cdot \\ &\qquad\qquad \cdot S_{l+1}(T|t_1, \ldots, t_l), \quad l = 1, \ldots. \end{aligned} \tag{4.3.1}$$

Therefore, for any given family of survival functions (and conditional densities), all the finite dimensional distributions of the temporal point process are uniquely identified. In particular, undertaking a slight shift of view, it is possible to define the *hazard function* of the k-th event T_k given the previous $k-1$ events as:

$$h_k(t \,|\, t_1, \ldots, t_{k-1}) = \frac{p_k(t \,|\, t_1, \ldots, t_{k-1})}{S_k(t \,|\, t_1, \ldots, t_{k-1})}, \tag{4.3.2}$$

where, by basic calculus, the following holds:

$$S_k(t \,|\, t_1, \ldots, t_{k-1}) = \exp\left\{ \int_{t_{k-1}}^{t} h_k(\tau \,|\, t_1, \ldots, t_{k-1}) d\tau \right\}.$$

The conditional densities can then be obtained as a function of the hazard functions through manipulation of (4.3.2):

$$\begin{aligned} p_k(t \,|\, t_1, \ldots, t_{k-1}) &= h_k(t \,|\, t_1, \ldots, t_{k-1}) \cdot S_k(t \,|\, t_1, \ldots, t_{k-1}) = \\ &= h_k(t \,|\, t_1, \ldots, t_{k-1}) \cdot \exp\left\{ \int_{t_{k-1}}^{t} h_k(\tau \,|\, t_1, \ldots, t_{k-1}) d\tau \right\}, \end{aligned} \tag{4.3.3}$$

and the Janossy densities of Equation (4.3.1) can finally be expressed in terms of the hazard functions only:

$$j_0(\emptyset \,|\, T) = \exp\left\{ \int_0^T h_1(\tau) d\tau \right\},$$

$$\begin{aligned} j_l(t_1, \ldots, t_l \,|\, T) = h_1(t_1) \prod_{k=2}^{l} h_k(t_k|t_1, \ldots, t_{k-1}) \cdot \\ \exp\left\{ \int_0^{t_1} h_1(\tau) + \sum_{k=2}^{l} \int_{t_{k-1}}^{t_k} h_k(\tau|t_1, \ldots, t_{k-1}) + \int_{t_l}^{T} h_{l+2}(\tau|t_1, \ldots, t_l) \right\}. \end{aligned}$$

The definition of an amalgam of all the hazard functions can remarkably ease the expression of the last formula. It results in the following piece-wise continuous function:

$$\lambda_c(t) = \begin{cases} h_1(t) & t \leq t_1 \\ h_k(t|t_1, \ldots, t_{k-1}) & t_{k-1} < t \leq t_k, \end{cases}$$

which is commonly known as the *conditional intensity function*[3]. Indeed, at each time point t, it has the very interesting interpretation as instantaneous intensity of the process given the observed past $\mathcal{H}(t) = \{t_i : t_i \leq t\}$. It can be proved to be equivalent to:

$$\lambda_c(t) = \lim_{dt \to 0} \frac{\mathbb{E}\left[N(t+dt) - N(t) \,|\mathcal{H}(t)\right]}{dt}, \tag{4.3.4}$$

and, by definition, it has a one-to-one relationship with the family of hazard functions of the process. The Janossy densities can then be expressed as a function of the conditional intensity function:

$$\begin{aligned} j_l(t_1, \ldots, t_l \,|\, T) &= \prod_{k=1}^{l} \lambda_c(t_k) \cdot \exp\left\{ \int_0^T \lambda_c(\tau) d\tau \right\} = \\ &= \prod_{k=1}^{l} \lambda_c(t_k) \cdot \exp\left\{ \Lambda_c(T) \right\}, \end{aligned} \tag{4.3.5}$$

where $\Lambda_c(t) = \int_0^t \lambda_c(\tau) d\tau$ is commonly known as the *compensator* and exists even when the conditional intensity function does not.

All this implies that $\lambda_c(\cdot)$ determines $h_k(\cdot)$ for all k, which in turn determine the Janossy densities for any l and so determine the probability structure of the point process uniquely. For instance, for $\lambda_c(\cdot) = \lambda$ the resulting process is an *Homogeneous Poisson* process (see Equation (4.2.1)); while $\lambda_c(\cdot) = \lambda(\cdot)$ that does not depend on the history of the process identifies the *Inhomogeneous Poisson* process (see Equation (4.2.3)).

For a more formal and detailed discussion of these results, the reader is referred to Chapter 7 of Daley and Vere-Jones (2007). However, to the end of this section, the take-home message is that a temporal point process can be defined uniquely by specifying its conditional intensity function. The latter provides a very flexible tool to define processes through an interpretable causal description with interesting implications for the specification of several temporal dependence structures. This is precisely the scheme through which the Hawkes process is defined.

4.3.2 The Hawkes process

The Hawkes process was introduced in Hawkes (1971b) as a new type of point process with the defining characteristic of exhibiting a *self-exciting* behavior. In the original paper it is described as the (simple) temporal point process induced by the following conditional intensity function:

$$\lambda_c(t) = \mu + \int_0^t g(t-\tau) dN(\tau), \tag{4.3.6}$$

where $\mu > 0$, $g(\cdot) : [0, \infty) \to \mathbb{R}^+$ and $N(\cdot)$ is the counting measure of the process itself. The constant term μ is the *background intensity* of the process, while $g(\cdot)$

[3] To not be confused with the *Papangelou conditional intensity*, introduced in Equation (4.1.5).

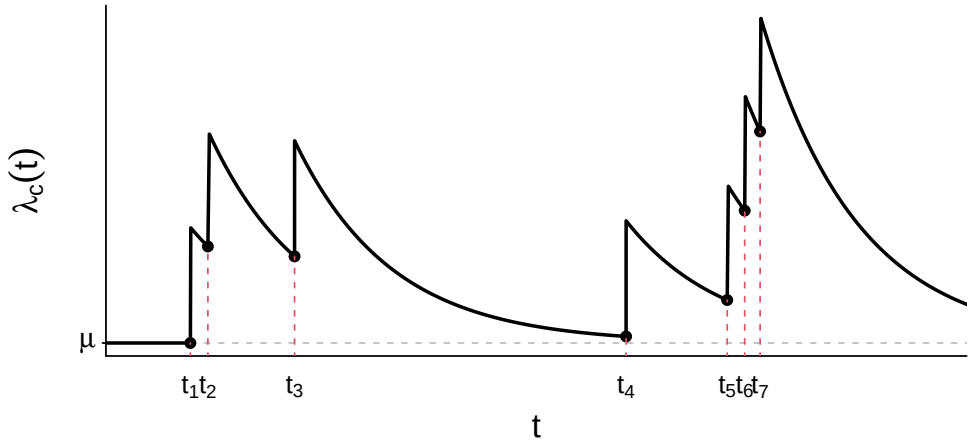


Figure 4.3. Example of Hawkes process realization over a bounded window and corresponding conditional intensity function $\lambda_c(\cdot)$

is generally known as the *excitation/triggering function* and determines the extent of the *self-excitation*. Equation (4.3.6) involves the computation of a stochastic integral on the past history of the point process (from which the self-dependence and the excitation). Therefore, marginally, the conditional intensity function is a random object that depends on the process's random realization. That writing, while consistent with the literature, may obscure the intuition behind it. A clearer expression can be provided conditionally on the observed sequence $\mathcal{T} = \{t_1, \dots, t_l\}$ over the window of interest $[0, T]$. Indeed, for a known past, the value of Equation (4.3.6) in $t \in [0, T]$ is no more a random object, but it is a deterministic function with the following form:

$$\lambda_c(t) = \mu + \sum_{t_i < t} g(t - t_i). \tag{4.3.7}$$

It is now evident how the conditional intensity of the process at time t is determined by its past history $\mathcal{H}_t$, where each previous event t_i contributes to $\lambda_c(t)$ through a function $g(\cdot)$ of the time elapsed since its occurrence. The function $g(\cdot)$ must be a positive function, usually monotone decreasing, that accounts for a sort of *memory effect*: events further back in time shall have smaller effect on the current intensity of the process. Figure 4.3 shows the values assumed by the conditional intensity function corresponding to a Hawkes process realization. After an initial time with constant rate μ, the intensity bumps of $g(0^+)$ just after the first event occurrence, and then slowly decays back to the background intensity as long as other events do not occur. It is completely intuitive how such behavior causes a cluster of events in time (as a consequence of a random event occurrence that increased the current rate), followed by times of apparent calm. Figure 4.3 also highlights how the conditional intensity is defined as left-continuous, coherently with the $t_i < t$ indexing in Equation (4.3.7). That is not as evident from the original expression in (4.3.6), but nonetheless, it is a fundamental property. Otherwise, the intensity in t would also be influenced by what happens in t itself, creating an inner circularity with no escape.

Generally speaking, the structure of $\lambda_c(\cdot)$ is quite flexible and allows for the various specification of both the background intensity μ and the excitation function $g(\cdot)$. In

its most basic version, the background rate μ is just constant and positive. Concerning the excitation function, a very common choice is the exponential parametrized by $\alpha > 0$ (instantaneous increase in the intensity) and $\beta > 0$ (exponential decay):

$$g(\tau) = \alpha \cdot e^{-\beta\tau}, \quad \tau \in \mathbb{R}^+.$$

The exponential decay, besides providing flexible and reasonable modeling for the memory effect, also sensibly simplifies the theoretical derivation of the spectral laws of the process (Hawkes, 1971b). Under this excitation shape, Hawkes (1971a) also extends the definition of the Hawkes process to a larger class of multivariate self and mutually exciting point process. These are not discussed here for the sake of brevity.

Another widespread choice for the excitation is the *power law function*:

$$g(\tau) = \frac{\alpha}{(c+\tau)^p}, \quad \tau \in \mathbb{R}^+,$$

where c, α and p are positive scalars. This power-law form is very well-known in the geological literature as the *Omori's law*, used to predict the rate of aftershocks caused by an earthquake (Ogata, 1999). Its utilization in the context of the Hawkes process is strictly related to the wide adoption of this last as a model for earthquakes aftershock sequences since Ogata (1988).

Many other different semi-parametric and non-parametric versions of excitation functions (piece-wise linear, splines, kernel smoothers) have been proposed in the literature. This includes applications in a great variety of fields, e.g.: finance (Bauwens and Hautsch, 2009; Bacry et al., 2015), neurology (Johnson, 1996), crime (Mohler et al., 2011; Porter et al., 2012), car accidents (Kalair et al., 2020), etc.. . All different alternatives cannot be discussed in detail, but the remainder of this Section will generally refer to an arbitrary function $g(\cdot)$.

Branching process representation Interestingly, Hawkes and Oakes (1974) provides an alternative *branching process* representation of the Hawkes process. This can be very useful as an interpretation tool and as a different point of view that can ease the derivation of some stability properties of the process. In practice, the events resulting from a Hawkes process can be partitioned in two disjoint processes: a *background process* of cluster centers, realizations of a homogeneous Poisson process with the constant rate μ; separate *offspring processes* of triggered events, one for each cluster, whose intensities are determined by $g(\cdot)$. Figure 4.4 shows an illuminating graphical representation of this branching structure.

Each event t_i that enters the system (either it is a background event or it is a triggered event) produces offsprings at future times $t > t_i$ with rate $g(t - t_i)$. Direct offsprings of background events are part of the *first-generation*. In turn, their offsprings comprise the *second-generation* and so on. Members of the union of all these generations are generally called *descendants* of the t_i arrival. In particular, if Z_i is the random number of offspring spawned by the event t_i, the following holds:

$$Z_i \sim \mathcal{P}oi(m_i), \qquad m_i = \int_{t_i}^{T} g(\tau - t_i)d\tau.$$

This would yield $m_i \approx \widetilde{m} = \int_0^\infty g(\tau)d\tau$ if the process continued indefinitely and, in particular, the shape of $g(\tau)$ determines the temporal/spatial distribution of the future offsprings over $[t_i, \infty]$ (after proper normalization over the bounded window). The long term mean $\widetilde{m}$ is known as the *branching ratio* of the process and determines whether its counting measure explodes or not.

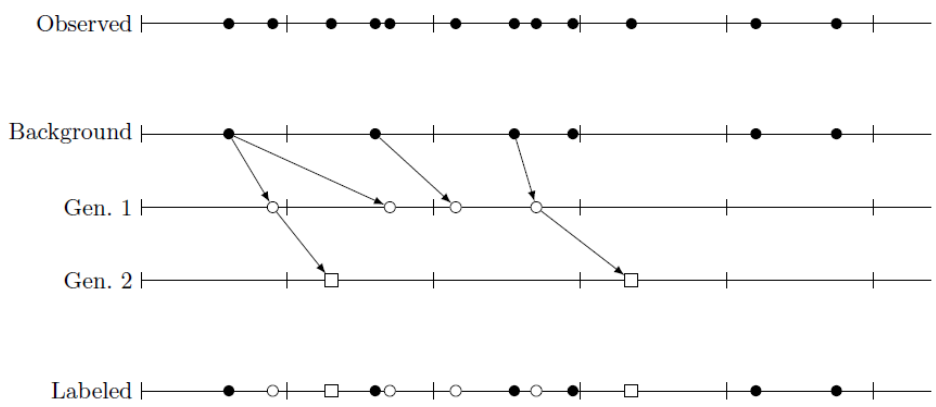


Figure 4.4. Image taken from Reinhart et al. (2018). At the top, a hypothetical observed self-exciting point process of events. Below, the separation of that process into a background process and two generations of offspring processes. Arrows indicate the cluster-trigger relationships; solid circles are background events, and open circles and squares are triggered events. At the bottom, the combined process with generations indicated by shapes and shading.

To see this, let us consider the behavior of the marginal intensity of the process, or rather of the conditional intensity function marginally on all the possible histories. Conditioning on the time of the first jump, by the Hawkes process specification of the conditional intensity function, the following holds:

$$\lambda(t) = \mathbb{E}\left[\lambda_c(t)\right] = \mu + \int_0^t g(t-\tau)\mathbb{E}\left[dN(\tau)\right], \tag{4.3.8}$$

and by definition:

$$\lambda(t) = \mathbb{E}\left[\lambda_c(t)\right] = \frac{\mathbb{E}\left[\mathbb{E}\left[dN(t)|\mathcal{H}(t)\right]\right]}{dt} = \frac{\mathbb{E}\left[dN(t)\right]}{dt}. \tag{4.3.9}$$

From Equation (4.3.8) follows that:

$$\mathbb{E}\left[dN(t)\right] = \lambda(t)dt,$$

and therefore, substituting in Equation (4.3.9):

$$\lambda(t) = \mu + \int_0^t g(t-\tau)\lambda(\tau)d\tau. \tag{4.3.10}$$

Equation (4.3.10) is a *renewal-tipe* equation that corresponds to the convolution $\lambda = \mu + g \circledast \lambda$. It has different solutions according to the value of $m = \int_0^t g(\tau)d\tau$. Grimmet et al. (2020) shows that it is possible to distinguish in three main cases: the *defective* case for $m < 1$; the *proper* case for $m = 1$; the *excessive* case for $m > 1$. For $m \leq 1$ the long term marginal rate (for $t \to \infty$) will settle about a positive and finite value. On the other hand, for the excessive case, the rate grows exponentially quickly and hence the number of events $N(t) \overset{t\to\infty}{\to} \infty$. In particular, the process is stationary if $\lambda(\cdot) = \lambda$, meaning that the process has a long-term mean. Under the stationarity assumption, derivation of the second-order properties and power

spectrum of the process is trivial, and we refer to Hawkes (1971b), Hawkes and Oakes (1974) and Laub et al. (2015) for their discussion.

Generalizations of the Hawkes process to consider non-linear effects in a temporal setting have been recently considered (Brémaud and Massoulié, 1996; Zhu, 2013), but will not be dealt within this work. On the other hand, the next section introduces a *spatio-temporal* extension of the Hawkes process which has been widely used in the study of clustered spatio-temporal patterns. For this last, being a more general version of the solely temporal model and being the model applied in Chapter 5, we will also introduce the likelihood of the corresponding model and the most common estimation methods.

4.3.3 The spatio-temporal Hawkes process

In some contexts, events happening along time occur at different locations in a bounded region $\mathcal{D}$. The spatio-temporal form of the Hawkes process extends the conditional intensity to account for spatial variability of the background rate and spatial propagation of the excitation function. As in Chapter 10 of Diggle (2013), a spatio-temporal point-process can be defined analogously to Equation (4.3.4) with the dependence on the location $\boldsymbol{s} \in \mathcal{D} \in \mathbb{R}^d$ made explicit in the conditional intensity expression:

$$\lambda_c(\boldsymbol{s}, t) = \lim_{d\boldsymbol{s}dt \to 0} \frac{\mathbb{E}\left[N\left(d\boldsymbol{s} \times dt\right) | \mathcal{H}(t)\right]}{d\boldsymbol{s} \cdot dt},$$

where $N(\cdot)$ is a counting measures over sets of $\mathcal{B}\left(\mathcal{Q}\right)$, with $\mathcal{Q} = \mathcal{D} \times [0, T]$. Thus, a *self-exciting spatio-temporal point process* is defined as one with conditional intensity:

$$\begin{aligned} \lambda_c(\boldsymbol{s}, t) &= \mu(\boldsymbol{s}, t) + \int_0^t \int_{\mathcal{D}} g(\boldsymbol{s} - \boldsymbol{x}, t - \tau) dN(d\boldsymbol{x} \times d\tau) = \\ &= \mu(\boldsymbol{s}, t) + \sum_{i:T_i < t} g(\boldsymbol{s} - \boldsymbol{S}_i, t - T_i), \end{aligned} \tag{4.3.11}$$

where $\{\boldsymbol{S}_i, T_i\}_{i=1}^{N(\mathcal{Q})}$ denotes the random set of locations and times of event occurrences. As for the temporal case, the function $g(\cdot, \cdot)$ shall be non-negative and, most likely, decay along both space and time. It is often a kernel or power law decay function, usually assumed to be separable in space and time:

$$g(\boldsymbol{\sigma}, \tau) = g_{\boldsymbol{s}}(\boldsymbol{\sigma}) \cdot g_t(\tau),$$

similarly to some covariance functions in spatio-temporal geo-statistical models (see Section 2.4). When a parametric form for the excitation is hard to establish a-priori, also non-parametric construction techniques have been considered (Marsan and Lengline, 2008; Mohler et al., 2011; Johnson et al., 2018; Zhuang and Mateu, 2019).

It is obvious that also the spatio-temporal version of the Hawkes process can be represented as a Poisson cluster process. The number of offsprings Z_i of each event $(\boldsymbol{s}_i, t_i)$ is:

$$Z_i \sim \mathcal{P}oi(m_i), \qquad m_i = \int_{t_i}^{T} \int_{\mathcal{D}} (\boldsymbol{s}_i) g(\boldsymbol{\sigma} - \boldsymbol{s}_i, \tau - t_i) d\boldsymbol{\sigma} d\tau,$$

where $\mathcal{D}(\boldsymbol{s}) = \left\{\boldsymbol{\sigma} \in \mathbb{R}^2 : \boldsymbol{s} - \boldsymbol{\sigma} \in \mathcal{D}\right\}$. If properly normalized over the whole region $\mathcal{Q}$, the excitation function induces a probability distribution for the location and times of the offspring events.

Likelihood and maximization Let us assume the realization of a self-exciting process with event locations $\mathcal{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$ and times $\mathcal{T} = \{t_1, \dots, t_n\}$ has been observed over the bounded space-time region $\mathcal{Q}$. If a reasonable parametric model for the conditional intensity function is available, then model-based inference (either in the frequentist or Bayesian setting) can be based on the computation and optimization/integration of the likelihood (or log-likelihood). As shown for Poisson processes in Section 4.2, the likelihood of a finite point process model over bounded domains coincides with the corresponding n-dimensional Janossy density. Hence, substituting the conditional intensity function in Equation (4.3.5) with (4.3.3), the result is:

$$\mathcal{L}_{\lambda_c}(\mathcal{S}, \mathcal{T}) = \prod_{i=1}^{n} \lambda_c(\boldsymbol{s}_i, t_i) \cdot \exp\left\{ - \int_0^T \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau \right\},$$

that on the log-scale becomes:

$$l_{\lambda_c}(\mathcal{S}, \mathcal{T}) = \sum_{i=1}^{n} \log\left(\lambda_c(\boldsymbol{s}_i, t_i)\right) - \int_0^T \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau. \tag{4.3.12}$$

This log-likelihood is not easy to evaluate. Indeed, while the first term is just a trivial finite sum of logarithms, the second involves computing the integral of the conditional intensity function over the $\mathcal{Q}$. Integration over bounded domains of various shapes cannot be computed analytically, but an approximation can be achieved through cubatures or other methods over a grid of $m_s \times m_s \times m_t$ points (Meyer et al., 2012). However, the conditional intensity is generally ill-behaved because of the discontinuities arising at each observation time and location (as an effect of the excitation). Therefore, to obtain a good approximation of the integral, the use of a fine grid is highly recommended. At the same time, that is unfortunate because the log-likelihood must be evaluated multiple times inside numerical optimization routines and the computational complexity steeply increases with the grid resolution. Optimization can easily become overwhelming in terms of computing time, even for moderate data-sizes.

In order to ease the approximation of the integral, Schoenberg (2013) suggests considering a large enough *buffer* zone so that the excitation would be mostly contained in the area. With only a little margin of error, the bounded integral over $\mathcal{Q}$ could be replaced with the unbounded one over $\mathbb{R}^2 \times \mathbb{R}^+$, which in some cases presents an analytical solution (e.g., Gaussian kernels). However, this can induce a significant bias in the estimation procedure and Lippiello et al. (2014) observes that the bias is more pronounced when applying the unbounded approximation in the temporal dimension. Since the spatial excitation shall decrease radially around each event, he proposes shifting to polar coordinates for its expression in order to improve the approximation's accuracy while considering not too fine grids for efficiency. This strategy however attains very little computational gain.

On top of all that, Veen and Schoenberg (2008) shows that the resulting log-likelihood, even if well-approximated, can be nearly flat for different forms of the conditional intensity function. That troubles numerical maximization algorithms and makes convergence slow or even to fail altogether in some instances. These issues also affect proper Bayesian estimation through *Markov Chain Monte Carlo* methods. Indeed, even if some alternatives based on a *Metropolis-Within-Gibbs* update have been proposed in Rasmussen (2013) and Loeffler and Flaxman (2018), as the data-size and the need for accuracy increases also these methods become soon impractical. The poor identifiability of the model is the cause of poor mixing of

the chains, which reach convergence only after suitable tuning and many iterations (i.e., of likelihood evaluations) or do not reach convergence at all (Ross, 2016). Simultaneously, it also often happens that different starting points lead to different stationary regions, jeopardizing any inferential conclusion on the true underlying process intensity.

Even if this looks like a helpless situation, a viable solution is available in the *cluster-process* representation of the self-exciting process. Indeed, if labels indicating whether any event is background or triggered were available, then the form of the corresponding log-likelihood would be less ill-conditioned and identified (Veen and Schoenberg, 2008). Such a label is never available in practice, but marginal estimation can be performed in the augmented space, including latent labels. In a frequentist setting, this can be tackled in the same spirit of the *Expectation-Maximization* algorithm for the mixture of distributions (Dempster et al., 1977), while in a Bayesian setting through a sequential MCMC update of latent quantities and parameters (Ross, 2016). This approach is based on the so-called *Complete Data Likelihood*, and more details about this latent variable representation are discussed in Section 4.3.4.

Simulations In order to test estimation algorithms to verify the coherence of the model with the observed dataset, it is generally valuable to be able to simulate fake data from the model itself.

A first ad-hoc simulation method for self-exciting point processes has been proposed in Ogata (1998). Basically, simulation over the spatio-temporal region $\mathcal{Q}$ is performed in two steps. The conditional intensity is integrated along the spatial dimension $\boldsymbol{s} \in \mathcal{D}$ and decomposed in the following product:

$$\lambda_c(\boldsymbol{s}, t) = \lambda_c(\boldsymbol{s}|t) \cdot \lambda_c(t),$$

where $\lambda_c(t) = \int_{\mathcal{D}} \lambda_c(\boldsymbol{s}, t) d\boldsymbol{s}$. This decomposition is usually straightforward under the space-time separability assumption. Even if both components' evaluation can be computationally intensive. All the event times are generated in advance according to the temporal marginal conditional model, while locations are simulated afterward conditionally on the simulated times. The latter encompasses the simulation from an inhomogeneous Poisson process in $[0, T]$ first, and distributing all the generated events according to $\lambda_c(\cdot|t)$ across $\mathcal{D}$ after. The actual procedure is a bit more convoluted, and all technicalities are detailed in Ogata (1988).

Later, Zhuang et al. (2004) proposes an alternative and more efficient simulation strategy based on the branching-representation of the process. First, background events are generated as an inhomogeneous Poisson process governed by the background intensity $\mu(\boldsymbol{s}, t)$ only. This yields a background catalogue of events $\mathcal{G}_0 = \{(\boldsymbol{s}_i, t_i)\}_{i=1}^{N_0}$. After that, a first generation offsprings catalogue $\widetilde{\mathcal{G}}_1 = \left\{(\widetilde{\boldsymbol{s}}_i, \widetilde{t}_i)\right\}_{i=1}^{\widetilde{N}_1}$ is created by simulating an inhomogeneous Poisson processes governed by the excitation function $g(\boldsymbol{\sigma}, \tau)$ for each event in $\mathcal{G}_0$. Only events of $\widetilde{\mathcal{G}}_1$ that are in $\mathcal{Q}$ are retained and the two catalogues are then joined into a common catalogue $\mathcal{G}_1 = \mathcal{G}_0 \cup \left(\widetilde{\mathcal{G}}_1 \cap \mathcal{Q}\right)$.

The same procedure is repeated for all the events in $\widetilde{\mathcal{G}}_1$ to produce the second-generation catalog $\widetilde{\mathcal{G}}_2$, which will be joined to $\mathcal{G}_1$ after the intersection with $\mathcal{Q}$. The procedure is iterated up to generation l, where the l-th generation catalog $\widetilde{\mathcal{G}}_l$ is the first one that does not contain any event in $\mathcal{Q}$.

Edge effects One crucial warning about the model estimation through the log-likelihood and the simulation procedures is that, even if the observed region is a bounded domain $\mathcal{Q}$, the underlying process may extend outside it. In these cases, the estimate of the conditional intensity function will be biased by boundary effects in similar ways to the *kernel density estimation* context (Silverman, 1986; Cowling and Hall, 1996). This issue is exacerbated when a self-exciting mechanism is taking place (Zhuang et al., 2004). Indeed, unobserved events just outside $\mathcal{Q}$ can produce observed offspring in the observed region that cannot but be erroneously attributed to the background from the observer point of view. That can sensibly bias upward the estimate of the background intensity at the edges. At the same time, events next to the boundaries will produce unobserved offsprings outside of $\mathcal{Q}$, which can sensibly bias downward the intensity and range of the excitation effect. That also evidently affects the comparability between simulated and real data because simulation will not consider the effect of points out of $\mathcal{Q}$.

Such biases can be significantly reduced by always considering the larger region $\mathcal{V} \supset \mathcal{Q}$ and using the area $\mathcal{V} \setminus \mathcal{Q}$ as a buffering zone. In the estimation step, those events will be considered as potential triggers or offspring of events in $\mathcal{Q}$, but will not be included in the likelihood evaluation. In the simulation, while simulating the process over the whole $\mathcal{V}$, only events over $\mathcal{Q}$ will be retained (Zhuang et al., 2004).

Marked process One of the original applications of the spatio-temporal Hawkes process was in earthquakes after-shock sequence modeling (Ogata, 1988). In that context, also the magnitude of each earthquake $\mathcal{K} = \{k_i\}_{i=1}^n$ is recorded as an additional observed feature. Generally (and reasonably), it is assumed to be an essential part of the process: the number and distribution of after-shocks (offsprings) may depend on it. Therefore, the need to include the magnitude in the model specification arises. That may also hold in other application fields, such as criminal activity or road accidents (the more serious the parent and the greater is the triggering ability). The effect of an additional observed variable can be trivially included as in the standard modeling framework of *marked spatio-temporal* point process (Chapter 6 of Daley and Vere-Jones (2007)). The *mark* distribution is usually specified at a higher level of the hierarchy, conditionally on the *ground process* conditional intensity, as follows:

$$\lambda_c^*(\boldsymbol{s}, t, k) = \lambda_c(\boldsymbol{s}, t) \cdot f_c(k|\boldsymbol{s}, t),$$

where $f_c(\cdot|\boldsymbol{s}, t)$ is the conditional density of the mark at time t and location $\boldsymbol{s}$ given the history of the process up to t. $\lambda_c(\cdot, \cdot)$ depends on $\mathcal{H}(t)$, which can include both times, locations and marks of previous events according to our (free at will) specification of the excitation function. Nevertheless, when this conditional writing holds and as long as marks are fixed and known variables, their presence does not change much in the discussion of general properties or estimation methods. Indeed, following Section 7.3 of Daley and Vere-Jones (2007), the log-likelihood would just include an additional term:

$$l_{\lambda_c^*}(\mathcal{S}, \mathcal{T}, \mathcal{K}) = \sum_{i=1}^n \log\left(\lambda_c(\boldsymbol{s}_i, t_i)\right) + \sum_{i=1}^n \log\left(f_c(k_i|\boldsymbol{s}_i, t_i)\right) - \int_0^T \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau.$$

For simplicity of notation, we will always refer to un-marked processes from here on.

4.3.4 The complete-data likelihood

The complete data-likelihood (i.e. including background/triggered labels) associated to the cluster representation of a self-exciting process has been first derived by

Veen and Schoenberg (2008).

Let us recall that we observed a realization of a Hawkes process at locations $\mathcal{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$ and time $\mathcal{T} = \{t_1, \dots, t_n\}$ over $\mathcal{Q}$. Let us now denote with $\mathcal{U} = \{u_i\}_{i=1}^n$ the latent label indicating whether event $(\boldsymbol{s}_i, t_i)$ is part of the background $(u_i = 0)$ or was triggered by a previous event $(\boldsymbol{s}_j, t_j)$ with $t_j < t_i$ $(u_i = j)$. Following the cluster representation discussed above, if $u_i = 0$, then the i-th event is a cluster center; otherwise, it is the descendent of a cluster center (direct or indirect offspring). Conditionally on the branching structure, the log-likelihood of Equation (4.3.12) can be expressed as:

$$l_{\lambda_c}(\mathcal{S}, \mathcal{T}|\mathcal{U}) = \sum_{i=1}^{n} I(u_i = 0) \log\left(\mu(\boldsymbol{s}_i, t_i)\right) + \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} I(u_i = j) \log\left(g(\boldsymbol{s}_i - \boldsymbol{s}_j, t_i - t_j) - \int_0^T \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau,$$

where $I(\cdot)$ is the indicator function of its argument. The branching structure dramatically simplifies the log-likelihood as each event's intensity comes only from its trigger or from the background. This does not avoid the computation of the integral of the conditional intensity $\lambda_c(\cdot)$, but favors the identification of background and triggering components easing the convergence of optimization and integration algorithms.

As pointed out above, this approach is not really viable as the labels u_i are never known. On the other hand, if both the conditional intensity components were known, it would be possible to obtain the background and triggering probabilities as:

$$\rho_{ij} = \mathbb{E}\left[u_i = j\right] = P(u_i = j) = \begin{cases} \frac{g(\boldsymbol{s}_i - \boldsymbol{s}_j, t_i - t_j)}{\lambda_c(s_i, t_i)} & t_j < t_i \\ 0 & t_j \geq t_i \end{cases}, \tag{4.3.13}$$

$$\phi_i = \mathbb{E}\left[u_i = 0\right] = P(u_i = 0) = 1 - \sum_{j : t_j < t_i} \rho_{ij} = \frac{\mu(\boldsymbol{s}_i, t_i)}{\lambda_c(s_i, t_i)}. \tag{4.3.14}$$

A Bayesian approach would sequentially sample from the posterior distribution of $\lambda_c(\cdot)$ given $\mathcal{U}$, and from the labels distribution given the conditional intensity function. The procedure would go on until stationarity and a satisfactory chain mixing is achieved. Suitable choice of the priors can favor efficient update of both terms (Ross, 2016).

In a frequentist setting, the probabilities of Equations (4.3.13) and (4.3.14) can instead be directly inserted into the complete-data log-likelihood and yield the so-called *expected* complete-data log-likelihood:

$$\mathbb{E}_{\mathcal{U}}\left[l_{\lambda_c}(\mathcal{S}, \mathcal{T}|\mathcal{U})\right] = \sum_{i=1}^{n} \phi_i \log\left(\mu(\boldsymbol{s}_i, t_i)\right) + \\ + \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij} \log\left(g(\boldsymbol{s}_i - \boldsymbol{s}_j, t_i - t_j) - \int_0^T \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau. \tag{4.3.15}$$

This, as its non-stochastic counterpart, is much easier to numerically maximize with respect to the shape of the conditional intensity function $\lambda_c(\cdot)$. A standard EM algorithm implementation would then alternate between maximizing (4.3.15) and updating the probabilities in (4.3.13) and (4.3.14) until some convergence criterion (e.g. increase in the log-likelihood) is met.

4.3.5 Stochastic declustering and reconstruction

Direct implementation of likelihood-based approaches is feasible whenever all the conditional intensity function components can be assigned a suitable parametric form. However, in most point pattern applications, expressing the intensity surface as a function of some parameters is far too limiting. Indeed, it is way more commonly fitted non-parametrically from the observed data through methods such as *kernel density* estimation (Silverman, 1986).

In the context of self-exciting processes, the same issue immediately arises for the expression of the background intensity $\mu(\cdot)$ (Zhuang et al., 2002). At the same time, in some cases, the researcher deems a non-parametric expression also for the excitation function $g(\cdot, \cdot)$. However, to build a kernel-density estimator of the two components, disentangling the background and the excitation contribution to the overall intensity at all locations become necessary. The first solution proposed in the literature is *stochastic declustering* (Zhuang et al., 2002), shortly followed by the *stochastic reconstruction* (Zhuang et al., 2004).

Stochastic declustering Stochastic declustering has been initially introduced in Zhuang et al. (2002) in the context of the *Epidemic-Type Aftershock Sequences* (ETAS) models. In this original application the authors consider fitting a non parametric background intensity (assumed to be constant across time), while keeping a parametric expression for the excitation function:

$$\lambda_c(\boldsymbol{s}, t|\theta) = \mu(\boldsymbol{s}) + \sum_{t_i<t} g_\theta(\boldsymbol{s} - \boldsymbol{s}_i, t - t_i). \tag{4.3.16}$$

Identification is eased by assuming space-time separability on the excitation function: $g_\theta(\boldsymbol{s} - \boldsymbol{s}_i, t - t_i) = g^t_{\theta_t}(t - t_i) \cdot g^s_{\theta_s}(\boldsymbol{s} - \boldsymbol{s}_i), \theta = \{\theta_t, \theta_s\}$.

The underlying idea of *stochastic declustering* starts from considering the *total spatial intensity function*:

$$m(\boldsymbol{s}) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \lambda_c(\boldsymbol{s}, \tau|\theta) d\tau, \tag{4.3.17}$$

where T is the length of the observation period. Since in practice only a finite window $[0, T]$ is ever observed, the limit can be ignored. The total intensity is a function of all the events, background and not. Hence, estimation of it does not require declustering and, assuming stationarity of the background over time, can be achieved through the following kernel approximation:

$$\hat{m}(\boldsymbol{s}) = \frac{1}{T} \left(\sum_{i=1}^{n} \widetilde{k}_b(\boldsymbol{s} - \boldsymbol{s}_i/b) \right),$$

where $\widetilde{k}_b(\cdot)$ is a suitably scaled kernel with bandwidth b as it is introduced in Equation (4.1.9). In particular, by substituting (4.3.16) in (4.3.17), the integral is approximately equal to:

$$\begin{aligned} m(\boldsymbol{s}) &\approx \frac{1}{T} \int_0^T \mu(s) + \sum_{t_i<t} g_\theta(\boldsymbol{s} - \boldsymbol{s}_i, \tau - t_i) d\tau = \\ &= \mu(\boldsymbol{s}) + \frac{1}{T} \sum_{i=1}^{n} g^s_\theta(\boldsymbol{s} - \boldsymbol{s}_i) \int_{t_i}^{T} g^t_\theta(\tau - t_i) d\tau, \end{aligned}$$

and therefore:

$$\begin{aligned}\mu(\boldsymbol{s}) &\approx m(\boldsymbol{s}) - \frac{1}{T}\sum_{i=1}^{n} g_\theta^s(\boldsymbol{s}-\boldsymbol{s}_i)\int_{t_i}^{T} g_\theta^t(\tau - t_i)d\tau = \\ &= m(\boldsymbol{s}) - \frac{1}{T}\gamma(\boldsymbol{s}-\boldsymbol{s}_i),\end{aligned} \tag{4.3.18}$$

where $\gamma(\cdot)$ represents the contribution of triggered events to the total spatial intensity function: the *cluster process intensity*. If the labels u_i indicating what events are triggered and what are background were known, $\gamma(\cdot)$ can be estimated by using only the *triggers*. Or, if at least the probabilities of Equation (4.3.14) and (4.3.13) were known, a weighted kernel estimate of $\gamma(\cdot)$ can be set-up as:

$$\begin{aligned}\hat{\gamma}(\boldsymbol{s}) &= \frac{1}{T}\sum_{i=1}^{n} P(u_i \neq 0)\cdot \widetilde{k}_b(\boldsymbol{s}-\boldsymbol{s}_i/b) = \\ &= \frac{1}{T}\sum_{i=1}^{n}(1-\phi_i)\cdot \widetilde{k}_b(\boldsymbol{s}-\boldsymbol{s}_i/b).\end{aligned}$$

This leads to the following kernel weighted estimator of the background intensity:

$$\begin{aligned}\hat{\mu}(\boldsymbol{s}) &= \hat{m}(\boldsymbol{s}) - \hat{\gamma}(\boldsymbol{s}) = \\ &= \frac{1}{T}\sum_{i=1}^{n}\phi_i\cdot \widetilde{k}_b(\boldsymbol{s}-\boldsymbol{s}_i/b),\end{aligned} \tag{4.3.19}$$

as long as the same kernel is used for both the total spatial intensity function and the cluster process intensity.

Equation (4.3.19) thus represents a weighted kernel estimation method for the background that works conditionally on $\{\phi_i\}_{i=1}^n$. However, in order to know these last quantities, both the excitation function and the background intensity knowledge are necessary. There is an inner circularity in this problem that can be solved in the EM algorithm's spirit, as in the pseudo-code here below.

Stochastic Declustering

0: Let $i = 1$ and set $\hat{\mu}_0(\boldsymbol{s}) = 1$ initially.

1: Given $\hat{\mu}_{i-1}(\boldsymbol{s})$, get an estimate θ_i of the parameters of the excitation function (4.3.16) through maximum likelihood.

2: Calculate the background probabilities $\{\phi_i\}_{i=1}^n$ using $\hat{\mu}_{i-1}(\boldsymbol{s})$ and θ_i in Equations (4.3.14) and (4.3.13).

3: Get a new non-parametric estimate of the background $\hat{\mu}_i(\boldsymbol{s})$ using the new background probabilities

4: If $\max_{\boldsymbol{s}} |\hat{\mu}_i(\boldsymbol{s}) - \hat{\mu}_{i-1}(\boldsymbol{s})| < \epsilon$, for $\epsilon < 0$ small, stop. Otherwise, go to step **1**.

The final result, particularly the background intensity estimate, depends on the choice of the bandwidth b for the kernel smoothing procedure. Zhuang et al. (2002) implements an adaptive approach, in which each point has an individual bandwidth selected so that at least n_p are included in the kernel (relevant) range. There is no

proper guidance about an optimal strategy, but the adaptive approach proves to work fine in practice.

Once the conditional intensity function has been estimated, it is possible to build stochastically declustered catalogs of events. Each event is assigned to a tree that connects it to its trigger (background or another event). Each event i is assigned either to the background or the triggering tree of another event j according to the final ϕ_i and $\{\rho_{ij}\}_{j\neq i}$. Given the randomness of the assignment procedure, results are stochastic and can change at every execution. While it may be seen as a drawback, the same authors argue that it is, instead, an advantage. Indeed, it permits quantifying the uncertainty of the declustering.

Stochastic reconstruction The first mention to *stochastic reconstruction* dates back to Zhuang et al. (2004) and was then more clearly formalized in Zhuang (2006). It consists of a very similar method to stochastic declustering in the spirit, but that allows for non-parametric estimation of the excitation function by inspecting the process's second-order properties. The necessity for such a non-parametric estimation arises especially out of the ETAS model context. Indeed, there is more than a century of literature about the triggering abilities of earthquakes that extrapolate valid parametric forms for the excitation function (e.g., *Omori's law*). Unfortunately, the same does not hold for other application fields, in which the introduction of an excitation effect is rather recent.

The first example of application of the stochastic reconstruction is in the context of crime modeling (Marsan and Lengline, 2008), in which the authors assume a constant background intensity and the excitation to be a step-wise constant function. Later, Mohler et al. (2011) generalize the *Marsan-Lengline* implementation by including the possibility to account also for spatial and temporal heterogeneity in the background. This algorithm has been after identified as an EM algorithm in fact by Fox et al. (2016a). We here introduce the main ingredients of the stochastic reconstruction, referring to the self-exciting model considered by Mohler et al. (2011):

$$\lambda_c(\boldsymbol{s}, t|\theta) = \mu(\boldsymbol{s}, t) + \sum_{t_i<t} g(x - x_i, y - y_i, t - t_i), \tag{4.3.20}$$

where x and y of $\boldsymbol{s} = (x, y)$ denote the two dimensions of $\mathcal{D}$, excitation is not necessarily isotropic and space-time separability is assumed both for the background intensity and the excitation function:

$$\mu(\boldsymbol{s}, t) = \mu_s(\boldsymbol{s}) \cdot \mu_t(t), \qquad g(x, y, t) = g(x, y) \cdot g(t).$$

The founding idea of the algorithm is that each event is partially caused by the background, and partially by the triggering effect of previous events. In particular, given an event i, the probabilities ϕ_i and $\{\rho_{ij}\}_{j=1,j\neq i}^{n}$ represent the expected contribution of the background (the first) and of other events' triggers (the second) to its occurrence. This scheme is summarized in Figure 4.5. Therefore, given a set of weights $\{\phi_i\}_{i=1}^{n}$ and $\{\rho_{ij}\}_{i,j=1,j\neq i}^{n}$, background and excitation can be smoothed through two weighted kernels. The background smoother considers points individually, and weighs each point by the background contribution:

$$\hat{\mu}_s(\boldsymbol{s}) = \sum_{i=1}^{n} \phi_i \cdot \widetilde{k}_{b_s}(\boldsymbol{s} - \boldsymbol{s}_i/b_s), \qquad \hat{\mu}_t(t) = \sum_{i=1}^{n} \phi_i \cdot \widetilde{k}_{b_t}(t - t_i/b_t), \tag{4.3.21}$$

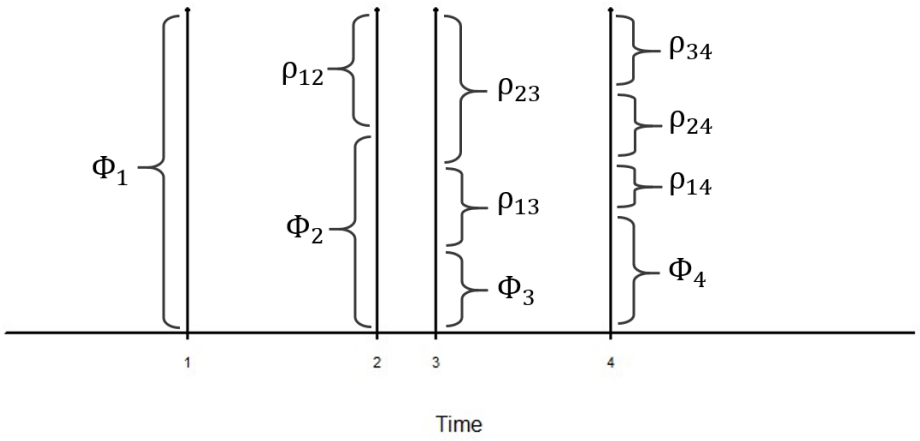


Figure 4.5. Contribution of background and triggering in 4 events from a self-exciting process

while the excitation function smoother considers points in pairs and each pair is weighted according to the probability that the second was triggered by the first:

$$\begin{aligned}\hat{g}_s(x,y) &= \sum_{i=1}^{n}\sum_{t_j<t_i}\rho_{ij}\cdot\widetilde{k}_{h_x,h_y}((x-(x_i-x_j))/h_x,\,(y-(y_i-y_j))/h_y),\\ \hat{g}_t(t) &= \sum_{i=1}^{n}\sum_{t_j<t_i}\rho_{ij}\cdot\widetilde{k}_{h_t}((t-(t_i-t_j))/h_t).\end{aligned} \tag{4.3.22}$$

As usual, all the kernels depend on a specific bandwidth and are scaled to integrate to 1 in the corresponding domain.

As for stochastic declustering, weights are needed to compute background and excitation, but background and excitation are needed to get the weights. The circularity can be solved in the same fashion by setting initial values for the background intensity and excitation function, updating the weights, recomputing the first, and so on. The algorithm is sketched here below.

Stochastic Reconstruction

0: Let $i = 1$ and set $\hat{\mu}_0(\boldsymbol{s}, t)$ and $\hat{g}_0(x, y, t)$ to some initial values (e.g. background equal to one, excitation with exponential decay in all dimensions).

1: Given $\hat{\mu}_{i-1}(\boldsymbol{s})$ and $\hat{g}_{i-1}(x, y, t)$ compute the background and triggering probabilities $\{\phi_i\}_{i=1}^n$, $\{\rho_{ij}\}_{i,j=1,j\neq i}^n$ from Equations (4.3.14) and (4.3.13).

2: Get the updated smoothing estimates of the background and excitation $\hat{\mu}_i(\boldsymbol{s})$, $\hat{g}_0(x, y, t)$ using Equations (4.3.21) and (4.3.22).

3: If convergence is reached (only small variations in the background and excitation), stop. Otherwise, go to step 1.

In this case, the choice of bandwidths plays a key role in determining the final results. However, at the moment, there is no general recommendation about their choice. Either automatic selection or adaptive methods borrowed from the more general

theory of kernel density estimation methods can be considered. According to the peculiarities of the data at hand, the choice is left to the researcher's sensibility.

This algorithm is not devoid of other issues. First of all, when the size of the dataset is large, it can get highly computationally expensive (especially in terms of the excitation function smoothing that scales with $\mathcal{O}(n^2)$). Secondly, the complete non-parametric fit of the two components makes proper identification hard, and different runs starting from different initial guesses can lead to potentially very different results. Last but not least, the stopping rule is vague (to put it mildly).

The more recent work by Zhuang and Mateu (2019) provides an analogous procedure that includes two relaxation coefficients and also includes a likelihood-based stopping rule. It has been successfully implemented for the estimation of a slightly more complicated model. The same model and estimation method has also been adopted in Kalair et al. (2020), and a more detailed description is provided in Chapter 5.

Chapter 5

A Periodic Spatio-Temporal Hawkes Model for road accidents

There is much work focusing on spatial and temporal analysis of traffic accidents, with a discussion of the evidence for spatial auto-correlation in accidents data given in Aguero-Valverde and Jovanis (2008). Here, the authors discuss previously used descriptive analysis methods, including K-function analysis and comparison to complete spatially random patterns, and detail how they showed evidence of spatial correlation among event locations. They furthered this by incorporating spatial components into an auto-regressive model, finding it improved upon models disregarding the spatial correlation in the data. Additional analysis is provided in Fan et al. (2018), where the authors analysed the evolution of event hot-spots through time on an urban network in China. They extended and applied kernel density estimation on networks, Moran's I (Moran, 1950) and Local Indicators of Mobility Associations (LIMA) to offer exploratory analysis of the dataset from multiple perspectives. Further spatial-temporal analysis was completed in Song et al. (2018), where the authors used Kulldorff's space-time scan statistics to determine statistically significant clusters of traffic accidents across the entire UK in 2016. They found two significant clusters, both in the north of the country, but conceded that they did not explicitly account for the network structure in their analysis. Al these works were focused on a descriptive analysis of the data, rather than formulating a model incorporating the observed features. Other tools and techniques for such descriptive spatial analysis have already been briefly introduced in Section 4.1.2, but more details on K-function analysis can be found in Dixon (2014), on Kulldorff's space-time scan statistic in Kulldorff (2001) and on LIMA in Rey (2016).

The approach proposed in this chapter is fundamentally different from the ones discussed above, as it seeks to achieve model-based inference for the clustered point patterns arising from road accidents data. Hence, we model the car accidents' dynamics as a point-process and do not limit the discussion to a descriptive analysis and discovery of locations with statistically significant clusters. After all, there is already some literature on the application of spatio-temporal point-process models to traffic events and road accidents are nothing but a particular kind of the former. For instance, Moradi and Mateu (2019) considers extensions of these models on linear networks for traffic event modeling. Considering road-networks in Houston, Medellin and Eastbourne, the authors investigated what features the network-constrained data showed. They find statically significant evidence that events on the network

did not follow a uniform spatial-temporal Poisson process but indicate an evident clustering of data in space and time. A point of particular interest to the approach we are introducing arises from the modeling of traffic crashes in Acker and Yuan (2019). Here, logistic regression and random forest models are used to predict the likelihood of event occurrences. In-particular, the models incorporates a significant *'cascading effect'* variable, in which the presence of an event showed significant influence on the likelihood of another nearby in space and time. Although this is not inserted in the point-process context, there is clearly a sense that cascading effects are a real component in some traffic data that one may want to incorporate into a model. From the results in Acker and Yuan (2019) is however unclear what time and length scales are associated with the cascading effect, and on what kind of roads this is more or less accentuated.

From this brief review of the literature it is clear that one could ask if point-process models incorporating a self excitation component may capture this apparent cascading effect. *Hawkes Process* based models, introduced in Section 4.3, are natural candidates to this end. As a matter of fact, they have already been successfully applied to a wide range of real-world problems. Historically, they have been initially used as a model for earthquakes' aftershock sequences (Ogata, 1988, 1999; Zhuang et al., 2004; Zhuang, 2011). There are indeed physical justifications and strong statistical evidence that an initial large earthquake leads to aftershocks, allowing for a clearly interpretable 'self-excitation' component at play. For long, the use of Hawkes Processes has been almost completely limited to the geological sector, but there has been a surge in the application of self-exciting point-process models to problems of various nature during the last two decades. For instance, their application has been discussed also in the context of crime modeling and forecasting (Mohler et al., 2011; Mohler, 2014a), where self-excitation can be seen (in practical terms) as retaliation or imitation crimes. They have also been considered in epidemic forecasting (Schoenberg et al., 2019), and modelling events on social-networks (Fox et al., 2016b). Finally, one of the most novel applications concerns the modeling of the COVID-19 spread, although this is still in its infancy (Lesage, 2020).

This surge of new application areas is probably related to the improvement of data-collection systems in different fields. Recording events at a such fine spatial and temporal scale to make point-process modeling feasible is now possible in almost any sector of scientific research, at reasonable costs. Nevertheless, at the author's knowledge, there is little on applying them to traffic events. One paper that does look at this is Li et al. (2018), where a self-exciting point-process model that could theoretically be fit to real data is proposed. However, this proposal is only verified on simulated data, thus giving only a snapshot of its potential fitting performance. Additionally, a somewhat similar idea is considered in Lim et al. (2016), where the authors consider traffic flow data to be 'events' and try to use self-excitation to model the idea that often traffic flow occurs in clusters. The authors use the Hawkes process to model the traffic flow in Sydney, but there is no clear conclusion as to if the model is statistically defensible and captures all features of the data, or if alternative traffic forecasting methods are preferable. There is still an enormous amount of work to be done applying this methodology to real data in order to understand what components of it are important, the actual amount of self-excitation present in traffic data, and the appropriate time and length scales it occurs on. Throughout the applications in Section 5.2 and 5.3, we try to offer some insightful discussion on these questions.

Let us recall that when talking about self-exciting point processes, there is some sense of a 'background' component, that models the typical behaviour, and a 'triggering' component that allows for self-excitation. There is much discussion as to

what functional form the components should take, in particular when referring to the triggering component. We argued in 4.3.2 that in the original applications on earthquakes after-shock sequences, one usually supposes some reasonable parametric forms on the triggering function drawing from the geological literature. Then, it is possible to determine which is most appropriate through inspection of the log-likelihood value or other information criteria. These choices have been initially extended also to other fields and included some form of exponential, Gaussian or power-law decay of triggering in time and space. However, recent work in Zhuang and Mateu (2019) proposes a spatio-temporal Hawkes process model for crime data without assuming any functional form on either the background and excitation components. Exploiting the methods originally introduced in Zhuang (2006) they show one can determine which events in the data appear to be background and which appear to be triggered, and then smooth data based on this to reconstruct the desired components. The general technique has been briefly explained in Section 4.3.5, but Zhuang (2006) considers a slightly more sophisticated model and estimation procedure, which is extensively illustrated in Section 5.1.4. This is particularly useful for novel applications in which there is no guidance in the literature about how the excitation shall decay along space and time, if any. Therefore, it comes useful in our modeling attempt of road accidents occurrences.

5.1 A periodic semi-parametric model

The final objective of our analysis is to describe the dynamic of traffic collisions (events) observed over a spatio-temporal domain $\mathcal{Q} = \mathcal{D} \times [0, T]$, where $\mathcal{D}$ denotes the spatial dimension. In general terms, the dimension and nature of $\mathcal{D}$ is arbitrary and peculiar to the application setting. Let us here consider the use case of $\mathcal{D} \subset \mathbb{R}^2$.

We deem to model the distribution of the number of car-crashes $N(B \times [t_1, t_2])$, where $B \subset \mathcal{D}$ and $[t_1, t_2] \subset [0, T]$. The founding assumption of the proposed approach is that the random mechanism governing the occurrences distribution over $\mathcal{Q}$ is a simple and (locally) finite spatio-temporal point process. Exploiting the order induced by the temporal dimension, following Section 4.3.1, it can be defined through the suitable specification of the corresponding conditional intensity function:

$$\lambda_c(\boldsymbol{s}, t) = \lim_{d\boldsymbol{s}, dt \to 0} \frac{\mathbb{E}\left[N(d\boldsymbol{s}, dt) \,|\mathcal{H}(t)\right]}{d\boldsymbol{s} \cdot dt}, \tag{5.1.1}$$

where $\mathcal{H}(t)$ is the history of the process up to time t.

The introduction to this chapter suggested that point patterns arising from road accidents present an irregular behavior in time and space, usually accompanied by a clustered structure. Hence, the form of Equation (5.1.1) shall take a form which is potentially able to account for all the causes of such inhomogeneity. First, the risk of a road accident occurring at an arbitrary point in space and instant in time is generally not spread uniformly along the observed region $\mathcal{Q}$. Indeed, this is definitely affected by risk factors such as the condition and shape of the roads, the current traffic level, the climatic conditions and other environmental factors that can vary, sometimes much, in the region of interest. Second, we can imagine that the occurrence of a car accident can cause sudden and unexpected slowdowns and traffic events, which in turn may increase the risk and trigger additional accidents in their immediate proximity (in space and time). While the former factor of variability can be easily modeled through the wide-known and applied inhomogeneous Poisson Process or Cox process, the second factor includes a very particular kind of inter-dependence never considered in previous modeling. As a matter of fact, the actual

presence of such a triggering is not yet certified in any study and it is an hypothesis that we would like to verify.

The Spatio-temporal Hawkes process discussed in Section 4.3 provides a very compelling model to include both components: a potentially space-time varying overall intensity that can be instantly modified by the triggering effect of other events. Let us report here below the conditional intensity function (Equation (4.3.11)) in the general spatio-temporal Hawkes process framework:

$$\lambda_c(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) + \int_0^t \int_{\mathcal{D}} g(\boldsymbol{s} - \boldsymbol{\sigma}, t - \tau) dN(d\boldsymbol{\sigma} \times d\tau),$$

where $\mu(\cdot, \cdot)$ is the background intensity and $g(\cdot, \cdot)$ is the excitation function accounting for the effect of all previous events. This work is largely inspired and based on the specification of these two components proposed in Zhuang and Mateu (2019), which is different from the typical version in two main aspects:

- includes a *periodic component* in the background rate;
- introduces *relaxation coefficients* to ease the model identification (and estimation).

Background The background component is a function describing the varying general risk of a car accident occurring at different times and locations. Whilst not always consistent with the reality of things, a very common working assumption is the one of space-time separability. Namely, space and time factors combine in a multiplicative fashion:

$$\mu(\boldsymbol{s}, t) = \mu_s(\boldsymbol{s}) \cdot \mu_t(t), \tag{5.1.2}$$

where $\mu_{\boldsymbol{s}}(\cdot)$ is the spatial background and $\mu_t(\cdot)$ is the temporal background. The spatial background is assumed to be static along time and accounts for different locations having differing rates of accidents. This can happen because of low visibility, higher presence of intersections, generalized higher level of traffic, etc. It practically accounts for the combined effect of unobserved spatial factors, that can vary irregularly (but possibly continuously) over $\mathcal{D}$. The temporal background is instead assumed to present some sort of regularity and is further modulated through three additional components:

$$\mu_t(t) = \mu_d(t) \cdot \mu_w(t) \cdot \mu_{tr}(t), \tag{5.1.3}$$

where $\mu_d(\cdot) = \mu_d(d(\cdot))$ is the *daily variation*, $\mu_w(w(\cdot))$ is the *weekly variation* and $\mu_{tr}(\cdot)$ is the eventual *long-term* trend; $d(t)$ and $w(t)$ represent the *time of the day* and *day of the week* corresponding to instant t, respectively. Daily and weekly seasonality is a ubiquitous feature of traffic data reflecting daily 'rush-hour' commuting patterns, and weekly differences such as between the 5-day working week to the 2-day weekend. Annual seasonality may also be present, but it is not considered in the original or this work since data spans no more than one year, and such variation must then be captured by the trend.

Excitation As for the background, also self-excitation is assumed to be separable in space and time. In particular, as specified in Section 4.3.5, Zhuang and Mateu (2019) considers a potentially anisotropic excitation in space. The excitation function can then be expressed as follows:

$$g(\boldsymbol{s} - \boldsymbol{s}', t - t') = g_s(x - x', y - y') \cdot g_t(t - t'), \tag{5.1.4}$$

where $\boldsymbol{s} = (x, y)$ and $\boldsymbol{s}' = (x', y')$. The potential anisotropy of the spatial excitation is implied by the dependence of the function value on the whole separation vector $\boldsymbol{s} - \boldsymbol{s}' = (x - x', y - y')$ and not just on the Euclidean distance $||\boldsymbol{s} - \boldsymbol{s}'|| = h$.

Relaxation coefficients Let us consider the specifications of the background intensity resulting from Equations (5.1.2) and (5.1.3), and of the excitation function in Equation (5.1.4). This would yield a Hawkes model with the following expression of the conditional intensity function:

$$\begin{aligned}\lambda_c(\boldsymbol{s}, t) = \mu_s(\boldsymbol{s}) \cdot \mu_d(t) \cdot \mu_w(t) \cdot \mu_{tr}(t) + \\ + \int_0^t \int_{\mathcal{D}} g_s(s_1 - u, s_2 - v) \cdot g_t(t - \tau) N(du \times dv \times d\tau),\end{aligned} \tag{5.1.5}$$

where $\boldsymbol{s} = (s_1, s_2)$. Pursuing a completely non-parametric estimation of such a complex model may suffer of a number of numerical and statistical complications. For instance, the multiplicative structure of the two components makes the single elements unidentifiable up to a multiplicative constant, jeopardizing the reliability and interpretability of the final estimates. The idea of Zhuang and Mateu (2019) to overcome this issue is to introduce two additional parameters $\mu_0 > 0$ and $A > 0$, named in the original paper *relaxation coefficients*. The first represents the *background overall level*, while the second the *excitation overall level*. Estimation of the single elements in (5.1.5) is then stabilized by forcing background and excitation elements to represent only relative variations with respect to these two general values: background elements are normalized to have average equal to 1, while excitation components are forced to integrate to 1. After these adjustments, the final expression of the model conditional intensity function becomes:

$$\begin{aligned}\lambda_c(\boldsymbol{s}, t) = \mu_0 \cdot \widetilde{\mu}_s(\boldsymbol{s}) \cdot \widetilde{\mu}_d(t) \cdot \widetilde{\mu}_w(t) \cdot \widetilde{\mu}_t(t) + \\ + A \cdot \int_0^t \int_{\mathcal{D}} \widetilde{g}_s(s_1 - u, s_2 - v) \cdot \widetilde{g}_t(t - \tau) N(du \times dv \times d\tau),\end{aligned} \tag{5.1.6}$$

where $\widetilde{\mu}_\cdot(\cdot)$ and $\widetilde{g}_\cdot(\cdot)$ represent the normalized versions of background and excitation elements.

Now, let us assume a point pattern with n points $\{(\boldsymbol{s}_i, t_i)\}_{i=1}^n$ over the region $\mathcal{Q}$ had been observed. We would like to pursue the estimation of every component of (5.1.6) in a data-driven way, taking advantage of the *stochastic reconstruction* algorithm. However, neither the relaxation coefficients and the periodic components of the background rate in the considered model formulation can be directly estimated by using the standard method introduced in Section 4.3.5. But, considering some little modifications based on the residual analysis developed in Zhuang (2006), it can be generalized to deal with the estimation problem at hand. The key point of the residual analysis methodology for our purposes is that the conditional intensity of a point process has the following property.

Property 5.1.1. *For any counting process $N(\cdot)$ equipped with an arbitrary conditional intensity function $\lambda_c(\cdot, \cdot)$ and a predictable process $f(t, x) \geq 0$ over a time domain $[T_1, T_2]$ and space domain S, we can write:*

$$\mathbb{E}\left[\int_{[T_1,T_2]\times S} f(\boldsymbol{\sigma}, \tau) dN(d\tau \times d\boldsymbol{\sigma})\right] = \mathbb{E}\left[\int_{T_1}^{T_2} \int_S f(\boldsymbol{\sigma}, \tau) \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau\right], \tag{5.1.7}$$

provided that the integrals on either sides exist.

Proofs and details are included in Daley and Vere-Jones (2003). Sections 5.1.1 and 5.1.2 outline how it is possible to exploit this property to define of suitable weights to smooth all the considered components (up to a proportionality constant, which is instead included in the respective relaxation coefficient).

5.1.1 Reconstructing background components

The smoothing of each background component is performed separately. The general technique is always the one of *weighted kernel averaging*, but suitable weights must be identified for each piece. The discussion of the appropriate smoothing for each piece is split in subsection, where the first focuses on the spatial background, and the second on the temporal background.

Spatial Background The background spatial term $\mu_s(\cdot)$ can be constructed as for the more typical separable model without periodicity. In order to prove that, it is possible to define the usual background weights of (4.3.14) as a function of time and location:

$$\phi(\boldsymbol{s}, t) = \mu_0 \frac{\widetilde{\mu}_s(\boldsymbol{s}) \cdot \widetilde{\mu}_d(t)\widetilde{\mu}_w(t)\widetilde{\mu}_t(t)}{\lambda_c(\boldsymbol{s}, t)}, \tag{5.1.8}$$

where $\phi(\boldsymbol{s}_i, t_i) = \phi_i, \ \forall\, i$ in the observed set.

Assuming that the background is sufficiently smooth, the number of points falling into a square $\Delta\boldsymbol{s} \subset \mathcal{D}$ is an approximation to the expected number of points in the same square. Hence, by Equation (5.1.7):

$$\begin{aligned}
\sum_{i=1}^{n} \phi(\boldsymbol{s}_i, t_i) I_{\Delta\boldsymbol{s}}(\boldsymbol{s}_i) &\approx \int_0^T \int_{\mathcal{D}} \phi(\boldsymbol{\sigma}, \tau)\lambda_c(\boldsymbol{\sigma}, \tau) I_{\Delta\boldsymbol{s}}(\boldsymbol{\sigma}) d\boldsymbol{\sigma} d\tau = \\
&= \int_0^T \int_{\mathcal{D}} \frac{\widetilde{\mu}_0\widetilde{\mu}_d(\tau)\widetilde{\mu}_w(\tau)\widetilde{\mu}_t(\tau)\widetilde{\mu}_s(\boldsymbol{\sigma})}{\lambda_c(\boldsymbol{\sigma}, \tau)} \lambda_c(\boldsymbol{\sigma}, \tau) I_{\Delta\boldsymbol{s}}(\boldsymbol{\sigma}) d\boldsymbol{\sigma} d\tau = \\
&= \widetilde{\mu}_0 \left(\int_0^T \widetilde{\mu}_d(\tau)\widetilde{\mu}_w(\tau)\widetilde{\mu}_t(\tau) d\tau \right) \left(\int_{\Delta\boldsymbol{s}} \widetilde{\mu}_s(\boldsymbol{\sigma}) d\boldsymbol{\sigma} \right) \approx \\
&\approx \widetilde{\mu}_0 \left(\int_0^T \widetilde{\mu}_d(\tau)\widetilde{\mu}_w(\tau)\widetilde{\mu}_t(\tau) d\tau \right) |\Delta\boldsymbol{s}| \widetilde{\mu}_s(\boldsymbol{s}) \propto \\
&\propto \widetilde{\mu}_s(\boldsymbol{s}),
\end{aligned} \tag{5.1.9}$$

where $|\Delta\boldsymbol{s}|$ is the measure of $\Delta\boldsymbol{s}$ that shall be taken small. Hence, this suggests the following writing:

$$\hat{\mu}_s(\boldsymbol{s}) \approx const \cdot \sum_{i=1}^{n} \phi_i I_{\Delta\boldsymbol{s}}(\boldsymbol{s}_i) \propto \sum_{i=1}^{n} \phi_i I_{\Delta\boldsymbol{s}}(\boldsymbol{s}_i), \tag{5.1.10}$$

that essentially yields a histogram estimate of $\widetilde{\mu}_s(\cdot)$. It is possible to pass from this histogram estimate to a smoother version through kernel smoothing according to some normalized kernel function $\widetilde{k}_{b_s}(\cdot)$. For instance, one may consider the two-dimensional Gaussian kernel:

$$\widetilde{k}_{b_s}(\boldsymbol{\sigma}) \propto \exp\left\{ -\frac{1}{2 \cdot b_s^2} ||\boldsymbol{\sigma}||^2 \right\}, \tag{5.1.11}$$

where $||\boldsymbol{\sigma}||$ is the norm of $\boldsymbol{\sigma}$. Hence, applying this smoothing, the final smoothed estimate is obtained as:

$$\hat{\mu}_s(\boldsymbol{s}) \propto \sum_{i=1}^{n} \phi_i \frac{k_{b_s}(\boldsymbol{s}-\boldsymbol{s}_i)}{\int_{\mathcal{D}(\boldsymbol{s})} k_{\omega_s}(\boldsymbol{s}-\boldsymbol{\sigma})d\boldsymbol{\sigma}} = \sum_{i=1}^{n} \phi_i \cdot \widetilde{k}_{b_s}(\boldsymbol{s}-\boldsymbol{s}_i), \tag{5.1.12}$$

where $\mathcal{D}(\boldsymbol{s}) = \{\boldsymbol{\sigma} \in \mathbb{R}^2 : \boldsymbol{s}-\boldsymbol{\sigma} \in \mathcal{D}\}$ is the domain on which the kernel must be normalized at each point $\boldsymbol{s} \in \mathcal{D}$ in order to correct the edge effects "on average": there is no mass leaking out of the considered domain.

Temporal Background The temporal background terms require the reconstruction of three components: the *daily* term $\widetilde{\mu}_d(\cdot)$, the *weekly* term $\widetilde{\mu}_w(\cdot)$ and the long-trend term $\widetilde{\mu}_t(\cdot)$. These can be re-constructed very similarly, and share a common justification behind the shape of the corresponding weight functions:

$$\phi^a(\boldsymbol{s},t) = \frac{\widetilde{\mu}_a(t)\widetilde{\mu}_s(\boldsymbol{s})}{\lambda_c(\boldsymbol{s},t)}, \quad \phi_i^a = \phi_a(\boldsymbol{s}_i,t_i), \tag{5.1.13}$$

with $a \in \{d,w,t\}$ and implicitly $\widetilde{\mu}_a(t) = \widetilde{\mu}_a(a(t))$. The function $a(\cdot)$ maps the raw time t to the corresponding point in the period identified by a. Let us consider a periodic domain on $[0,m_a]$, i.e. partition the interval $[0,T]$ in sub-intervals (periods):

$$[0,T] = [0,m_a] \cup (m_a, 2\cdot m_a] \cup \cdots \cup (T_a \cdot m_a, T],$$

where $T_a = \lfloor T/m_a \rfloor$. Any t can be mapped into the corresponding interval point as $a(t) = t - m_a \left\lfloor \frac{t}{m_a} \right\rfloor$. For instance, if t is on the scale of hours, then $m_d = 24$ and $m_w = 168$, while $m_t = T$ in any case.

Then, similarly to the spatial background, let us consider a sub-interval Δt of $[0,T]$ and observe that:

$$\begin{aligned}
\sum_{i=1}^{n} \phi^a(\boldsymbol{s}_i,t_i) I_{\Delta t}(t_i) &\approx \int_0^T \int_{\mathcal{D}} \phi_a(\boldsymbol{\sigma},\tau)\lambda_c(\boldsymbol{\sigma},\tau) I_{\Delta t}(\tau) d\boldsymbol{\sigma} d\tau = \\
&= \int_0^T \int_{\mathcal{D}} \frac{\widetilde{\mu}_a(\tau)\widetilde{\mu}_s(\boldsymbol{\sigma})}{\lambda_c(\boldsymbol{\sigma},\tau)} \lambda_c(\boldsymbol{\sigma},\tau) I_{\Delta t}(\tau) d\boldsymbol{\sigma} d\tau = \\
&= \left(\int_{\mathcal{D}} \widetilde{\mu}_s(\boldsymbol{\sigma}) d\boldsymbol{\sigma}\right) \cdot \left(\int_{\Delta t} \widetilde{\mu}_a(\tau) d\tau\right) \approx \\
&\approx \left(\int_{\mathcal{D}} \widetilde{\mu}_s(\boldsymbol{\sigma}) d\boldsymbol{\sigma}\right) \cdot |\Delta t| \widetilde{\mu}_a(t) \propto \\
&\propto \widetilde{\mu}_a(t),
\end{aligned} \tag{5.1.14}$$

where $|\Delta t|$ is the measure of Δt that shall be taken small. Equation (5.1.14) is a histogram estimate of the temporal background components $\widetilde{\mu}_a(\cdot)$, that can be smoothed considering a suitable normalized kernel function $\widetilde{k}_{b_a}(\cdot)$. The kernels are indexed by a since different components can have different bandwidths and are normalized over different domains $[0,m_a]$. For instance, one may consider Gaussian kernels for all the temporal components:

$$\widetilde{k}_{b_a}(\tau) \propto \exp\left\{-\frac{1}{2\cdot b_a^2}\tau^2\right\}. \tag{5.1.15}$$

The final smoothed estimate is:

$$\hat{\mu}_a(t) \propto \sum_{i=1}^{n} \phi_i^a \frac{k_{b_a}(a(t)-a(t_i))}{\int_{\mathcal{T}_a(t)} k_{b_a}(a(t)-\tau)d\tau} = \sum_{i=1}^{n} \phi_i^a \cdot \widetilde{k}_{b_a}(a(t)-a(t_i)), \tag{5.1.16}$$

where $\mathcal{T}_a(t) = \{\tau \in \mathbb{R} : a(t) - \tau \in [0, m_a]\}$ is the domain on which the kernel must be normalized at each point $t \in [0, T]$.

5.1.2 Reconstructing excitation components

The reconstruction of the temporal and spatial excitation components is obtained by *weighted Kernel smoothing* of all the observed pairwise space and time lags. Given two arbitrary points $\boldsymbol{p} = (\boldsymbol{s}, t)$ and $\boldsymbol{p}' = (\boldsymbol{s}', t')$, the spatial lag is evaluated separately along the two dimensions through the lag vector is $\boldsymbol{p} - \boldsymbol{p}' = (s_1 - s_1', s_2 - s_2', t - t')$.

The required weights are the triggering probabilities already introduced in Equation (4.3.13), which can be expressed as a function of any pair as:

$$\rho(\boldsymbol{s}, t, \boldsymbol{s}', t') == \begin{cases} \frac{A \cdot \widetilde{g}_s(s_1 - s_1', s_2 - s_2') \cdot \widetilde{g}_t(t - t')}{\lambda_c(\boldsymbol{s}', t')} & t < t' \\ 0 & t \geq t', \end{cases} \tag{5.1.17}$$

with $\rho(\boldsymbol{s}_i, t_i, \boldsymbol{s}_j, t_j) = \rho_{ij} \; \forall \, i, j$.

Theoretical justification to this weights is analogous to the ones of the background components, and relies on the fact that for any fixed $\boldsymbol{p} = (\boldsymbol{s}, t)$ the weight function of Equation (5.1.17) is deterministic (hence predictable). Therefore, fixing one observation $(\boldsymbol{s}_i, t_i)$, (5.1.7) can be applied in order to justify the weights of the temporal excitation:

$$\begin{aligned} \sum_{j=1}^{n} \rho_{ij} I_{\Delta t}(t_j - t_i) &\approx \int_0^T \int_{\mathcal{D}} \rho(\boldsymbol{s}_i, t_i, \boldsymbol{\sigma}, \tau) \lambda_c(\boldsymbol{\sigma}, \tau) I_{\Delta t}(\tau - t_i) d\boldsymbol{\sigma} d\tau \\ &= A \cdot \int_{t_i}^T \int_{\mathcal{D}} \frac{\widetilde{g}_t(\tau - t_i) \widetilde{g}_s(\sigma_1 - s_1, \sigma_2 - s_2)}{\lambda_c(\boldsymbol{\sigma}, \tau)} \lambda_c(\boldsymbol{\sigma}, \tau) I_{\Delta t}(\tau - t_i) d\boldsymbol{\sigma} d\tau \\ &= A \cdot \left(\int_{t_i}^T \widetilde{g}_t(\tau - t_i) I_{\Delta t}(\tau - t_i) d\tau \right) \left(\int_{\mathcal{D}} \widetilde{g}_s(\sigma_1 - s_1, \sigma_2 - s_2) d\boldsymbol{\sigma} \right). \end{aligned}$$

Replacing $u = \tau - t_i$ then:

$$\begin{aligned} \sum_{j=1}^{n} \rho_{ij} I_{\Delta t}(t_j - t_i) &\approx A \cdot \left(\int_0^{T - t_i} \widetilde{g}_t(u) I_{\Delta t}(u) du \right) \left(\int_{\mathcal{D}} \widetilde{g}_s(\sigma_1 - s_1, \sigma_2 - s_2) d\boldsymbol{\sigma} \right) = \\ &= A \cdot \left(\int_{\Delta t} \widetilde{g}_t(u) du \right) \left(\int_{\mathcal{D}} \widetilde{g}_s(\sigma_1 - s_1, \sigma_2 - s_2) d\boldsymbol{\sigma} \right) \approx \\ &\approx A \cdot |\Delta t| \cdot \widetilde{g}_t(t) \cdot \left(\int_{\mathcal{D}} g_s(\sigma_1 - s_1, \sigma_2 - s_2) d\boldsymbol{\sigma} \right) \propto \\ &\propto \widetilde{g}_t(t), \end{aligned} \tag{5.1.18}$$

where the last passage does not depend on the time or location $(\boldsymbol{s}_i, t_i)$. Therefore:

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \rho_{ij} I_{\Delta t}(t_j - t_i) \right) \approx n \cdot \sum_{j=1}^{n} \rho_{ij} I_{\Delta t}(t_j - t_i),$$

and the histogram estimate of the temporal excitation function results to be:

$$\hat{g}(t) \propto \sum_{i,j=1}^{n} \rho_{ij} I_{\Delta t}(t_j - t_i). \tag{5.1.19}$$

Similarly, the histogram estimate of the spatial excitation can be obtained as:

$$\hat{g_s}(s_1, s_2) \propto \sum_{i,j=1}^{n} \rho_{ij} I_{\Delta s_1}(s_{j1} - s_{i1}) I_{\Delta s_2}(s_{j2} - s_{i2}), \tag{5.1.20}$$

where $\boldsymbol{s}_i = (s_{i1}, s_{i2})$ and $\boldsymbol{s}_j = (s_{j1}, s_{j2})$.

As usual, these histogram estimates can be converted into a smooth version via some suitable kernel function. For the temporal triggering histogram estimate of Equation (5.1.19) one may consider a uni-variate Gaussian kernel as the one of Equation (4.1.10):

$$\hat{g}_t(t) \propto \sum_{i,j=1}^{n} \rho_{i,j} \frac{k_{h_t}\left(t - (t_i - t_j)\right)}{\int_0^{T-t_i} k_{h_t}\left(\tau - (t_i - t_j)\right) d\tau} = \sum_{i,j=1}^{n} \rho_{i,j} \cdot \widetilde{k}_{h_t}\left(t - (t_i - t_j)\right), \tag{5.1.21}$$

where $\widetilde{k}_{h_t}(\cdot)$ denotes the normalized kernel.

The spatial triggering histogram estimate is instead estimated through composition of two uni-dimensional Gaussian kernels over the two spatial dimensions. This yields the following:

$$\hat{g}_s(\sigma_1, \sigma_2) \propto \sum_{i,j=1}^{n} \rho_{i,j} \frac{k_{h_s}\left(\sigma_1 - (s_{i1} - s_{j1})\right) \cdot k_{h_s}\left(\sigma_2 - (s_{i2} - s_{j2})\right)}{\int_{\mathcal{D}} k_{h_s}\left(\sigma_1 - (s_{i1} - s_{j1})\right) \cdot k_{h_s}\left(\sigma_2 - (s_{i2} - s_{j2})\right) d\boldsymbol{\sigma}}, \tag{5.1.22}$$

where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ and the composition kernel is scaled altogether.

5.1.3 Determining relaxation coefficients

Once all the functions involved in the expression of the conditional intensity function have been estimated up to a proportionality constant, we must recall that the model includes two relaxation coefficients μ_0 and A. These are supposed to globally adjust for the missing constants and pull background (the first) and intensity (the second) toward the most suitable level for the observed pattern. Conceptually, these specify how much or little of the triggering and background components respectively enters the conditional intensity function. It should be noted that A can be interpreted as the proportion (on average) of the impact of the triggering function on the total intensity: a high A value suggests data is dominated by triggering and the converse for a small value. In particular, for $A > 1$ the counts will eventually explode for a large enough time-window (see Section 4.3.2).

The most natural way to estimate the value of μ_0 and A that best fit the observed point pattern would be *Maximum Likelihood Estimation*, conditionally on the non-parametrically estimated background and intensity functions.

The log-likelihood function of $\{(\boldsymbol{s}_i, t_i)\}_{i=1}^n$, realization of a finite and simple spatial-temporal point process over a bounded domain $\mathcal{Q} = \mathcal{D} \times [0, T]$, is known and can be evaluated as:

$$\log(\mathcal{L}_{\lambda_c}) = \sum_{i=1}^{n} \log\left(\lambda_c(\boldsymbol{s}_i, t_i,)\right) - \int_0^T \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau, \tag{5.1.23}$$

for whatever $\lambda_c(\cdot,\cdot)$. Replacing the conditional intensity function form of Equation (5.1.6) into it:

$$\begin{aligned}
\log(\mathcal{L}_{\mu_0,A}) &= \sum_{i=1}^{n} \log\left(\mu_0\widetilde{\mu}_d(t_i)\widetilde{\mu}_w(t_i)\widetilde{\mu}_t(t_i)\widetilde{\mu}_s(\boldsymbol{s}_i) + A\sum_{t_j<t_i} g_t(t_i-t_j)g_s(\boldsymbol{s}_i-\boldsymbol{s}_j)\right) + \\
&\quad - \int_0^T\int_{\mathcal{D}} \mu_0\widetilde{\mu}_d(\tau)\widetilde{\mu}_w(\tau)\widetilde{\mu}_t(\tau)\widetilde{\mu}_s(\tau) + A\sum_{t_j<\tau} g_t(\tau-t_j)g_s(\boldsymbol{\sigma}-\boldsymbol{s}_j)d\boldsymbol{\sigma}d\tau \\
&= \sum_{i=1}^{n} \log\left(\mu_0\widetilde{\mu}_d(t_i)\widetilde{\mu}_w(t_i)\widetilde{\mu}_t(t_i)\mu_s(\boldsymbol{s}_i) + A\sum_{t_j<t_i} g_t(t_i-t_j)g_s(\boldsymbol{s}_i-\boldsymbol{s}_j)\right) + \\
&\quad - \widetilde{\mu}_0\int_0^T\int_{\mathcal{D}} \widetilde{\mu}_d(\tau)\widetilde{\mu}_w(\tau)\widetilde{\mu}_t(\tau)\widetilde{\mu}_s(\tau)d\boldsymbol{\sigma}d\tau - A\int_0^T\int_{\mathcal{D}}\sum_{t_j<\tau} g_t(\tau-t_j)g_s(\boldsymbol{\sigma}-\boldsymbol{s}_j)d\boldsymbol{\sigma}d\tau \\
&= \sum_{i=1}^{n} \log\left(\mu_0\widetilde{\mu}_d(t_i)\widetilde{\mu}_w(t_i)\widetilde{\mu}_t(t_i)\mu_s(\boldsymbol{s}_i) + A\sum_{t_j<t_i} g_t(t_i-t_j)g_s(\boldsymbol{s}_i-\boldsymbol{s}_j)\right) + \\
&\quad - \mu_0\int_0^T\int_{\mathcal{D}} \widetilde{\mu}_d(\tau)\widetilde{\mu}_w(\tau)\widetilde{\mu}_t(\tau)\widetilde{\mu}_s(\tau)d\boldsymbol{\sigma}d\tau - A\sum_{j=1}^{n}\int_{t_j}^T\int_{\mathcal{D}} g_t(\tau-t_j)g_s(\boldsymbol{\sigma}-\boldsymbol{s}_j)d\boldsymbol{\sigma}d\tau.
\end{aligned}$$

The following computations are performed conditionally on all the background and intensity components but μ_0 and A, hence for the sake of brevity let us denote:

$$\begin{aligned}
U &= \int_0^T\int_{\mathcal{D}} \mu_d(\tau)\mu_w(\tau)\mu_{tr}(\tau)\mu_s(\boldsymbol{\sigma})d\boldsymbol{\sigma}d\tau, \\
G &= \sum_{j=1}^{n}\int_{t_j}^T\int_{\mathcal{D}} g_t(\tau-t_j)h_s(\boldsymbol{\sigma}-\boldsymbol{s}_j)d\boldsymbol{\sigma}d\tau.
\end{aligned}$$

Computing the partial derivatives of the log-likelihood with respect to A and μ_0 the result is:

$$\begin{aligned}
\frac{d\log(\mathcal{L}_{\mu_0,A})}{d\mu_0} &= \sum_{i=1}^{n}\frac{\mu_d(t_i)\mu_w(t_i)\mu_{tr}(t_i)\mu_s(\boldsymbol{s}_i)}{\lambda_c(t_i,\boldsymbol{s}_i)} - U, \\
\frac{d\log(\mathcal{L}_{\mu_0,A})}{dA} &= \sum_{i=1}^{n}\frac{\sum_{t_j<t_i} g_t(t_i-t_j)g_s(\boldsymbol{s}_i-\boldsymbol{s}_j)}{\lambda_c(t_i,x_i)} - G.
\end{aligned}$$

Finally, setting these equal to 0 and solving the system one attains the following iterative system of equations:

$$\begin{aligned}
\psi_i^{(\zeta)} &= \frac{\mu_0^{(\zeta)}\mu_d(t_i)\mu_w(t_i)\mu_{tr}(t_i)\mu_s(x_i)}{\mu_0^{(\zeta)}\mu_d(t_i)\mu_w(t_i)\mu_{tr}(t_i)\mu_s(\boldsymbol{s}_i) + A^{(\zeta)}\sum_{t_j<t_i} g_t(t_i-t_j)g_s(\boldsymbol{s}_i-\boldsymbol{s}_j)} \\
A^{\zeta+1} &= \frac{n-\sum_{i=1}^{n}\psi_i^{(\zeta)}}{G} \\
\mu_0^{\zeta+1} &= \frac{n-A^{(\zeta+1)}G}{U},
\end{aligned} \tag{5.1.24}$$

which can be iteratively solved in order to optimize all components in the log-likelihood while saving the computation time of numerical optimization.

Alternatively, one may consider numerical optimization methods of any sort. Exploiting (5.1.24) to determine suitable starting points and then searching for the optimum numerically can result in faster and more stable results.

5.1.4 Stochastic Reconstruction

The procedures described in Section 5.1.1 and 5.1.2 allow to reconstruct background and excitation components given the relaxation coefficients, which are necessary in order to compute the weights of Equations (5.1.8), (5.1.13) and (5.1.17). At the same time, Section 5.1.3 outlines an optimization method to determine μ_0 and A given the reconstructed background and excitation components.

This circularity can be solved in the spirit of the EM algorithm, by alternating between finding the optimal relaxation coefficients given background and excitation forms and viceversa. We refer again to the original work Zhuang and Mateu (2019) for a version of the stochastic reconstruction implementing this exact logic. Its description is detailed in Algorithm 4.

Algorithm 4: Semi-parametric Stochastic reconstruction

Input: Initial guesses $\mu_s^{(0)}, \mu_d^{(0)}, \mu_w^{(0)}, \mu_t^{(0)}, g_t^{(0)}, g_s^{(0)}, A^{(0)}, \mu_0^{(0)}$
Set $k = 1$
while *Not converged* **do**
- Compute $\phi_i^d, \phi_i^w, \phi_i^t, \phi_i, \rho_{i,j}$ for all valid (i, j) using Eq. (5.1.8), (5.1.13) and (5.1.17) using the current components
- Reconstruct $\mu_s^{(k)}, \mu_d^{(k)}, \mu_w^{(k)}, \mu_t^{(k)}, g_t^{(k)}, g_s^{(k)}$ using Eq. (5.1.12), (5.1.16), (5.1.21), (5.1.21) and (5.1.22) with the current weights.
- Determine optimal $A^{(k)}$ and $\mu_0^{(k)}$ using Eq. (5.1.24) given the current components
- Check *convergence*

Output: Optimized components $\mu_d, \mu_w, \mu_{tr}, \mu_s, g, h, A, \mu_0$

Convergence can be checked using different criteria. Zhuang and Mateu (2019) suggests fixing a small value $\epsilon > 0$ and stop if the increase in the log-likelihood with respect to the previous iteration is lower than that, i.e. if:

$$\log(\mathcal{L}_{\lambda_c^{(k)}}) - \log(\mathcal{L}_{\lambda_c^{(k-1)}}) < \epsilon, \text{ with } \epsilon \text{ small.}$$

Eventually, after a reasonable number of iterations (which is application-dependent), the log-likelihood increments flatten and a (local) maximum has been found. There is no guarantee that the optimum is global, therefore multiple runs starting from different initial guesses are suggested. For those interested in an implementation of the original methodology detailed in Zhuang and Mateu (2019), this can be found at `https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

One main drawback of this algorithm resides in its computational complexity, that is mostly affected by the smoothing of the excitation function (that scales quadratically with the number of points) and the likelihood maximization. One idea to reduce the burden of the former, is to assume that triggering gets negligible after a pre-specified distance in time and/or space (excitation *tapering*). This can sensibly reduce the number of eligible pairs for the smoothing. Mohler (2014b) also suggests to perform each iteration only on a random subset of all the points. However, while this reduces the burden of each iteration, may require a larger number of iterations to reach convergence.

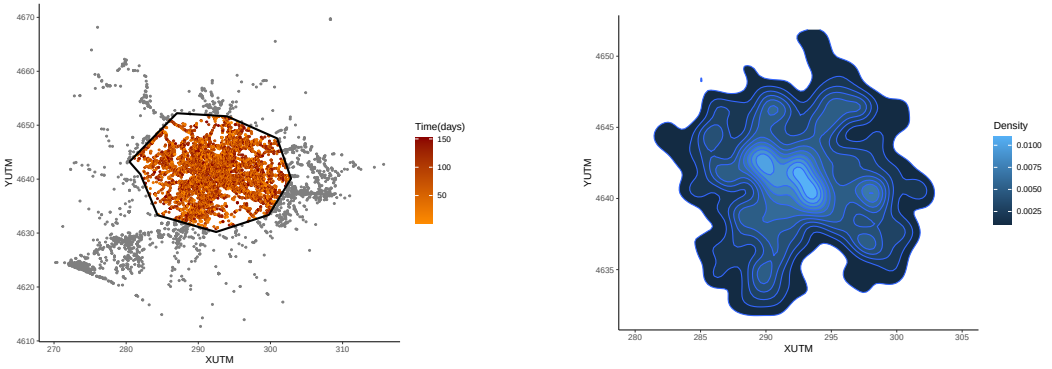
The next sections describe two applications to real data that present some slight variations with respect to the more general methodology outlined until now. The first one is on road accidents occurred on the urban network of the City of Rome; it is very preliminary and will see further developments in the near future. The second one is on car accidents occurred on the M25 London Orbital, a British *motorway*; it is final and has been recently published in Kalair et al. (2020).

5.2 Preliminary application to road-accidents in Rome

Since the second half of the 21st century, car accidents have determined a significant portion of the global death toll. The *World Health Organization* (WHO) estimates that approximately 1.35 million people are cut short every year as a result of a road traffic crash, putting them at the 9-th position among all the causes of deaths ranked by number of victims. In particular, they are the leading cause of death for young people aged between 18 and 35 years in a great number of Nations, among which Italy. According to the latest annual report of the "Italian Statistical Office", 172,183 road accidents resulting in death or injury occurred in Italy during 2019, of which 3,173 were deaths (within 30 days). The social cost of the road accidents for the same year, calculated on the basis of parameters indicated by the *Ministry of Infrastructure and Transport*, at constant 2010 values, is equal to 16.9 billion euros (ACI and ISTAT, 2019).

It is therefore obvious how during the last decades the need for accurate modeling of car accidents occurrences has become a priority for national and supra-national organizations. This is needed in order to identify the major causes of the crashes, verify the impact of implemented policies and drive future ones (La Torre et al., 2007). However, apart from the widespread and very common epidemiological studies based on medical records, poor effort has been addressed toward direct modeling of the spatial distribution of these events. This is probably due to the historical lack of data about the exact space-time location of the vehicles at the moment of the accident, whose retrieval has always been a challenge. Recent technological developments are finally enabling straightforward and cheap ways of recording these information. For instance, since 2014, road authorities of the City of Rome have been equipped with a GPS device and are assigned the task to record time and location of every car accident on which their intervention is required. These geo-referenced records are collected monthly and, in the spirit of the *open-data* era we are living in (`https://opendefinition.org/`; Dominici (2015)), published for public use at `https://dati.comune.roma.it/catalog/dataset?tags=Incidenti&groups=sicurezza-urbana`. This gold-mine of public data, whose structure was before reserved to small ad-hoc studies, has already drawn some attention (Alaimo Di Loro, 2016; Comi et al., 2018; Alaimo Di Loro et al., 2019).

This work investigates the presence of a spatial and temporal excitation mechanism between road accidents occurring on the Rome urban road network. To the author's knowledge, this is technically the first attempt to apply the spatio-temporal Hawkes process as a model for spatial and temporal clustering in this context (Li et al. (2018) considers only time). The following application is the result of a collaboration with Prof. Zhuang Jiancang during the LML 2019 Summer School and, as specified in the title, it is only a preliminary attempt that requires further investigation and adaptation of the original model. This has been included in this dissertation since it presents some simple but interesting ideas and it is the milestone that allowed the following production of the paper Kalair et al. (2020), later exposed in Section 5.3.



(a) Territory of Rome

(b) Density in urban area

Figure 5.1. Spatial distribution of the road accidents occurred in the territory of Rome between May and September 2017: all events colored by time over the whole territory of Rome (left); spatial density of the events over the urban area (right).

5.2.1 Data

The data considered for the following application have been downloaded from the *Open Data* website of the City of Rome[1]. On this web-page, a great variety of data-sets resulting from administrative inquiries are made available for download, with a *Creative Commons* licence that allows for their free visualization and utilization for any scope.

As mentioned in the introduction, the City of Rome has been publishing monthly data-sets that contain all the road accidents occurred in its territory during each month of every year since 2006. Each dataset contains all the road accidents a police patrol intervened at, but does not include all the cases in which the involved subjects achieved a friendly agreement. In particular, all accidents occurred on the *Grande Raccordo Anulare* (road-ring circling Rome) are not included in the dataset.

The records are not referred one-to-one to every accident, but to every involved subject. In practice, to each road accident correspond as many records as the individuals involved in it. These individuals may be pedestrians, bikes or vehicles. Nevertheless, different records referred to the same accidents can be easily recognized thanks to a common protocol-ID which is accident-specific. Every record is also accompanied by different features that describe the context and characteristics of the event, as well as general information about the involved person. The relevant information on the accident is retrieved from the report compiled by the Police officer at the scene and generally includes: date and time, road name and closest civic number, accident nature, road type, weather and traffic conditions, lighting, number of involved people, deceased and injured, and many other individual specific variables. Since 2014, each record also reports the exact geo-referenced location of the crash. This yields a point pattern and, hence, makes the data viable for analysis through point processes.

This preliminary application considers all the road accidents recorded between May and September 2017, with the aim to verify whether the crashes exhibit some clustering and, possibly, triggering effect. The latter is verified through the fit of the non-parametric spatio-temporal Hawkes process with periodic background model, using the *Stochastic Reconstruction* algorithm. The reason why only a subset of all the available data is being used lies in the high computational cost of Algorithm

[1] `https://dati.comune.roma.it/od/en/progetto.page`

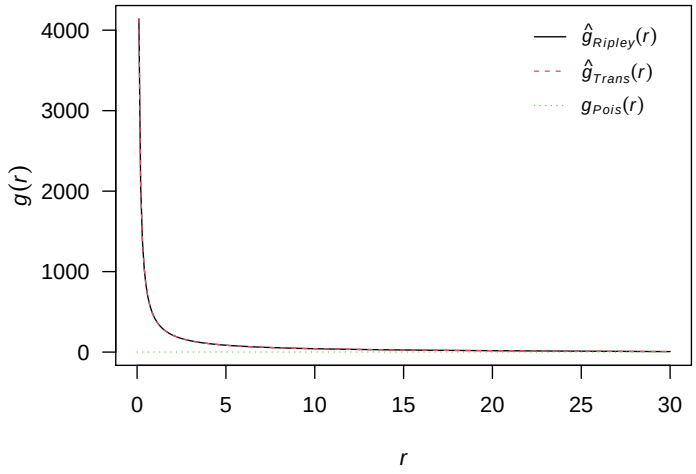
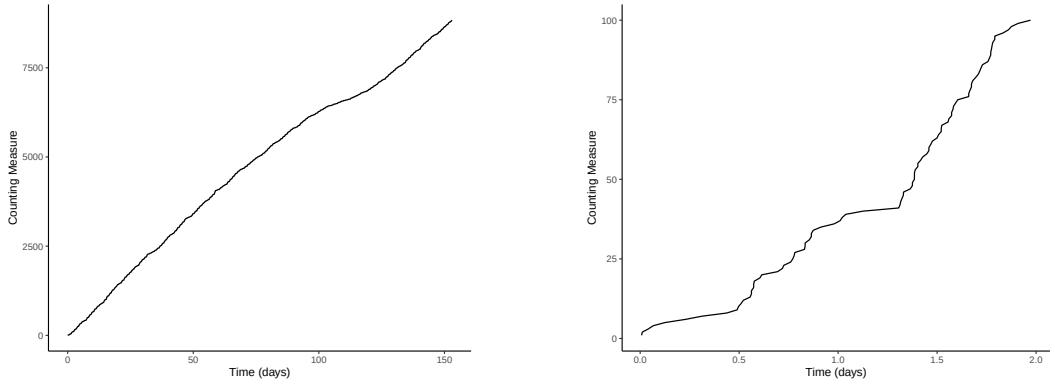


Figure 5.2. Estimated spatial Pair correlation functions at lag r with *Ripley*'s (full black) and *translation* (dashed red) correction, as compared to the Homogeneous Poisson correlation function (dotted green).

4, that scales quadratically with the number of events. Indeed, the reduced subset already counts $n = 11365$ records and requires the algorithm to run for several hours on a standard computing machine before reaching convergence.

More computational power and time may be deployed, but the final application is on hold for the implementation of some (later discussed) improvements. As a matter of fact, the method is not yet refined to deal properly with road accidents data recorded on a urban network. The current version operates under the working assumption that the events occur over a compact surface $\mathcal{S} \subset \mathbb{R}^2$, while road accidents actually take place on a much different domain: the geometrically structured space of the road network. Ad-hoc methods and adaptation of the general methodology to the case at hand are currently under development.

A graphical representation of the spatial distribution of events is shown in Figure 5.1. The accidents in 5.1a are divided between those occurred in the urban (scale of reds) and in the suburban area (gray). While all road accidents will be involved in the fitting procedure, the suburban area is used only as a *buffering zone*: events happening there will be exploited in the computation of the smoothed estimates of the various components, but won't be considered in the likelihood maximization of Algorithm 4. The use of such a buffering zone is done in order to alleviate the edge effects of the smoothing estimates, both in the background and in the triggering (see Section 4.3.5 and Zhuang et al. (2004)). The urban accidents are $\approx 78\%$ of all the road accidents, with a total number of $n^* = 8835$. Figure 5.1b provides a naive non parametric estimate of the road accidents density in the urban area, and it highlights how the spatial distribution is not homogeneous. This can be due to variations of traffic and road conditions across the city. However, previous studies on road accidents (see the Introduction to Chapter 5) proved that these do not just exhibit inhomogeneity, but also clustering in space. This can be further verified by computing the pair correlation function (or *g-function*), theoretically introduced in Section 4.1.1 at Equation (4.1.4), and whose estimate has been briefly described in Section 4.1.2 at Equation 4.1.7. Figure 5.2 shows how the observed spatial point pattern presents an evident clustering behavior as compared to a standard Poisson process.



(a) Whole period **(b)** Two random days

Figure 5.3. Empirical counting measure of the road accidents occurred in the urban area od Rome between May and September 2017

The clustering effect is way more evident using a visual inspection along the temporal dimension. Figure 5.3 shows the increments in the counting measure for every unit increment of time. An homogeneous process would show an approximately linear behavior. An inhomogeneous process would present smooth accelerations of the increments on some sections. A clustering process, instead, would present (apparently random) sudden increments in the counting measure, followed by flat sections (Diggle and Ribeiro Jr, 2007). Figure 5.3a presents all the data on the original scale: the behavior seems coherent with an inhomogeneous process, with some spikes barely visible. Figure 5.3b zooms in, focusing on just the first two days: the behaviour of the counting measure highlights the evident clustering pattern of the events.

We would like to verify if some of it can be explained through a triggering effect of the car accidents on others.

5.2.2 Model variations

Before fitting the periodic non-parametric spatio-temporal Hawkes process model for the considered data, we consider some slight variations with respect to its original version. The first one acts at the model specification level, and consists of modifying the smoothing of the spatial excitation enforcing isotropy (i.e. triggering decays with the euclidean distance). The second instead introduces some ad-hoc edge corrections on the smoothing of the different model components.

Isotropic spatial excitation In the original semi-parametric estimation of the spatial excitation function introduced in Section 5.1.2, the excitation is a function of two arguments $g_s(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}^+$. It allows the triggering effect to decay differently along the two geo-spatial dimensions. Despite this freedom, the results included in Zhuang and Mateu (2019) and a preliminary application to the car accidents in Rome returned an estimated spatial excitation with an approximately isotropic behavior. Apparently, as the intuition would suggest, the decay in the excitation from one point to the other depends on the relative positions between the two points only through the euclidean distance between those $d(\boldsymbol{u}, \boldsymbol{v}) = ||\boldsymbol{u} - \boldsymbol{v}||$. Motivated by this, we may consider performing the smoothing on the observed pairwise euclidean distances, rather than on independent distances along the two dimensions as in Equation

(5.1.22). One may think this could be easily achieved by computing all the pairwise Euclidean distances $\{d_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j||, i, j = 1, \dots, n,\ j \neq i\}$ but, before proceeding with the smoothing, some geometric considerations must be carefully made. Indeed, given one point $\boldsymbol{s}$ on a two-dimensional surface, points at different distances from it lie on different sub-spaces with different sizes. The smaller the distance and the smaller is the circle on which such distance could have been observed. Hence, given a distance d, the histogram estimator shall be suitably re-scaled by $(2\pi d)^{-1}$, i.e. the size of the space on which such distance lies:

$$\hat{g_s}(d) \propto \sum_{i,j=1}^{n} \frac{\rho_{ij}}{2\pi d_{ij}} I_{\Delta d}(d_{ij}), \tag{5.2.1}$$

where $d \in \mathbb{R}^+$. Therefore, the corresponding kernel smoothed estimate of Equation (5.1.22) becomes:

$$\hat{g}_s(d) \propto \sum_{i,j=1}^{n} \rho_{i,j} \frac{\widetilde{k}_{h_s}\left(d - (d_{ij})\right)}{2\pi d_{ij}},$$

where $h_s > 0$ is the bandwidth. This approach enforces exact isotropy (i.e. radial decay), yielding a more parsimonious specification of the spatial excitation function, less prone to over-fitting and likely to be a better fit for the data excitation mechanism.

Furthermore, it also provides the additional advantage of shrinking the computational burden of the estimation process both from a memory and a flop counts point of view. Indeed, the number of point-wise pairs distances goes from $2 \cdot n^2$ (each pair along the two dimensions) to n^2, requiring less memory and fewer kernel evaluations.

Edge correction It is well known that kernel density estimates are biased near the boundary on truncated domains, with discussion of this in Chiu (2000) and references within. *"On average"* corrections such as kernel normalization alleviate this problem, which by the way remains important when the density does not vanish to 0 at the boundaries. Various solutions have been proposed in the literature, each having its own implications. Hence, different corrections have been considered for different components of the model.

Let us consider the two triggering functions, truncated at their lower boundary 0. When the function is expected to keep the same behavior as long as it approaches the domain bound, the reflection correction method 'mirrors' (or reflects) the estimate over such boundary (Hominal and Deheuvels, 1979; Schuster, 1985; Silverman, 1986; Jones, 1993). Let $x_1, x_2, \dots x_N$ be the set of observed points in $[0, \infty)$. A compelling and practical way to attain reflection about 0 consists of introducing extra points $-x_1, -x_2, \cdots - x_N$. The smoothed estimate $\hat{\nu}_r(\cdot)$, using a bandwidth $b > 0$, is then:

$$\hat{\nu}_r(x) = \frac{1}{N} \sum_{i=1}^{N} \left[\widetilde{K}_b\left(x - x_i\right) + \widetilde{K}_b\left(x + x_i\right)\right]. \tag{5.2.2}$$

Provided the kernel is symmetric and differentiable, some easy manipulation shows that the estimate will always have zero derivative at the boundary. It is clear that it is not usually necessary to reflect the whole data set: if x_i is sufficiently large, the reflected point $-x_i$ will not be felt in the calculation of $\hat{\nu}_r(x)$ for $x \geq 0$, and only points close to 0 shall be reflected. This correction may be reasonably applied to the excitation functions lower boundaries.

Other components in the proposed model are not truncated at a specific point, but rather periodic over some domain. For instance, this is true for the daily

and weekly background components. These two components do not strictly need boundary correction, because they do not actually have any bound, but rather repeat the same pattern over and over with some periodicity. Let us consider the daily background as an example. Events times t_i are observed on the whole domain $[0, T]$ and can be mapped onto some periodic domain $[0, M_d]$, where M_d represents (for instance) the number of minutes in a day. The mapped times are then given by $m_i = t_i - M_d \left\lfloor \frac{t_i}{M_d} \right\rfloor \in [0, M_d]$. All the mapped points end up in $[0, M_d]$, which we refer to as the *reference section*. Nevertheless, the exact same pattern supposedly repeats itself at any other previous or following section $[l \cdot m_d, (l+1) \cdot m_d]$, $l \in \mathbb{Z} \setminus \{0\}$. These repeated patterns contain points that shall contribute to our kernel estimate $\hat{\nu}_p(\cdot)$ defined over $[0, M_d]$, and their contribution can be especially relevant near the boundaries. As a result, a thorough periodic estimate would require to evaluate an infinite sum, accounting for the contribution of all the repeated data-points in all the sections $[l \cdot m_d, (l+1) \cdot m_d]$, $l \in \mathbb{Z}$. However, as for the reflection correction, the decay with distance of most kernels (e.g. the Gaussian kernel) implies that the contribution to $\hat{\nu}_p(\cdot) : [0, M_d] \to \mathbb{R}^+$ of points too far from the bounds is negligible. Generally, restricting the evaluation to only the previous period $[-M_d, 0]$ and following period $[M_d, 2M_d]$, yields indistinguishable results from the complete sum (for a small enough bandwidth). Hence, the smoothed contribution of each point in $\{m_i\}_{i=1}^N$ to the function $\hat{\nu}_p(\cdot)$ in any point $m \in [0, M_d]$ shall account for m_i, but also for $m_i^- = -M_d + m_i$ and $m_i^+ = M_d + m_i$, weighted through an appropriate kernel $\widetilde{K}_b$ with bandwidth b and normalized over $[-M_d, 2M_d]$. These contributions can be incorporated into the estimate as:

$$\hat{\nu}(m) = \frac{1}{3N} \sum_{i=1}^{N} \widetilde{k}_b \left(m - m_i^- \right) + \widetilde{k}_b \left(m - m_i \right) + \widetilde{k}_b \left(m - m_i^+ \right), \tag{5.2.3}$$

with $m \in [0, m_d]$. Note that the kernel must be normalized over the entire domain, including the extended points, but only values of the function evaluated in the reference period section $[0, M_d]$ are retained. Hence, there is no need for further boundary corrections since there is no interest in the behavior at the extended boundary. Trivial examples of edge corrections results are available in the Appendix B.1.

Two different corrections for the triggering components (reflection) and the daily and weekly periodic components (periodic) have been proposed here. Let us recall that the contribution of each point must be further scaled by the corresponding weights introduced in Equations (5.1.17) and (5.1.13) (as in Equations (5.1.21) and (5.1.16)).

5.2.3 Results

Bandwidth Selection and Model Choice In order to verify the ability of the triggering component to explain part of the clustering behavior in the analyzed dataset, models with and without triggering have been fit to the data and performances are compared in terms of the resulting likelihood. The validity of the best model is then further verified at the end of Section 5.2.3.

Let us denote the excitation component at $(\boldsymbol{s}, t) \in \mathcal{D} \times \mathcal{T}$ as $J(\boldsymbol{s}, t) = \int_0^t \int_{\mathcal{D}(\boldsymbol{s})} g_t(t - \tau) g_s(||\boldsymbol{s} - \boldsymbol{\sigma}||) d\tau d\boldsymbol{\sigma}$. The four following models have been considered:

Trend. Non-periodic but non-stationary Poisson model: $\lambda(\boldsymbol{s}, t) = \mu_0 \cdot \mu_{tr}(t) \mu_s(\boldsymbol{s})$;

Model	$\hat{\mu}$	$\hat{A}$	log-likelihood
Trend	0.1670		−18883.85
Trend + Triggering	0.1534	0.0915	−18674.64
Trend + Periodic	0.1670		−17303.34
Trend + Periodic + Triggering	0.1582	0.0592	−17216.65

Table 5.1. Log-likelihood values for models with various components.

Trend + Triggering. Non-periodic Hawkes model: $\lambda(\boldsymbol{s}, t) = \mu_0 \cdot \mu_{tr}(t)\mu_s(x, y) + A \cdot J(\boldsymbol{s}, t)$

Trend + Periodic. Periodic Poisson model: $\lambda(\boldsymbol{s}, t) = \mu_0 \cdot \mu_{tr}(t)\mu_d(t)\mu_w(t)\mu_s(\boldsymbol{s})$

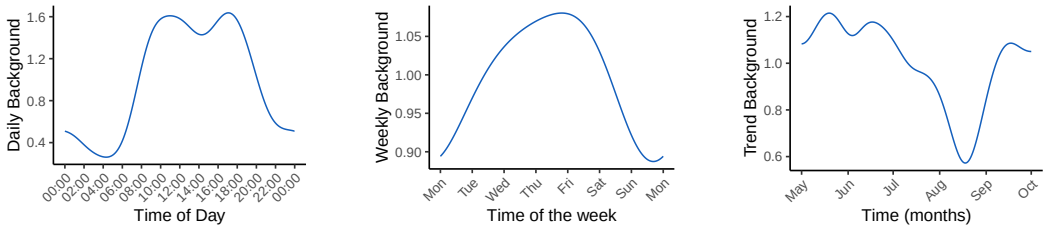
Trend + Periodic + Triggering. Periodic Hawkes model: $\lambda(\boldsymbol{s}, t) = \mu_0 \cdot \mu_{tr}(t)\mu_d(t)\mu_w(t)\mu_s(x, y) + A \cdot J(\boldsymbol{s}, t)$.

All the models are fitted non-parametrically through Algorithm 4 and the resulting estimates and performances depend upon the kernels and bandwidths chosen for the smoothing. Let us point out how all the model variations introduced in Section 5.2.2 (i.e. edge corrections and isotropic excitation) are considered in this application. Appendix B.2 compares the results obtained using the isotropic excitation with those that would have been obtained using the anisotropic one.

The Gaussian kernel form is used for all components, but the corresponding bandwidths must be chosen according to the smoothness degree required by each of those. Generally speaking, automatic selection of the bandwidths in kernel density estimation is still an open-problem. Most approaches are mainly based on heuristics, model-based plug-in estimators, cross-validation methods or Bayesian fitting procedures (Silverman, 1986; Heidenreich et al., 2013).

Given the absence of a valid cross-validation procedure for our model, we here rely mainly on heuristics and graphical inspection of the results for the choice of the bandwidths. These define the impact range of each observation on the estimated intensity at each point of its domain. It shall roughly represent the resolution at which the estimated components can vary. In principles, considering different bandwidths in different models may jeopardize the comparison process (just as much as comparing models exploiting different sets of covariates). Nevertheless, as long as all common components of the four different models share the same bandwidths, this comparison can be considered *fair*. Hence, the bandwidth of each component are fixed throughout all models

For the *spatial background*, we decided for a varying bandwidth approach (as in Zhuang et al. (2004)). A standard bandwidth of $b_s = 0.3$ has been chosen (given that coordinates are recorded in UTM, this amounts to ≈ 300 meters), with the additional constraint that if less than 10 other points are included in its range (≈ 600 meters) then the bandwidth is increased until that condition is met. This is done in order to alleviate the edge effects on points close to the border and also to stabilize the intensity estimation in areas with sparsely observed points. For the temporal components, the application offers natural choices with time-scales inherent to the system and the corresponding resolution requirements. Fixed bandwidths of 7 for the trend, 1 for the weekly component and 0.05 for the daily component, with days as the temporal unit, have been adopted. The trend shall capture variations across different weeks, the weekly component across different days, while the daily component across different hours of the day. It is very important that the chosen bandwidths



(a) Daily Background **(b)** Weekly Background **(c)** Trend Background

Figure 5.4. Estimated Temporal Background components

allow each temporal component to adapt only to variations on its inherent scale, without generating confounding among them. The chosen bandwidths seems to have reasonable ratios to this end. Finally, the temporal and spatial bandwidths for the triggering functions have been chosen to be 0.15 minutes and 0.2 meters respectively. Variations on all of these values have been considered in an initial tuning process, and visual inspection of the resulting estimates found those listed to provide a reasonable compromise of model interpretability and identify known components of traffic flow.

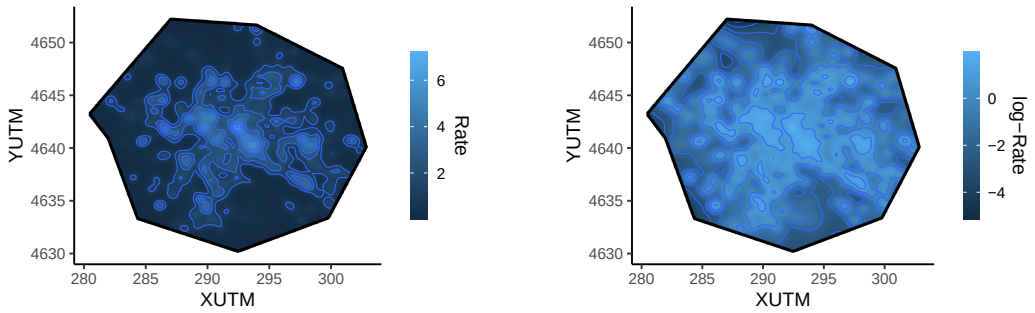
The models are compared in terms of the log-likelihood values, computed as in Equation (5.1.23), with a larger value suggesting a better model. Results for this are given in table 5.1. Note that using the log-likelihood to judge the goodness of fit ignores model complexity: more and more components could be iteratively added to any model and see increasingly marginal improvements as a more complex (free) model is attained. Therefore, as a metric, it is not completely reliable and prone to favor over-fitting. At the current state of the art, cross-validation metrics or penalized methods have not been yet developed in this context. We trust we averted over-fitting issues by selecting bandwidths that produced estimates with a smooth shape. In particular, all these present reasonable and interpretable behaviour that seems to catch marginal but still present features of the phenomenon of interest.

Table 5.2 shows how the model including only spatial background and trend provides the worst fit. The periodic component is the one that brings the greater improvement in terms of model fitting, while the inclusions of triggering effect is able to further improve on the log-likelihood value. It is interesting to notice that if the periodic component is missing, then the triggering effect increases in magnitude. It is probably trying to account for some of the previously unexplained heterogeneity.

The *periodic Hawkes* model is then picked as the best one, and all comments and analyses are referred to it only.

Background Analysis The relaxation coefficient of the background is estimated at $\hat{\mu} = 0.1582$. Integrating it over the whole spatio-temporal area, an overall background magnitude of $\approx \hat{\mu} \cdot |\mathcal{D}| \cdot |\mathcal{T}| = 8055$ is obtained. This, rescaled by the integrated overall intensity $\int_0^T \int_{\mathcal{D}} \hat{\lambda}_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau$, yields a value of 0.94 and therefore explains the 94% of all the accidents occurring in the urban road network of Rome. As expected, it is the major drive of car accidents. In particular, the inclusion of the periodic component improved the log-likelihood of a great amount, highlighting its utmost importance. The estimated temporal background components are shown in Figure 5.4.

Figure 5.4a provided results very close to the expected ones. There is indeed a reducing effect (i.e. rate < 1) during the night-time. The estimated background



(a) Spatial Background **(b)** Log-Spatial Background

Figure 5.5. Estimated Spatial Background Components

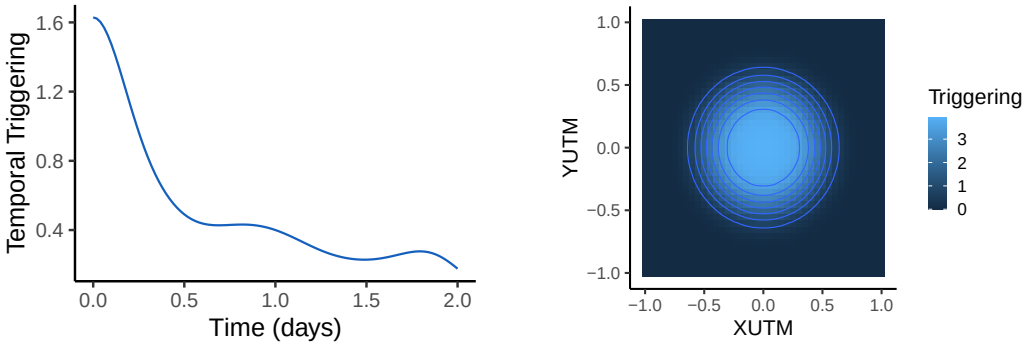
increases in the morning and presents a first peak between 8 : 00 and 12 : 00, morning hours during the ones people get out of home in order to get to work. A second peak is instead really evident between 17 : 00 and 18 : 00, which is well-known by the citizens of Rome as the real rush hour during the one most of the workers head home from their job-place. Figure 5.4b shows the weekly behavior. It presents low values during the weekend, which gradually increase during the week. The peak is observed during Thursday and Friday. This may be explained by a number of reasons, e.g.: work-traffic combined with more people going out for leisure increasing the overall traffic; large number of people going out late during night-time, which is a well-known risk factor for road accidents. This could be a talking point for further analysis intended to identify these properly. Finally, Figure 5.4c shows the overall trend in the considered 5 months period. Except for some slight changes from one month to the other, a first drop in the rate is observed in July and a very evident one during August. Also this behavior is completely reasonable. Rome is famous to get empty during the summer holidays, with off-site students and workers reaching their home-towns and Roman citizens leaving the city. The amount of traffic is very low during these months and, therefore, we also expect a substantially diminished road accident risk.

Figure 5.5 shows the spatial background, with Figure 5.5b representing it on the log-scale in order to highlight spatial variations. The most of the car accidents happen in the inner area, and this may also be due to the increased probability of a police patrol intervention. Nevertheless, there are also some spots in the outer area that present a sensibly higher risk[2]. Experts with deeper knowledge on urban traffic dynamics and the Rome road conditions may be addressed toward inspection of these areas for recognizing possible structural flaws.

Triggering Analysis Each single car accident contributes to the overall integrated intensity of $\approx \hat{A} = 0.0592$. Therefore, with good approximation, the triggering effects explains $\approx 6\%$ of the car accidents occurring in the urban road network of Rome. The resulting functions are visualized in Fig. 5.6.

Figure 5.6b shows clearly how the spatial triggering is limited to a 500 meters radius. Over that range, the triggering effect is estimated to be practically null. The temporal triggering behavior is shown in Figure 5.6a. In its initial section it steeply decreases with time as expected. However, after half a day (≈ 0.5 on the plot scale) the decrease slows down. The function actually gets unnaturally bumpy and decays

[2]We recall that road accidents occurring on the GRA are excluded from the analysis



(a) Temporal Triggering **(b)** Spatial Triggering

Figure 5.6. Estimated excitation functions

to 0 only after two days. The excitation magnitude in this last section is really low as compared to the background components, so this is not worrisome in terms of the model fitting. Nevertheless, some care must be taken in order to interpret properly this result. Indeed, the temporal triggering effect is likely to be finished after (less than) half a day, with the long-tail being just a residual and negligible effect.

Model Validation To validate if the model is actually describing properly the underlying intensity of the point process, one often follows the work of Ogata (1988), Brown et al. (2002) and Zhuang (2006).

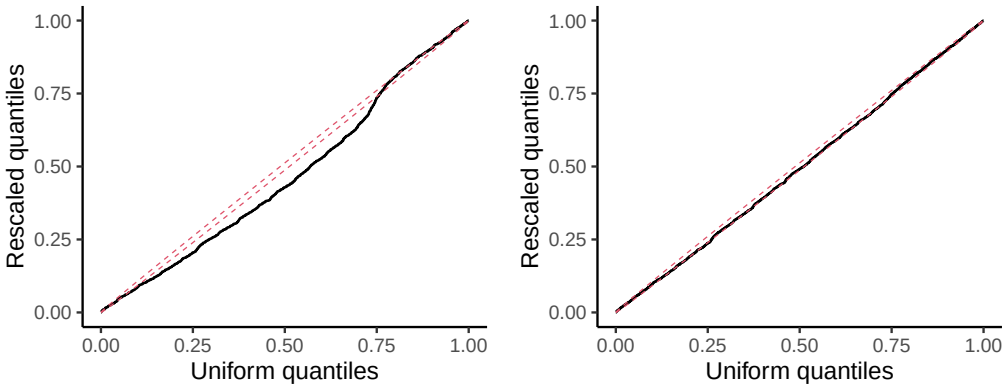
Specifically, the transformed time sequence of the observed pattern:

$$t_i \to v_i = \int_0^{t_i} \int_{\mathcal{D}} \lambda_c(\boldsymbol{\sigma}, \tau) d\boldsymbol{\sigma} d\tau, \quad i = 1, \ldots, n, \tag{5.2.4}$$

is such that the resulting times $\{v_i\}_{i=1}^n$ shall follow a unit rate Poisson process if the model is correctly specified. Once the transformed sequence has been obtained, as in Zhuang and Mateu (2019), one can test its adherence to the unit rate Poisson process in different ways. Times shall be uniformly distributed on the time interval, and this can be tested through the Kolmogorov-Smirnov test (KS-test) or visually checked through for quantile-quantile (QQ) plots. For the latter, it is possible to obtain confidence bands exploiting the fact that the order statistics of a uniform distribution follow a beta distribution[3]. Obviously, this holds exactly when the original times are rescaled for the true intensity. When the intensity is estimated, convergence is slower but still proved to hold (Schoenberg, 2002). The KS-test performed comparing the original times *empirical CDF* with the uniform distribution, rejects the hypothesis with a very low p-value $p < 0.001$. If the same test is performed on the rescaled times, the p-value $p = 0.044$ is still significant but much larger. Validation of the alternative models considered for this application are not discussed in details here, but all yield p-values < 0.01 [4]. Therefore, while not entirely satisfactory, the validation of the picked model is mostly defensible among all the considered alternatives. In particular, looking at Figure 5.7, we can see how the QQ-plot resulting from the original times

[3]The distribution of the k-th order statistic has parameters k and $n + 1 - k$, with n being the number of sample points

[4]Let us point out that this test is not taking into account the additional variability due to the estimation of the intensity (it is not the *true intensity*), and then it is probably returning too severe conclusions



(a) Original times **(b)** Rescaled times

Figure 5.7. Validation results on the original times and rescaled times. Model results are shown with ——; confidence bands at 99% with: - - -.

is (almost) completely outside the confidence bands at the 99% level. On the other hand, all the quantiles of the rescaled times in Figure 5.7b never deviate significantly from the target.

Hence, the chosen model looks really promising and especially improves substantially on standard alternatives. Some additional modeling effort may be needed in order to further improve on the current results (e.g. include unobserved heterogeneity, Network structure, etc.).

5.3 Application to Car-accidents on the M25 London Orbital

The application in the previous section, in principles, deals with car accidents occurred on an urban road network. However, the road network structure is neglected and methods to include it are currently under development. The application in this chapter instead restricts the observed region to a single road: a *smart motorway*. It can be represented as a continuous ring-road, not an urban network, and makes completely reasonable considering a standard distance measure rather than say a graph distance such as the one that should be considered for the Rome car accidents application.

We here want to stress the point that the choice of this dataset is not to simplify the fitting procedure onto continuous space rather than a network, but instead it is to offer some insight into a much discussed topic of smart motorway safety. The United Kingdom has one of the lowest per-capita death rates from traffic accidents in the world, estimated by the World Health Organization (2018) at 3.1 per 100,000 of population in 2016. Nevertheless, 1782 deaths and 25,484 serious injuries resulted from accidents on UK roads in 2018 Department for Transport (2019). Aside from the direct human cost of serious accidents, indirect economic costs result even from relatively minor crashes. This is because crashes, collisions and breakdowns can cause severe congestion leading to significant drops in the efficiency of the road transport network. For these reasons, there is an imperative to further reduce the accident rate on UK roads. However, traffic accidents are rare in absolute terms and are not distributed uniformly across the network. Further rate reductions are therefore likely to require targeted interventions.

Targeted interventions might try to improve safety at specific locations where the accident risk is known to be high compared to the baseline or might try to mitigate against particular mechanisms that are known to account for a significant proportion of accidents. Infrastructure modifications to improve the safety of accident-prone junctions is an example of the first type of intervention. The deployment of the Motorway Incident Detection and Automatic Signalling (MIDAS) system to reduce the number of secondary accidents on motorways is an example of the second. Secondary accidents occur when a driver fails to react appropriately to the disruption caused by an existing accident leading to a subsequent accident upstream of the first one. MIDAS uses a network of induction sensors, known as loops, embedded in the road surface to automatically detect queues and then warn upstream drivers of the danger ahead via roadside signage, and is one component of the UK 'smart motorways' infrastructure, see Highways England (2020) for further details. Most of the data, including data on accidents and congestion, is publicly available via the National Traffic Information Service (NTIS).

In this application, main topic of Kalair et al. (2020), data from NTIS repository are considered and used to model the distribution of motorway car accidents as a spatio-temporal process comprised of a background component and a self-excitation component. The focus is on the M25 London Orbital, one of the busiest motorways in the UK. The objectives of the study are two-fold. The first is to quantify how accident risk on the M25 varies in space and time relative to the baseline. The second is to use the self-excitation component of the process to quantify the likely contribution of secondary accidents to the observed totals. We would like this model to be helpful in addressing the question of how best to target interventions when the baseline accident rate is low in absolute terms. We therefore perform extensive in-sample and out-of-sample validation to verify the models performance on seen and unseen data.

5.3.1 Data Collection and Pre-Processing

The data for this study is taken from the National Traffic Information Service (NTIS)[5]. NTIS provides both historic and real-time traffic data for all roads in the UK that lie on the 'Strategic Road Network.' This network includes all motorways and major A-roads in the UK but we choose to focus our analysis on one of the British busiest motorways, the M25 London Orbital, pictured in Fig. 5.8. Inside NTIS, roads are represented by a directed graph, with edges (referred to as links from now on) being segments of road that have a constant number of lanes and no slip roads joining or leaving. All links that lie on the M25 have been extracted in the clockwise direction, yielding a subset of the road network to collect data for. Placed along links are physical sensors called 'loops', which record passing vehicles and report averaged quantities each minute.

The most relevant components of NTIS to this work are event flags that are manually entered by traffic operators. These flags specify an event type, for example accident or obstruction, the start and end time and the link the event occurred on. All accidents and obstructions that occurred on the chosen links between September 1st 2017 and September 30th 2018 are selected. However, the considered model requires more fine-grained locations for events than just the link it occurred on. Hence, further localization becomes necessary. This pre-processing step is highly specific to this dataset, and hence only a general overview of it is offered here. More

[5]Technical details of the NTIS data feeds are available at `http://www.trafficengland.com/services-info`

Figure 5.8. The M25 London Orbital, roughly 180 kilometres in length. The Dartford crossing, located in the east, is a short segment of road that we do not have data for.

details are instead given in the supplementary material. To understand why this is needed, one should note that links vary in size from around 500 to 10000 meters. However, the average distance between successive loop sensors is roughly 500 meters. As a result, the data is not initially appropriate to model with point-processes, but one can use the time-series provided by individual loop sensors to determine higher resolution location data for events. Given an event flag and link, the event is exactly localized in-between the two loop sensors on that link that show the largest drop in speed and rise in occupancy when going from the sensor downstream to upstream. This is motivated by existing work in traffic accidents' detection, with an example being Payne et al. (1976).

5.3.2 Model variations

There are three main changes. Firstly, the spatial triggering mechanism is one-dimensional and unidirectional: secondary accidents cannot occur downstream of the primary crash. Secondly, monotonicity of the triggering functions, $g_t(\cdot)$ and $g_s(\cdot)$ in Equation (5.1.6), is enforced as a constraint to help with identifiability and discourage over-fitting. Thirdly, boundary correction is applied to the kernel density estimates to reduce bias as for the previous application. This is not detailed here since it has already been illustrated in Section 5.2.2. Let us only add that also the spatial background is here assumed to be periodic, as the M25 is an almost continuous ring around London. Therefore, the same sense of a periodic function having no boundary can be accounted for in the spatial background in the same way, altering Equation (5.1.12). From Fig. 5.8, a small section of the M25 is not a motorway and reports no data, but it is negligible in comparison to the wider motorway, so assuming a spatial background on a ring is reasonable. In the remainder of this section, further detail is provided on these changes. For computational speed, the triggering after 12 hours is set to 0, which is informed by considering time-scales similar to the worst recorded traffic jams in the U.K, detailed in INRIX (2019).

One-Dimensional and Unidirectional Spatial Triggering In this application, the space is not bi-dimensional but a uni-dimensional circle and the traffic only flows in one direction. Thus, the spatial coordinates $\{\boldsymbol{s}_i\}_{i=1}^n$ are no more pairs, but

single values $\{x_i\}_{i=1}^n \in [0, X]$ which also respect an order (according to the flow direction). Therefore, the work in Zhuang and Mateu (2019) must be slightly adapted to reconstruct a uni-dimensional and uni-directional spatial triggering function. The derivations of Section 5.1.2 are modified as follows.

Let us re-define the triggering probabilities of Equation (5.1.17) as:

$$\rho\left(x,t,x',t'\right) = \begin{cases} \frac{Ag_s(x'-x)g_t(t'-t)}{\lambda_c(x',t')} & x < x' \text{and } t < t' \\ 0, & \text{otherwise,} \end{cases}$$

with $\rho(x_i, t_i, x_j, t_j) = \rho_{ij}\ \forall\, i, j$.

Then, the histogram estimate of the spatial excitation can be obtained as:

$$\hat{g_s}(x) \propto \sum_{i,j=1}^{n} \rho_{ij} I_{\Delta x}(x_j - x_i),$$

where $x \in \mathbb{R}^+$. The histogram estimate can then be converted into a smooth version via a suitable kernel function, that in this case shall be uni-dimensional. As for the temporal triggering, the uni-variate Gaussian kernel of Equation (4.1.10) is considered:

$$\hat{g_s}(x) \propto \sum_{i,j=1}^{n} \rho_{i,j} \cdot \widetilde{k}_{h_s}\left(x - (x_i - x_j)\right),$$

where $\widetilde{k}_{h_s}(\cdot)$ denotes the normalized kernel.

Enforcing Triggering Functions monotonicity We note that this model aims to explain any residuals from the background process using the triggering component. As such, the model has significant freedom to adapt to the data. We seek to limit this freedom by ensuring that the triggering component truly reflects increased rates of events on a short-time scale (compared to the periodic components) in the wake of a particular event. A natural way to do this and extension of the original methodology is to enforce the triggering functions to be monotonic. This is a reasonable constraint that shall favor interpretability of the triggering functions, whilst still allowing them to be constructed by the data. Triggering decaying in space and time suggests we expect reduced influence of an event as we move further from its location and as time passes. The work in Hall and Huang (2001) details a practical and sound method to achieve such monotonic estimate. Let us recall again that in standard kernel smoothing one can write the smoothed estimate as:

$$\hat{\nu}(x) = \frac{1}{N} \sum_{i=1}^{N} \widetilde{K}_b\left(x - x_i\right) y_i, \tag{5.3.1}$$

with some normalized kernel $\widetilde{K}$, some bandwidth b and observed values $\{y_i\}_{i=1}^n$ at locations $\{x_i\}_{i=1}^n$. In the context of the triggering functions, the locations are differences in event locations or event times, and all y-values equal 1 (however the technique works for more general cases). Equation (5.3.1) does not place any constraint on the estimate $\hat{\nu}(x)$, however in practice there are many situations where some minimal structure must be enforced.

Hall and Huang (2001) generalized Equation (5.3.1) to incorporate a weight p_i to each data-point used in the smoothing to *'adjust'* the initial smoothed fit to be

monotonic. One writes this adjusted fit as:

$$\hat{\nu}_{mono}(x \,|\, p_1, \ldots, p_N) = \frac{1}{N} \sum_{i=1}^{N} p_i \cdot \widetilde{K_b}\,(x - x_i)\, y_i. \tag{5.3.2}$$

The goal is now to choose a weight p_i for each data-point i used in the construction of the function, whilst altering the original estimate as little as possible. There are an infinite number of sets of $\{p_1, p_2, \ldots, p_N\}$ one could choose to enforce a monotonic function, and to identify a unique solution, the set that is the closest to the uniform distribution $\left\{\frac{1}{N}, \ldots, \frac{1}{N}\right\}$ is chosen. One possible distance measure, which is then used in the real application, to compare the p_i's to a uniform distribution is:

$$D_0(p_1, \ldots, p_N) = -\sum_{i=1}^{N} \log(N p_i). \tag{5.3.3}$$

Adding this step in our model fitting encompasses solving an additional optimization problem, determining each p_i value to produce a monotonic triggering function. The optimization problem is specified as:

$$\begin{aligned} \min_{p_1,\ldots,p_N} \quad & D_0(p_1, \ldots, p_N) \\ \text{s.t.} \quad & \frac{d\hat{\nu}_{mono}(x \,|\, p_1, \ldots, p_N)}{dx} \leq \epsilon \\ & p_i \geq 0 \;\text{ for all } i \in [1, 2, \ldots, N] \\ & p_i \leq 1 \;\text{ for all } i \in [1, 2, \ldots, N] \\ & \sum_{i=1}^{N} p_i = 1. \end{aligned} \tag{5.3.4}$$

Since the triggering is enforced to decay with increasing time and distance, ϵ constraints the gradient of the triggering functions to be below some value.

It should be noted that in practice, constraining the triggering function only alters the resulting functions on the seasonal results and it is probably related to the reduced size of the subsets of data. When fitting to the entire year, the resulting functions are monotonic even without enforcing monotonicity.

5.3.3 Results

The main analysis is reported in this Section. Additional results, obtained on subsets of data and aimed at further verifying the validity of the application, are reported in Section C.3.

Bandwidth Selection, Model Selection & Prevalence of Triggering Same arguments of Section 5.2.3 hold here. The daily, weekly and trend bandwidths are chosen to be 60, 10×60 and $60 \times 24 \times 14$ minutes respectively. The choice of daily bandwidth is selected due to the 'rush-hour' behaviour in the UK typically varying on a time-scale of around an hour, whilst the weekly and trend components capture variation across larger time-scales. The spatial bandwidth is chosen as 5500 meters, which appears small enough to capture differing features across the M25, whilst not introducing superfluous oscillations. This is also larger than the uncertainty we would expect in event localization, therefore accounting for potential

Model	$\hat{A}$	Log-Likelihood
Fixed Rate Poisson Process	-	-28861.55
Daily + Weekly Background	-	-28028.05
Daily + Weekly + Trend Background	-	-27929.60
Daily + Weekly + Triggering	0.068495	-27864.48
Daily + Weekly + Trend + Triggering	0.065462	-27781.38

Table 5.2. Log-likelihood values for models with various components.

uncertainty in the data. Finally, the temporal and spatial bandwidths for the triggering functions are chosen to be 30 minutes and 500 meters respectively. Also here we considered variations on all of these values, finding those listed provided a reasonable compromise of model interpretability and identify known components of traffic flow.

Models containing variations of the discussed components are compared through the corresponding log-likelihood scores given by:

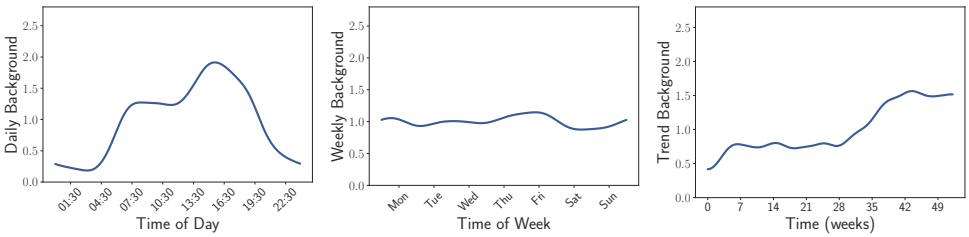
$$\log(L) = \sum_{i=1}^{n} \log\left(\lambda_c(t_i, x_i)\right) - \int_0^T \int_0^X \lambda_c(x,t) dx dt, \tag{5.3.5}$$

with a larger value suggesting a better model. The first is a homogeneous Poisson process, used as the simplest reference model one could construct. The other compared models include: daily and weekly background components, daily, weekly and trend background components, daily, weekly background components and triggering, and daily, weekly, trend background components and triggering. Final results are given in table 5.2. The residuals are inspected in order to further check the specification of the models. The results are then validated in sample for the whole dataset, while we also provide a successful attempt of out-of-sample validation on seasonal subsets.

Table 5.2 shows how the constant rate Poisson process is by far the worst model, and adding periodic daily and weekly components to this implies the largest improvement in log-likelihood. Including a trend and triggering component further improves on this score. The parameter A can be interpreted as the proportion of the impact of the triggering function on the total intensity. For the optimal model incorporating all components, this means about 6.55% of events appear to be the result of triggering, in practical terms about 100 events. One may want to consider this as an upper bound along with an appropriate time-scale.

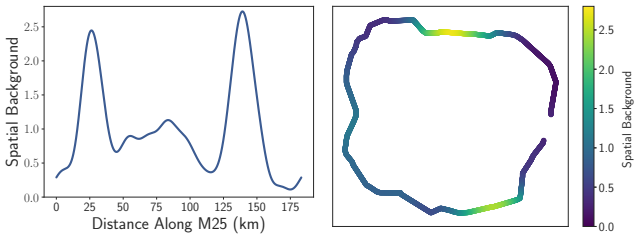
Background Analysis The background component of the model is strong, as seen when inspecting the changes in log-likelihood from table 5.2. In this model this constitutes a daily, weekly, spatial and trend component. All the temporal background components are visualized in Fig. 5.9.

Fig. 5.9a shows that the daily background increases to an initial peak during the morning rush hour, then remains roughly constant, before rising again to a peak at around 4pm, and decaying after this. From Fig. 5.9b, there appears to be much less variation in the intensity across the week compared to all other identified components, but a slightly higher intensity on Thursdays and Fridays, and the lowest on Tuesdays and Saturdays. Finally, we see a small increase in the trend during the first 7 weeks of the data, then it remains reasonably flat until week 28, where it begins to rise again. Around week 40, it stabilizes again. This could be due to an



(a) Daily Background **(b)** Weekly Background **(c)** Trend Background

Figure 5.9. Temporal Background Components fit to 1 year of data.



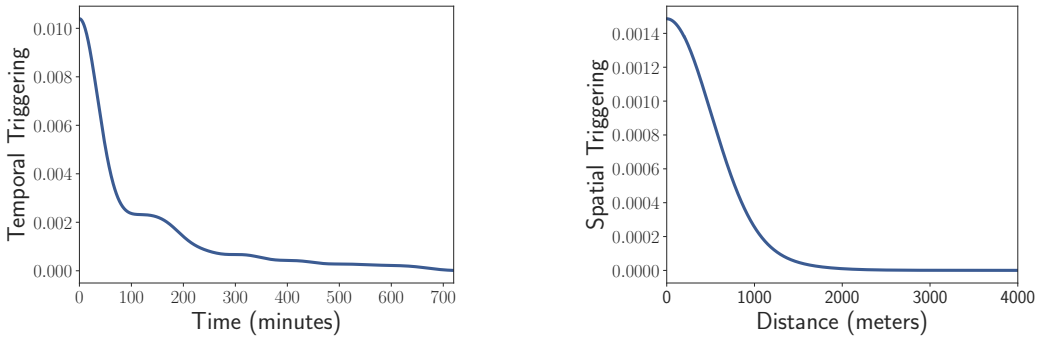
(a) Spatial Background **(b)** Spatial Background overlaid on a map of the M25.

Figure 5.10. Spatial Background Components fit to 1 year of data.

increase in actual event intensity, or perhaps to a more comprehensive reporting after a certain point, a larger number of operators and so forth. That said, no changes in reporting are known to us.

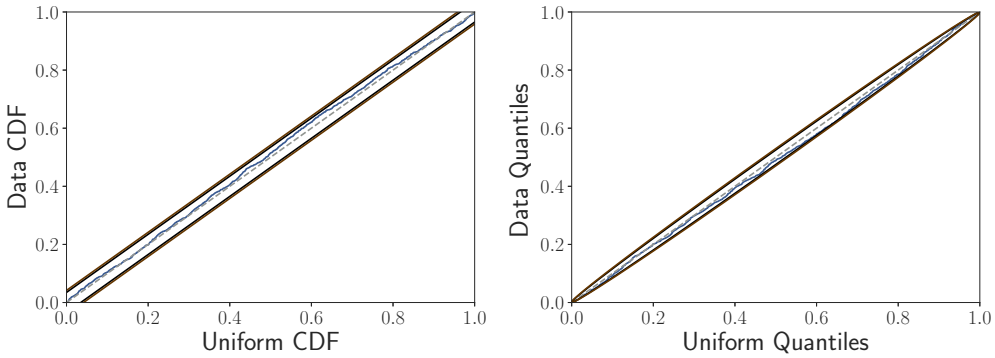
As-well as the temporal background, also *where* events are most common around the M25 is of utter interest. The spatial background visualized in Fig. 5.10 indicates a clear spatial structure. There are are indeed two distinct peaks in Fig. 5.10a, and a smaller spread out peak in-between these two. The largest peak, around 140 kilometres along the motorway, is located near the 'Potters Bar' junction, infamous location where multiple secondary roads converge onto the M25 and cause much traffic distress. The second largest, around 25 kilometres into the motorway, is located between where the M25 meets the M26, and where the M25 meets the M23. This background intensity, imposed onto a map schematic of the real M25 is plotted in Fig. 5.10b. In particular, we investigated if the temporal background in the vicinity of these spatial peaks differed to that across the entire M25, but found only minor changes. Interested readers can find this analysis in the supplementary material.

Triggering Analysis Triggering does appear to improve the log-likelihood of the model and, as clear from Table 5.2, it explains around 6.55% of the events in the data. The resulting functions are visualized in Fig. 5.11. From Fig. 5.11b, it is clear that spatial triggering is limited to around 2-kilometres, after which non-zero values are absent. However, the temporal triggering pictured in Fig. 5.11a appears to have quite a 'long tail' in the sense that it decays over a very long range. However, there is a clear time-scale in this result of around 100 minutes. As a result of this, one should take 6.55% as an upper bound of sorts. Combining this information with the identified time and length scales is a very informative conclusion to draw.



(a) Temporal Triggering **(b)** Spatial Triggering

Figure 5.11. Triggering functions, fit to 1 year of data.



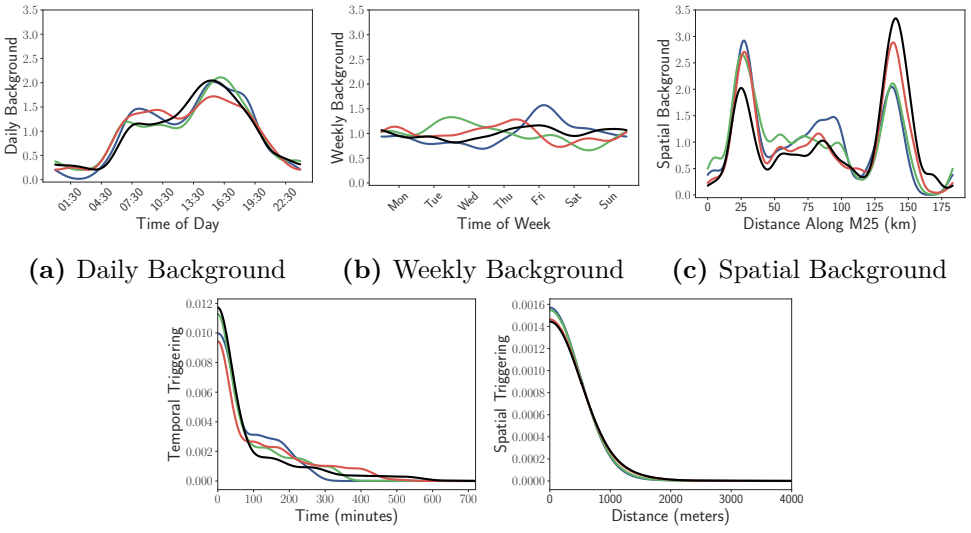
(a) CDF Validation **(b)** Quantile-Quantile Validation

Figure 5.12. Validation results for 1 year of data. Model results are shown with —, 95% limits: —, 99% limits: — and a reference line: - - -.

Model Validation The validation of the estimated model is again based on the inter-time intervals $\{\widetilde{v}_i = v_{i+1} - v_i,\ i = 1, \ldots, n-1\}$. As for the previous application, these are supposed to follow an exponential distribution with parameter 1 and the validation will be based in this hypothesis as in Brown et al. (2002). Given that the expected distribution of the $\widetilde{v}_i$ is known, compare empirical and theoretical cumulative distribution functions (CDF) can be compared and tested for significant differences, as-well as the observed and expected quantiles. One can generate confidence bounds for the comparison of two CDFs by inversion of the Kolmogorov-Smirnov statistic. Furthermore, these can be back-transformed to the Uniform distribution using the *inverse-transform theorem*:

$$z_i = 1 - e^{-(\Lambda_i - \Lambda_{i-1})}, \tag{5.3.6}$$

, so that the Beta bands for the *qqplot* can be used. Figures 5.12a and 5.12b show the transformed inter-times sequence as compared with the target CDF and quantiles. It is clear from Fig. 5.12 that the model is statistically defensible when inspecting both the CDF and QQ plots of the results. Some of the quantiles in Fig. 5.12b are just on the edge of acceptable, but do not deviate outside of the confidence bands. These results show that the model seems well specified, and correctly recognizes the underling spatio-temporal intensity behavior of car crashes.



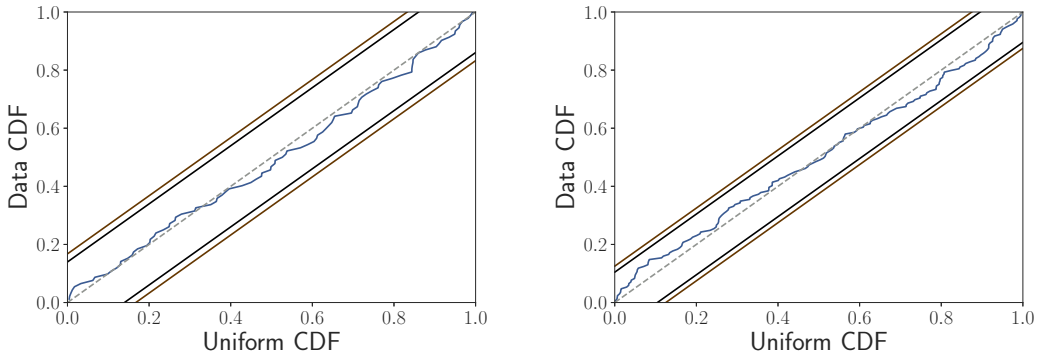
(a) Daily Background **(b)** Weekly Background **(c)** Spatial Background

(d) Temporal Triggering **(e)** Spatial Triggering

Figure 5.13. Background and triggering components compared across different 3-months datasets. Datasets are: 09/2017-11/2017: —, 12/2017-02/2018: —, 03/2018-05/2018: — and 06/2018-08/2018: —.

Do Components Change with Season? Given the results for the year of data, one can question how resilient the background and triggering components are by inspecting subsets of the data. To test this, data are partitioned into 3-months seasonal periods, and the model is separately fit to each subset. The estimated components are then overlayed in Fig. 5.13. It is clear from our results in Fig. 5.13 that there are varying amounts of consistency in components of the model. Generally, Fig. 5.13a shows that the daily background component is constant throughout the year, the only variation being in the 3-months of data starting on 03/2018, where the morning peak is a little more pronounced than any of the other datasets, and the evening is a little less. It is difficult to make conclusions about the weekly component, when only 12 instances of each day of the week for a given 3-month subset are available. Of particular interest is the spatial background through time, shown in Fig. 5.13c, where the peak around 140 kilometres along the M25 actually becomes more pronounced as time progresses. Both the peak at 25 and 140 kilometres are present throughout all subsets of data, but later periods appear to show that the peak around Potters Bar is more significant later in the dataset compared to earlier. It is unclear if a physical change occurred leading to this, but would be of interest to investigate with more data. Finally, the temporal and spatial triggering functions are generally quite consistent across all datasets. A somehow long decay is still present in the temporal triggering, but reasonable time-scales of around 100 minutes remain evident. Considering the triggering in each of these subsets, A values of 0.034, 0.063, 0.068 and 0.075 are attained for the periods starting 2017/09, 2017/12, 2018/03 and 2018/06, respectively. More data would be necessary in order to verify if these values vary periodically throughout the year, or if they are increasing over time.

We note that, for small temporal subsets of data, the trend component becomes less important in the model. Indeed, Fig. 5.9c shows how the trend is almost flat for long periods, suggesting we can omit it for some temporal subsets. As a result, a sort of out-of-sample model validation can be performed, something we are not



(a) Out of sample CDF Validation, fitted to data 12/2017-02/2018, validated on data for 03/2018.

(b) Out of sample CDF Validation, fitted to data 06/2018-08/2018, validated on data for 09/2018.

Figure 5.14. Out of sample model validation for two datasets. Model results are shown with —, 95% limits: —, 99% limits: — and a reference line: - - -.

aware of being done previously in the literature for this type of model. Once the trend is omitted, all the other components are assumed to be constant through times. Then, if such consistency actually exists, the model trained on some 3-months subset of data shall be transferable to any other 3-months subset (excluded from the training process). Rescaling the times of the out-of-sample data shall provide us with a validation metric on unseen data. Fig. 5.14a and 5.14b shows two examples of this procedure. It is clear from both images in Fig. 5.14 that on short time-scales, the model does provide satisfactory performance out-of-sample, and hence is able to catch inherent features of the road accidents cascading dynamics. It is not possible however to apply this exact method on longer time-scales due to the clear trend and varying spatial background, unless robust parametric could be derived for this components so that extrapolation would be feasible.

5.4 Conclusions and further developments

We have argued how the Spatio-Temporal Hawkes process with periodic background can be profitably fitted to road accidents data in order to detect and inspect the magnitude of primary and secondary (i.e. triggered) road accidents. The methodological framework is based on the work by Zhuang and Mateu (2019), which originally introduced the semi-parametric stochastic reconstruction method as a fitting procedure for a version of the aforementioned model, applying it to crimes and retaliation crimes. In this chapter, we analyzed two different data-sets about road accidents and introduced some innovations in the model specification and in the fitting procedure of both.

We first performed a preliminary analysis of the spatio-temporal intensity of road accidents on the urban network of the city of Rome, between May and September 2017. We wanted to test if some of the inhomogeneous and clustering behavior exhibited by the point pattern could be explained by periodic components and/or a Hawkes excitation mechanism. The comparison of models with increasing complexity highlights the presence of both a strong periodic component and a non-negligible triggering effect. The obtained periodic estimates seem to well-represent the typical routine of the average Rome inhabitant: generalized lower levels during night-time

and higher level during day-time, with the latter showing two peaks at the typical Roman rush hours; weekly effect that sharply distinguishes week-days from week-ends. Finally, the general trend indicates how road accidents are way less frequent during the month of July and, even less, of August. This is not surprising at all, given the greatly reduced presence of people (and traffic) during such times of summer holidays. The spatial background shows that most of the road accidents occur in the central area, but some others non-negligible peripheral seem to appear. We found that approximately the $5-6\%$ of the road accidents have been triggered by others. This does not take into account the gravity of the triggering event, while it is likely that the mildest urban road accidents have a little, if any, triggering ability. With careful interpretation, reasonable time and length scales can be derived from the estimated temporal and spatial range of the excitation: half a day for the temporal triggering, and 500 meters for the spatial triggering.

The second application analyzed the spatio-temporal variation in the accident rate on London's M25 motorway over a period of one year, with the final aim of distinguishing between primary and secondary events. This variation is found to be strongly inhomogeneous. The temporal variation shows a strong daily double peak structure reflecting commuting patterns superimposed in a weaker weekly variation with a peak on Fridays and a trough on Saturdays. This pattern of temporal variation remains stable over the data period. The spatial variation shows two primary peaks in intensity. The first and largest is in the vicinity of the Potters Bar Interchange. The other is in the vicinity of Junctions 5 and 6 where the M26 and M23 join the M25. The peak at Potters Bar appears to increase in intensity during the data period and is more pronounced when we condition on the most significant events in terms of impact on traffic speed. We found that $6-7\%$ of the observed road accidents are most probably secondary events under the assumptions of our model. Plausible time and length scales emerge for the range of the triggering effects: 100 minutes in the temporal triggering, and 1 kilometer for spatial triggering. From these figures, we conclude that the effects of secondary events is a small but detectable feature of the M25 traffic accidents data set. Hence, we suggest that, on the M25, the main attention for the reduction of traffic accident rates shall be addressed toward accurate policies for the specific times or 'hot-spot' locations showing large intensity peaks. In this application, the existing work has been extended to limit the freedom of the self-excitation components, along with considering unidirectional spatial triggering. We further proposed an out-of-sample validation method which is viable when the trend component is roughly constant over time. Applying this to our dataset showed we have avoided over-fitting the model, despite its significant freedom.

All the introduced advances are not specific to the application discussed here and can be used when models of this form are applied to different domains with similar practical considerations. Generally speaking, the modeling framework is novel in the context of traffic crash analysis, and the methodological refinements were addressed toward ensuring that the model captured all the relevant properties of the application.

In terms of future work, it would be very useful to adapt the methodology to fit properly point patterns evolving on road networks. We trust that the planar surface misspecification may be the main reason why validation on the Rome accident data was not completely successful. However, directly extending our methodology to account for this structure appears difficult. Whilst smoothing of point-processes on a network has been explored recently, for example in Moradi and Mateu (2019), further complication remains regarding smoothing of the triggering functions on a network. When we take data-points in $\mathbb{R}^N$ and compute distances between them, we

attain a value in $\mathbb{R}^+$, and differences between these distances lie in $\mathbb{R}$. However, the same is not necessarily true for differences between distances computed on a network, which still attain real values, but belong to the sub-space stemming from their network-constrained spatial structure. Taking this into account makes smoothing the triggering functions complex and deserves further attention in future work. The adoption of semi-parametric forms (polynomial, splines, etc.) is currently under consideration. Additionally, it would also be very informative to repeat the analysis for a major UK road without the MIDAS system or smart motorway features, since we expect this infrastructure to reduce the risk of secondary events (all other things equal). One could also extend the out-of-sample validation to include a trend component. To do so, a parametric form of the trend is required, as the the current non-parametric estimate cannot be extrapolated out of the fitting window. Finally, one could think about the potential incorporation of additional covariates that may influence traffic accident occurrence (e.g. road and weather conditions) or their triggering ability (e.g. crash severity).

Chapter 6

Final discussion

This thesis covered different topics in the spatial and spatio-temporal modeling context. It initially provided a historical introduction to the motivations that raised the need for the development of probabilistic and statistical tools able to account for the temporal and spatial distribution of phenomena. In particular, we highlighted how these tools and methods have been undergoing great changes during the last decades, dictated by the evolution of data collection tools that modified temporal and geo referenced datasets both in terms of accuracy, frequency, and size. We focused on two of the many branches of spatial statistics: continuous spatial variations and point patterns.

Chapter 2 zoomed in on the modeling of outcomes varying continuously over some space but sampled only at a finite number of locations. We introduced and motivated spatial interpolation and the geo-statistical approach, both in the spatial and spatio-temporal context, providing estimation methods and tools from the context of *Spline Regression* and *Bayesian Hierarchical Modeling* of Gaussian processes. These two approaches are usually seen as antithetic, but they can actually be profitably combined if due care is taken. Chapter 3 indeed presents an example in which splines are used to model geographical variations and Gaussian processes to drive temporal dependence, in a context where the two components are actually separated at a physical level and could be independently identified. The same work is currently published (pre-print) in Alaimo Di Loro et al. (2021), while undergoing the review process. The disentanglement of the two effects appears to be completed successfully, providing satisfying predictive performances and interpretable (reasonable) conclusions on the behavior of both. Moreover, the same application dealt with a Big-Data problem. Hence, the strategies introduced in Section 2.5 have been profitably adapted and developed to overcome all the corresponding computational issues (Datta et al., 2016a). Improvements both in predictive performance and computational saving have been validated and verified, both on simulated and real data. The proposed model is novel in the context of physical activity analysis, and can potentially be a game-changer for the new track and monitoring technologies. Indeed, it provides a decently efficient and precise tool to impute gaps and predict activity levels at unobserved times and locations. No-cost tracking devices such as smartphones currently present the problem of sampling at non-constant frequencies and with (sometimes) low accuracy. Integrating these data with a reliable model can really ease future data collection for large studies in a free-living environment.

Chapter 4 concentrated on the modeling of point patterns recorded on continuous domains. We introduced some of the most common and useful summary statistics (both theoretical and empirical) to describe and highlight the relevant properties of

the underlying point process. We further introduced some of the most basic and common probabilistic models for point patterns, along with some of their properties. After that, Section 4.3 focused on the Hawkes process and its extension to the spatio-temporal context. We described its defining properties and gave an ex-cursus of the current state of the art for its fitting and estimation in a spatio-temporal setting. Chapter 5 presents a particular spatio-temporal Hawkes model, characterized by a multiple periodic component over time. It revealed itself very useful to model phenomena that are likely to present a cyclic pattern. We first introduced its original structure and estimation procedure as it is in the seminal paper Zhuang and Mateu (2019) and then motivated its utilization in the context of road accidents. Section 5.2 presented a preliminary application on the Rome urban road network, while Section 5.3 a conclusive application to road accidents that occurred on the M25 London Orbital (Kalair et al., 2020). In both cases, we presented some original adaptation of the methodology to the problem at hand, that proved to be useful in order to make the model results more stable, robust, and reliable. The first application included some useful edge correction (also adopted in the other application) and an alternative specification of the spatial excitation function enforcing isotropic-ness. Surprisingly, the latter resulted in an improved model fit even if it is a less-parametrized model (see Appendix B.2). This has been justified by considering that the simpler structure of the resulting log-likelihood favors the individuation of the actual maxima. However, in the current state, the model neglects the network geometry of the underlying space and issues in the final validation are likely to be caused by this. Current and future work is aimed at solving this. In the second application, the adaptation concerns the uni-dimensionality and directionality of the spatial dimension, being it a single stream continuous ring road. Furthermore, we also proposed a variation on the estimation procedure that forces the excitation component to decrease monotonically. In the considered non-parametric estimation framework of the excitation function, this alleviated over-fitting issues while favoring interpretability. The model is indeed validated successfully, both in-sample and out-of-sample (with a novel method), proving its ability to actually catch the relevant features of the data.

To conclude, all the presented applications showed how proper consideration of temporal and spatial components can sensibly boost modeling efforts at different levels. It indeed improves not only the predictive performances but most importantly on the reliable and accurate explanation of the data generation mechanism. This allows validating our results with common sense and scientific knowledge and understanding of the analyzed phenomenon. This last point is yet what makes statistical analysis superior with respect to some of the ever more (and often carelessly) used machine learning algorithms, that are lately trying to colonize also the domain of spatial statistics. These have proved indeed effective in extracting information from geo-referenced data, but their rules and conclusions are mostly based on inductive reasoning. The statistical methods instead rely on deductive reasoning, where theory and hypotheses are verified by collecting and testing data. In this way, actual knowledge can be distilled from the information contained in the data, allowing to anticipate the un-observed (or rarely observed) by incorporating it in a common theoretical framework. The dimension and complexity of the contemporaneous spatio-temporal datasets cannot be an excuse to neglect some of the most relevant and important information we have about an outcome or resort to less insightful techniques. The current tools and methods of spatial statistics must evolve (and are evolving) to deal with these new data-sources, and this evolution has to be pushed and favored, trying as much to keep the pace of the relentless technological development of our era.

Appendix A

Temporal NNGP and the Collapsed algorithm

A.1 Additional simulation experiments

We carried out two additional experiments to test the reliability of our algorithm and verify comparative performances with the Sequential NNGP as it is implemented in the `spNNGP` package (Finley et al., 2017a). We did not consider the Response NNGP because it does not recover the latent component. The first one is described in Section A.1.1 and includes simulated observations for one single individual; the second one includes simulated observations for multiple individuals and is described in Section A.1.2. Codes to reproduce the following results and additional comparative analyses of NNGP versus the full GP model are available at `https://github.com/minmar94/EfficientTNNGPforActigraph`.

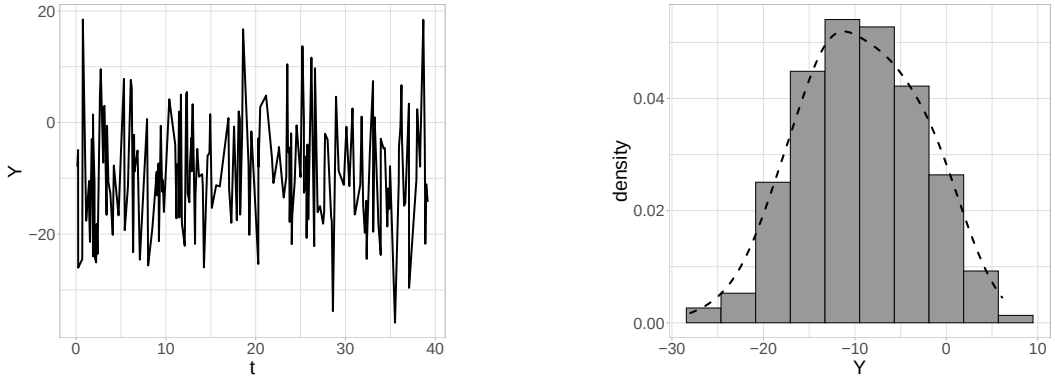
A.1.1 Experiment 1

We generated observations $\{y(t_j)\}_{j=1}^{\top}$ for $K = 1$ individual, using $T = 10^5$ time-points, where each $t_i = \sum_{h=1}^{i-1} \delta_h$, and $\delta_h \sim Exp(5)$, $\forall\, h$. The model included an intercept β_0 and 3 covariates, x_1, x_2 and x_3 all drawn from $\mathcal{N}(0, 1)$, with associated slopes β_1, β_2 and β_3. We modeled the covariance structure between any two simulations at time-points t and t' using the exponential covariance function:

$$\mathbf{Cov}_\theta \left[Y(t), Y(t')\right] = c_\theta(t, t') = \sigma^2 e^{(\phi|t-t'|)}, \quad \sigma^2, \phi \in \mathbb{R}^+, \tag{A.1.1}$$

where σ^2 represents the variance of the process (sill), ϕ is the decay in temporal correlation (range) and τ^2 the residual variance (nugget). In this data generation step the parameters have been set to the following values: $\beta_0 = -1.878$, $\beta_1 = 0.326$, $\beta_2 = -0.302$, $\beta_3 = 1.182$, $\sigma^2 = \phi = \tau^2 = 1$. A chunk of the simulated trajectory and its density can be observed as an example in Figures A.1a and A.1b, respectively.

We fitted the model on the simulated data using our *Collapsed NNGP* implementation, specifically optimized for the temporal setting, while fitting the *Sequential* NNGP using the `spNNGP` package. The latter, while generally used for fitting spatial (i.e. two-dimensionals) models, can be adapted to the temporal (uni-dimensional) case by providing a set of locations where t is one of the coordinates and the other is fixed to a constant value (e.g. $\{\widetilde{\boldsymbol{s}}_j\}_{j=1}^{\top} = \{(t_j, 0)\}_{j=1}^{\top}$). In our implementation, the intercept and slope regression parameters were given a vague normal prior distribution $\mathcal{N}(0, 10^6)$. The variance components, σ^2 and τ^2, were both assigned an inverse



(a) **(b)**

Figure A.1. Simulated uni-dimensional Gaussian process (a) and its density (b).

Param. (True)	**Collapsed NNGP**			**Sequential NNGP**		
	Point	Interval	ESS	Point	Interval	ESS
β_0 (−1.88)	-1.87	(-1.89, -1.85)	4999	-1.87	(-1.89, -1.85)	57
β_1 (0.33)	0.33	(0.32, 0.34)	4999	0.33	(0.32, 0.34)	1285
β_2 (−0.30)	-0.30	(-0.31, -0.29)	4999	-0.30	(-0.31, -0.3)	1365
β_3 (1.18)	1.18	(1.17, 1.19)	4999	1.18	(1.17, 1.19)	1342
σ^2 (1)	1.00	(0.97, 1.03)	472	1.00	(0.97, 1.03)	294
ϕ (1)	0.99	(0.95, 1.04)	496	0.99	(0.95, 1.04)	65
τ^2 (1)	1.01	(0.99, 1.03)	457	1.01	(0.99, 1.03)	165
Metric	Out-of-sample	In-sample		Out-of-sample	In-sample	
Coverage	0.96	0.99		0.95	0.99	
RMSPE (r)	0.39 (1.19)	0.20 (0.85)		0.39 (1.19)	0.20 (0.85)	
PIW	4.68	4.46		4.68	4.46	
Run time (h)	1.77			1.86		

Table A.1. Parameter estimates, predictive validation and fitting times (hours) on the simulated dataset for all the considered models.

Gamma prior $\mathcal{IG}(2,2)$, and the decay parameter ϕ was ascribed a $\Gamma(1,1)$. On the other hand, the `spNNGP` assumes a flat prior on the intercept and slope coefficients and a uniforma $\mathcal{U}(a,b)$ prior on the decay parameter ϕ. In this experiment we fixed $a = 0.5$ and $b = 30$. All the models were trained on the same random sample composed of the 70% of the total observations, while the the remaining 30% have been excluded to assess the out-of-sample predictive performances in terms *Relative Mean Squared Prediction Error* (RMSPE), *Root Mean Squared Prediction Error* (rMSPE), *Coverage*, *Predictive Interval Width* (PIW). We ran the 10000 MCMC iterations, fixing the number of neighbours $m = 10$. The first 5000 simulations have been dropped as burn-in, while the last 5000 have been retained for estimation and prediction purposes. No thinning has been considered. Results are summarized in Table A.1. The two approaches provide identical outputs, both in terms of estimation and prediction. However, our implementation is faster than its competitor (at least in the context of the temporal setting) and provides way better performances in terms of Effective Sample Size (ESS).

$T \times 10^3$	Algorithm	Min	q_{025}	Median	Mean	q_{975}	Max
1	Collapsed	0.01	0.01	0.02	0.02	0.03	0.03
	Sequential	0.12	0.12	0.13	0.14	0.18	0.18
2	Collapsed	0.03	0.03	0.03	0.04	0.06	0.07
	Sequential	0.25	0.25	0.26	0.27	0.31	0.34
4	Collapsed	0.06	0.06	0.07	0.09	0.16	0.16
	Sequential	0.50	0.50	0.52	0.64	1.21	1.21
8	Collapsed	0.13	0.14	0.30	0.26	0.32	0.41
	Sequential	1.01	1.01	2.34	1.99	2.41	2.56
16	Collapsed	0.27	0.28	0.60	0.46	0.63	0.64
	Sequential	2.02	2.02	4.65	3.46	4.75	4.77
32	Collapsed	0.55	0.56	1.23	1.17	1.28	1.37
	Sequential	4.08	4.09	9.40	8.87	9.59	10.16
64	Collapsed	1.01	1.03	2.46	1.90	2.79	2.85
	Sequential	2.51	7.49	18.71	14.43	20.13	20.69
100	Collapsed	1.60	1.61	1.67	1.68	1.87	1.99
	Sequential	11.68	11.74	11.93	12.01	12.80	13.87

Table A.2. Time (in seconds) of one MCMC iteration for the two considered algorithms with increasing sample size (T) and fixed $m = 30$.

Computation time evaluation

In order to delve more into the computational aspect, we *quantified* the *linearity* of all the algorithms: by construction, the fitting time should increase linearly with the sample size. We split observations in $l = 1, \dots, 5$ different fitting windows $\{t_1, \dots, t_{T_l}\}$ with increasing sizes $T_l = \{\{2^l\}_{l=0}^{6} \cup \{100\}\} \times 10^3$ and the computation time of one sampler iteration, fixing $m = 30$ for all the considered algorithms has been recorded for $\widetilde{M} = 100$ times. Figure A.2 shows that all algorithms scale linearly with the sample size. However, our implementation of the collapsed NNGP, whilst pointed out as generally less efficient than its competitors in Finley et al. (2019), scales with a rate of $\approx 0.376 \cdot 10^{-4}$ per data-point, while the Sequential NNGP scale with a rate equal to $\approx 4.5736 \cdot 10^{-4}$, which is sensibly higher.

For an exact numerical analysis, results are reported in Table A.2.

Additionally, we also wanted to quantify the computational advantage of our NNGP-based collapsed algorithm over the standard MCMC update of the full GP model (Cressie and Wikle, 2015), where the latter is already implemented in `R` through `spLM()` function of the `spBayes` package (Finley et al., 2013). Since Datta et al. (2016a) proved that the NNGP approximation with 30 neighbours provides almost exactly the same inference of the full GP, we fixed $m = 30$ and built again different sets of data with increasing number of data-points, but this times with $T_l = 100, 1000, 5000, 10000$ (sizes have been reduced to comply with the slow update of the Full GP). Results, which are summarized in Figure A.3, showed that for $T_l = 100$ the computational time difference between the full GP and the collapsed NNGP is negligible. However, as the size increases, the saving of time increases exponentially: 15 seconds (per iteration) when $n = 5000$, 122 seconds per iteration when $n = 10000$. For the last scenario, which is the more realistic in a MCMC inference context, this means that the collapsed NNGP will provide us with the same results 14 days in advance with respect to the Full GP model.

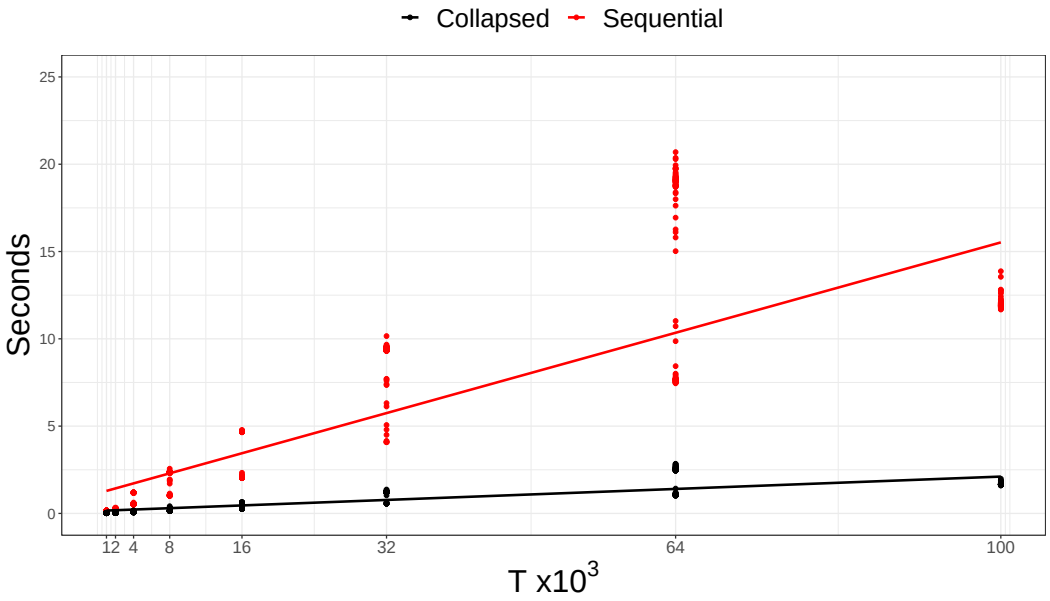


Figure A.2. Time elapsed (in seconds) for 1 MCMC iteration for the two considered algorithms with increasing sample size T and fixed $m = 30$.

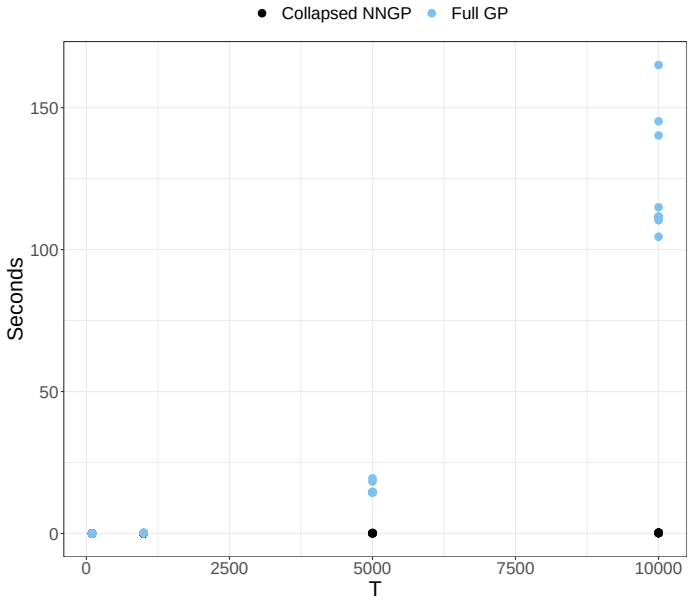


Figure A.3. Time elapsed (in seconds) for 1 MCMC iteration for the Collapsed NNGP and the Full GP with increasing sample size T.

A.1.2 Experiment 2

The aim of this experiment is to verify the ability of our algorithm in recovering the true parameters and to determine if pooling information from multiple individuals can help in improving the accuracy of the estimates. Comparison with the *Sequential NNGP* is not feasible, since it does not allow the contemporary fitting of multiple Gaussian processes with common parameters. Thus, we compare performances of the Pooled NNGP (that's how we will refer to the proposed collapsed in what follows) with the single models estimated separately for each individual.

We generated $2 \cdot 10^4$ observations for $K = 5$ individuals, using the same scheme of Experiment 1 (total of 10^5 data-points). Results are presented in Table A.3. The model also included 3 covariates and an intercept for each individual drawn from independent $N(0, 1)$. Observations were then generated as described in Section A.1.1. The simulated data was split into two sets: 70% composed the train set for estimation purposes, while the remaining 30% was used to asses model predictive performances. RMSPE, coverage of the predictive 95% credible intervals and their mean width were used as measures of the goodness of fit. For all the models, the intercept and slope regression parameters were given a flat normal prior distribution $\mathcal{N}(0, 10^6)$. The variance components, σ^2 and τ^2, were both assigned an inverse Gamma $\mathcal{IG}(2, 2)$ priors, and the decay parameter ϕ received a Gamma prior $\Gamma(1, 1)$. Visual diagnostic check (running mean, traceplots, autocorrelation) is performed and results are available in the supplementary materials (not added yet). The advantage of pooling information from multiple individuals for the estimation of common parameters, while the independence assumption among them still holds, is evident according to all criteria. First of all, there is a sensible gain in the estimation accuracy of the common parameters. Indeed, while the true value of the parameters are included in the intervals also considering one single individual at a time, the widths of 95% credible intervals are sensibly smaller when we pool information together. Furthermore, some slight advantage is also visible for prediction purposes, where the Pooled NNGP provides larger coverage and smaller RMSPE. Additionally, thanks to parallelization of the code, there is almost no loss in terms of the computational time required for the fitting: ≈ 40 minutes to fit one individual VS ≈ 55 minutes to fit the pooled model.

A.2 DAG-based approximation of additive Gaussian process for spatio-temporal modeling

The proposed methodology has been devised under the pressing request of being efficient despite the the huge amount of data and the potentially very complex dependence structure among them. This precluded us from considering since the beginning an ad-hoc spatio-temporal dependence structure that would be coherent with the nature and generative process of the data (for the reasons exposed in Section 3.2.1. For this reason, the spatial effect has been unloaded on the mean term in Section 3.2.4, keeping a simple and easily manageable form of the covariance structure.

In this Section, we propose an alternative model that replaces the spatial spline terms with a more flexible additive Gaussian process. It is only an attempt to propose a viable model and devise an efficient estimation procedure that, if proved reliable, can be a direct future development of the currently adopted methodology.

		Individuals				
Param.(true)	Pooled NNGP	1	2	3	4	5
β_{01}(-9.39)	-9.41 (-9.46, -9.37)	-9.41 (-9.46, -9.37)				
β_{02}(1.63)	1.59 (1.54, 1.64)		1.59 (1.54, 1.64)			
β_{03}(-1.51)	-1.53 (-1.57, -1.48)			-1.54 (-1.58, -1.49)		
β_{04}(5.91)	5.91 (5.86, 5.96)				5.91 (5.86, 5.96)	
β_{05}(-0.92)	-0.80 (-0.85, -0.76)					-0.81 (-0.86, -0.76)
β_1(6.48)	6.48 (6.47, 6.49)	6.48(6.46,6.5)	6.48(6.46,6.50)	6.48(6.46,6.50)	6.46(6.44,6.48)	6.49(6.47,6.51)
β_2(6.76)	6.75 (6.74, 6.76)	6.74(6.72,6.76)	6.75(6.74,6.77)	6.75(6.73,6.77)	6.75(6.73,6.77)	6.76(6.74,6.78)
β_3(-1.46)	-1.46 (-1.47, -1.45)	-1.47(-1.49,-1.45)	-1.48(-1.50,-1.46)	-1.47(-1.48,-1.45)	-1.45(-1.47,-1.43)	-1.47(-1.49,-1.45)
σ^2(1)	0.98 (0.96, 1.01)	0.94(0.88,1.00)	1.02(0.953,1.085)	0.93(0.88,0.99)	1.03(0.97,1.1)	1.01(0.94,1.08)
ϕ(1)	1.01 (0.97, 1.06)	1.00(0.91,1.11)	1.03(0.93,1.14)	1.06(0.95,1.17)	0.94(0.85,1.04)	0.99(0.9,1.11)
τ^2(1)	1 (0.99, 1.02)	1.04(1.00,1.08)	0.98(0.94,1.02)	1.00(0.967,1.04)	1.00(0.96,1.04)	0.99(0.95,1.03)
Coverage	0.95 (0.99)	0.95 (0.99)	0.95 (0.99)	0.95 (0.99)	0.96 (0.99)	0.95 (0.99)
RMSPE	0.012 (0.006)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
rMSPE	1.22 (0.84)	1.24 (0.83)	1.24 (0.83)	1.22 (0.85)	1.22 (0.85)	1.23 (0.84)
PIW	4.67 (4.44)	4.99 (4.44)	4.94 (4.43)	4.93 (4.43)	4.95 (4.44)	4.94 (4.44)
Fitting time	0.59	0.33	0.30	0.35	0.34	0.37

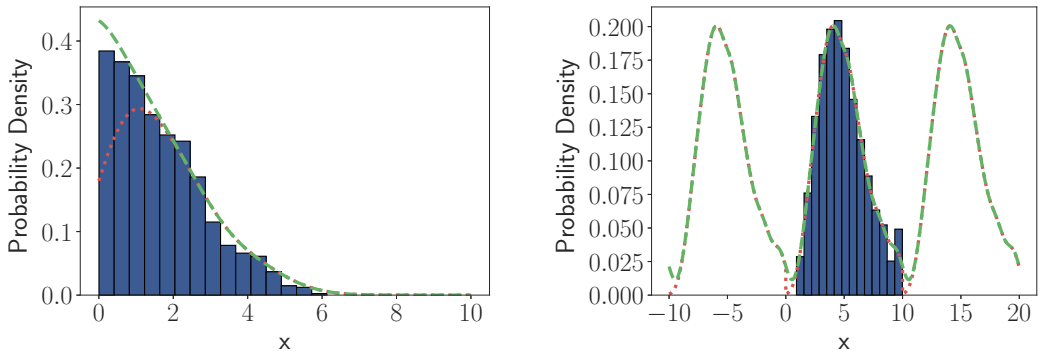
Table A.3. Parameter estimates (credible intervals ($q_{0.025}, q_{0.975}$)), *out-of-sample* prediction error and fitting times (hours) on the simulated dataset for the pooled and the single models.

Appendix B

The Spatio-Temporal Hawkes process on Rome car accidents

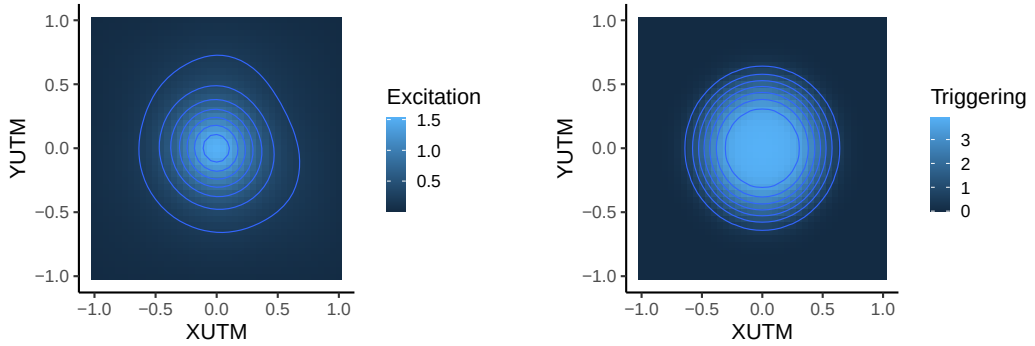
B.1 Examples of Boundary Correction

In the main text and in the previous section to this, we discussed methods of boundary correction. Specifically, we use mirrored correction when the domain is truncated, and for periodic domains we include influence from points lying one period either side of the domain in question. It can be useful to visualize what impact such methods are actually having on an estimate, so we generate simulated data to show exactly how they impact the estimates. In Fig. B.1, we show examples of two datasets. In Fig. B.1a, we show a data-set truncated at 0, and fit two density estimates to it: one with and one without mirrored boundary correction. Data are generated from a half normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$. Each density estimate is fit with a bandwidth of $b = 0.75$. In Fig. B.1b, we show a dataset generated on the domain $[0, 10]$, and imagine that we are trying to construct a periodic function from it, with $[0, 10]$ being the reference period. Data are generated from a gamma distribution with shape $\alpha = 5$ and scale $\beta = 1$. Each density estimate is fit with a bandwidth of $b = 0.5$. Again, we show an example of a density fit without any boundary correction, and one with the additional periodic components included.



(a) Mirrored correction example. **(b)** Periodic correction example

Figure B.1. Examples of boundary corrections. We show the data as a histogram, the non-corrected fits with (····) and the boundary corrected fits with (- - -).



(a) Anisotropic **(b)** Isotropic

Figure B.2. Comparison of Anisotropic and Isotropic Triggering functions estimates

Inspecting Fig. B.1a, we see a clear drop in the estimate at the truncation point of 0 without correction. Additionally, from Fig. B.1b, we see again a drop at the two ends of the periodic domain without any correction. These are removed when we apply our correction methods.

B.2 Isotropic and anisotropic excitation

Here, we want to check how much the results would have varied if the more general anisotropic form was considered for the spatial excitation function. First of all, we compare the two alternatives in terms of overall fit on the data. Again, this is quantified in terms of the log-likelihood, whose values are reported in Table B.1 Surprisingly, the anisotropic excitation yields a lower value of the log-likelihood, exhibiting a worse fit on the data. Nevertheless, comparing it to the other models in Table 5.1, we can see how it still provides better fits than the non-triggering models.

The excitation component magnitude, represented by the parameter A, is slightly reduced. Apparently, the euclidean distance based excitation (i.e. isotropic) fits better the triggering caused by road accidents and provides a better explanation of this dynamic. Comparing the two estimate spatial excitation in Figure B.2, we notice how the isotropic presents a more peaked behavior at low distances (near 0). The anisotropic instead is flatter. Aside from that, the two functions in Figures B.2a and B.2b, look really alike. Deviations of the anisotropic version from the isotropic mainly arise because the first is defined on the unit square instead of the unit circle. Moreover, the isotropic obviously represents a perfectly radial decay. The anisotropic seems to approximate such behavior and, by construction, cannot possibly achieve it (unless a very refined grid is chosen).

Therefore, the improvement in fit may be explained by the actual presence of a radial decay in the road accidents dynamic, which would sound completely natural. In addition, the isotropic model specification is way more parsimonious than the

Model	$\hat{\mu}$	$\hat{A}$	log-likelihood
Anisotropic	0.1591	0.038	−17264.33
Isotropic	0.1582	0.0592	−17216.65

Table B.1. Log-likelihood values for the isotropic and anisotropic models.

anisotropic one (the spatial excitation function has less freedom to adapt) and shall reassure us against over-fitting issues.

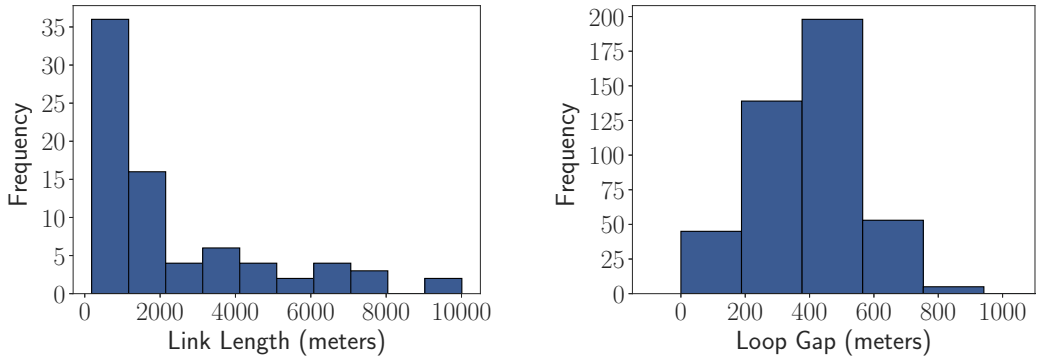
Appendix C

The Spatio-Temporal Hawkes process on NTIS car accidents

C.1 Problems With Link Specification of Events

To understand why we can localize events, we detail the two scales of data being provided by NTIS. The first is at the loop-level, with time-series of speed, flow and occupancy being reported each minute for each sensor in the network. The second is at the link-level, where aggregated data from all sensors that lie on a link is provided each minute, specifically speed, flow, occupancy and travel time. We extract both of these sets of data across our studied domain, and apply a 5-minute rolling average to smooth out the time-series.

When an event is recorded on our network, we are told which link it has occurred on, however NTIS links vary vastly in size, with the distribution of link lengths given in Fig. C.1a, along with the distribution of gaps between successive loop sensors across the entire network in Fig. C.1b.



(a) Distribution of link lengths through the M25.

(b) Distribution of gaps between loop sensors on the M25.

Figure C.1. Comparison of link lengths and distance between successive loop sensors. We see that there is wide range of link lengths, with a small number being close to 10 kilometers in length. On the other hand, distances between successive loop sensors are highly concentrated between 0 and 800 meters.

The wide range of link lengths shown in Fig. C.1a clearly indicate that further localization needs to be performed to narrow down a specific point on the link before

we can apply point-process methodology. However, there is hope that the loop sensors are fine-grained enough in space to provide an estimate of this, with Fig. C.1b showing that any point in the network should have a loop sensor less than 400 meters away. This is a far higher level of detail than initially provided by NTIS events, so we decide to refine our data further by using the time-series provided by the loop sensors. This is not a trivial task however as we do not have labeled data on which we can train a model, and instead must define sensible criteria that indicate the location of a traffic accident. This is an extensive task, and detailed in section C.2

C.2 Event Localization Methodology

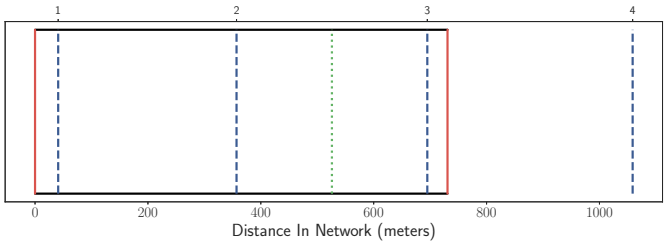
We note from the outset that detection of traffic events in both space and time is an active research field, and various approaches are being considered using an increasing variety of data sources. We are in an atypical situation in that we know a temporal and spatial window in which an event occurred, and are only searching for a more fine-grained spatial location within this window. As such, we do not aim to completely solve the event detection problem using inductive loop data, rather we try and develop an effective methodology to take a given window with a known event in and argue what single set of sensors the event may lie between. From discussions with industry experts, we consider the following properties to be clear signatures of a significant traffic event:

a) upstream of an event location, speed will be decreased and occupancy will be increased compared to the seasonal values

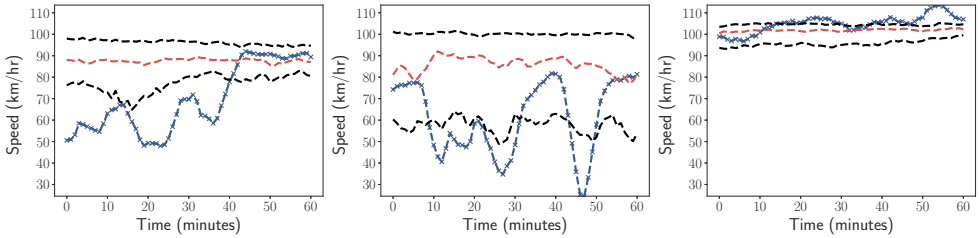
b) downstream of an event location, speed may still be decreased, but less so than upstream of the accident

c) downstream of an event location, occupancy will be decreased compared to the seasonal values

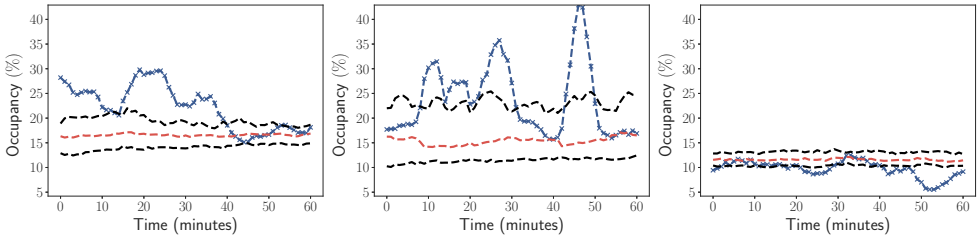
Given the industry expert criteria, we first develop simple seasonal models for each loop sensor in our network. There is clear seasonality on the weekly scale in traffic data, with commuting days in the U.K. being Monday to Friday, and Saturday and Sunday having less vehicles on the road. Taking this as the leading seasonal component, we construct a simple seasonal model by taking all data on a given weekday at a given time of day, and then finding the median value, using this as a reasonable seasonal estimate. Doing so, we have a model for each weekday, at the minute level, fit to each loops data separately. We produce one such seasonal model for each traffic variable on a loop. We then consider what is reasonable to develop with our available data. Ideally, we would design and validate some localization methodology incorporating spatial-temporal information from the loop data, inferring a location from the behaviour of all loops. However, we have no data with the ground truth locations, so we cannot reasonably develop such a model. Instead, we can use a simple 'rule of thumb' approach based on existing methodology. Numerous historic methodologies in traffic theory Payne et al. (1976), Gall and Hall (1989) compare adjacent sensors to determine two points, one where the data appears to show an accident, and another where the data does not. We can do the same, and given we know there must be an event in the given window, we can simply ask at what point do we see the largest discrepancy for two adjacent sensors. First, we consider any two sensors in our network, $i-1$ and i for all $i \in \{2, \cdots, N\}$. We



(a) Simple Diagram of the link with sensors marked (- - -), the localized event location (····), and the start and end of the link (——).



(b) Loop 1 - speed **(c)** Loop 2 - speed **(d)** Loop 3 - speed



(e) Loop 1 - occupancy **(f)** Loop 2 - occupancy **(g)** Loop 3 - occupancy

Figure C.2. An example result of our localization procedure. We plot the data (- - -), seasonal median (- - -) and the 20th and 80th percentiles for the particular traffic variable (- - -).

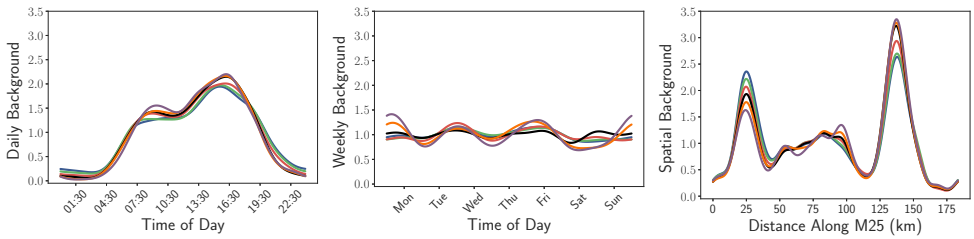
denote the residual series for speed and occupancy on a sensor i as RS_i and RO_i respectively. We then compute the spatially differenced values:

$$\Delta RS_i = RS_i - RS_{i-1}, \quad \Delta RO_i = RO_i - RO_{i-1}. \tag{C.2.1}$$

We then define an 'event impact score' between the two loops as:

$$EIS_i = \Delta RS_i - \Delta RO_i. \tag{C.2.2}$$

Since ΔRS_i will likely be positive and ΔRO_i will likely be negative if an event occurs between $i-1$ and i, then we look for the pair of loops at which the value of EIS_i is largest, and place our localized event halfway between these two loops. There is of course huge scope to improve upon such a model, but without data to validate more advanced approaches it is sufficient to provide a method that agrees with existing literature and common sense rules. Figure C.2 shows an example of localizing an event based on our simple methodology applied to loop sensor data. The contained plots give a sense of how much variation there is in the data. This is the first link in our network, so we show the next loop sensor along for a sense of scale. We see that sensors 1 and 2 have large drops in speed and increases in occupancy, however



(a) Daily Background **(b)** Weekly Background **(c)** Spatial Background

Figure C.3. Background components compared across significant events, varying what speed decrease is required to define an event as significant. Thresholds are: 0%: —, 10%: —, 20%: —, 30%: —, 40%: — and 50%: —.

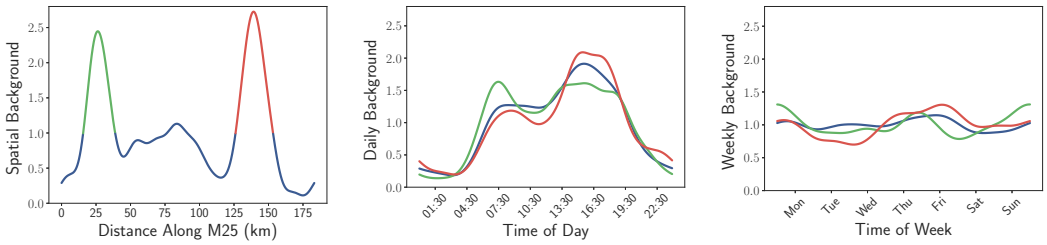
sensor 3 appears reasonably seasonal. Our methodology has then placed the event between sensors 2 and 3.

C.3 Additional analyses

C.3.1 Do Components Change for Significant Events?

Whilst all events flagged in NTIS should correspond to actual traffic accidents, many of them may not have had a significant impact on the traffic state. Consider the case where two vehicles have a minor collision, create no debris, and the drivers pull into the closed hard-shoulder to exchange insurance information. In such a case, there should be little impact on flow, travel time and speed. Additionally, if a vehicle breaks down and pulls into the hard-shoulder, and the road is far from capacity, we should again see little drop in average speed. To consider only the behaviour of events that have some significant impact, we inspect the link-level data, which contains significantly less noise than the loop-level. For a given event window, we consider the largest percentage drop in speed between a simple historical median segmentation profile and measured values across the entire window. If this percentage drop is above some threshold, then we say the event caused a significant impact on the traffic state. Further discussion of this seasonal model is given in the supplementary material. As we raise this threshold, we isolate more extreme events but discard so much data that it is no longer reasonable to fit a model. To retain enough data for fitting, we consider only thresholds between 0 and 50%. We split our dataset into subsets containing only events that lead to a speed decrease of at-least 0%, 10%, ..., 50%, then re-run our model fitting on these subsets. The resulting background components are visualized in Fig. C.3.

Fig. C.3a shows the daily background is reasonably stable across different thresholds of significance. As we raise the threshold, the morning and evening peak structure becomes clearer, and the periods very early and late in the day are lowered. Fig. C.3b shows that weekends have lower intensity than weekdays as we raise the threshold for significance. This is likely due to demand being significantly lower on weekends compared to weekdays, and hence when an event does occur, there is less chance of an queue forming as the road is further from capacity than on a weekday. Finally, the spatial background in Fig. C.3c clearly shows that the second peak, identified around 140 kilometres along the M25, appears to experience more significantly impactful events than the first peak, observed around 25 kilometres along the M25. This suggests that not only is Potters Bar a 'hot-spot' for events, but also that the events here are some of the more extreme on the network. Analysis



(a) Visualization of hotspot locations. **(b)** Comparison of daily backgrounds. **(c)** Comparison of weekly backgrounds.

Figure C.4. Identification of hotspot locations, and comparison of temporal background components around them. In C.4a, we show the spatial locations of hotspot 1 (—) and hotspot 2 (—). In C.4b and C.4c, we show the results for the entire dataset (—), hotspot 1 (—) and hotspot 2 (—).

of triggering during different significance thresholds shows that the time-scales and A values remain consistent throughout. The only visible difference in the triggering functions is that for larger thresholds, the functions decay more quickly.

C.3.2 Temporal Background Analysis Around Peaks in Spatial Intensity

Whilst we have seen that the background component of our model has two clear peaks in it, we can also question if the daily and weekly background components differ significantly in the vicinity of these peaks, compared to their behaviour across the entire motorway. To investigate this, we consider each peak separately. Firstly, we define two 'hotspots' surrounding the two peaks in spatial intensity. These locations start where the spatial background component has a value above 1, and end where it then falls back below 1. We then isolate a set of events that occur in each spatial hotspot. For the form of our model, this suggests we are isolating the spatial location where there is a increase in the rate of events compared to the average spatial location. We then re-fit our model using only this subset of events, and question what resulting daily and weekly background components arise. We visualize our results in Fig. C.4. Specifically, we visualize the subsets by location in Fig. C.4a, the daily background in Fig. C.4b and the weekly background in C.4c.

From Fig. C.4, it is clear that the temporal background around each hotspot is reasonably similar to that across the entire M25. The daily background around the first hotspot has a slightly more pronounced peak in the morning and lower peak in the evening compared to the entire network. The second hotspot has a slightly higher peak in the evening. Both hotspots have a similar weekly component. Overall, it seems reasonable to conclude that behaviour at these two hotspots is not fundamentally different to that across the entire motorway.

Bibliography

Abdalla, N., Banerjee, S., Ramachandran, G., Stenzel, M., and Stewart, P. A. (2018). Coastline kriging: a bayesian approach. *Annals of work exposures and health*, 62(7):818–827.

Abel, M., Hannon, J., Mullineaux, D., and Beighle, A. (2011). Determination of step rate thresholds corresponding to physical activity intensity classifications in adults. *Journal of Physical Activity and Health*, 8(1):45–51.

Aberg, S., Lindgren, F., Malmberg, A., Holst, J., and Holst, U. (2005). An image warping approach to spatio-temporal modelling. *Environmetrics: The official journal of the International Environmetrics Society*, 16(8):833–848.

ACI and ISTAT (2019). ROAD ACCIDENTS: Year 2019. Technical report, Italian government.

Acker, B. and Yuan, M. (2019). Network-based likelihood modeling of event occurrences in space and time: a case study of traffic accidents in Dallas, Texas, USA. *Cartography and Geographic Information Science*, 46(1):21–38.

Aguero-Valverde, J. and Jovanis, P. P. (2008). Analysis of road crash frequency with spatial models. *Transp. Res. Rec.*, 2061(1):55–63.

Aguilar-Farias, N., Peeters, G., Brychta, R. J., Chen, K. Y., and Brown, W. J. (2019). Comparing actigraph equations for estimating energy expenditure in older adults. *Journal of sports sciences*, 37(2):188–195.

Alaimo Di Loro, P. (2016). Exact Approximate MCMC for Hidden Markov models: An Application to Car Accidents. Master's thesis, University of Rome La Sapienza.

Alaimo Di Loro, P., Ciminello, E., and Tardella, L. (2019). Hidden Markov Model estimation via Particle Gibbs. *Book of Short Papers SIS 2019*, pages 829–835.

Alaimo Di Loro, P., Mingione, M., Lipsitt, J., Batteatte, M. C., Jerrett, M., and Banerjee, S. (2021). Efficient bayesian hierarchical modeling and analysis for physical activity trajectories using actigraph data. *arXiv preprint arXiv:2101.01624*.

Allcroft, D. J. and Glasbey, C. A. (2003). A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):487–498.

Anderes, E. and Chatterjee, S. (2009). Consistent estimates of deformed isotropic gaussian random fields on the plane. *The Annals of Statistics*, pages 2324–2350.

Armstrong, M., Chetboun, G., and Hubert, P. (1993). Kriging the rainfall in lesotho. In *Geostatistics Tróia'92*, pages 661–672. Springer.

Azizpour, S., Giesecke, K., and Schwenkler, G. (2018). Exploring the sources of default clustering. *Journal of Financial Economics*, 129(1):154–183.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns: (with discussion). *Australian & New Zealand Journal of Statistics*, 42(3):283–322.

Bai, Y., Song, P. X.-K., and Raghunathan, T. (2012). Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(5):799–824.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Bang-Jensen, J. and Gutin, G. Z. (2008). *Digraphs: theory, algorithms and applications*. Springer Science & Business Media.

Barbian, M. H. and Assunção, R. M. (2017). Spatial subsemble estimator for large geostatistical data. *Spatial Statistics*, 22:68–88.

Bauwens, L. and Hautsch, N. (2009). Modelling financial high frequency data using point processes. In *Handbook of financial time series*, pages 953–979. Springer.

Bde Boor, C. (2001). A practical guide to splines, revised edition.

Bedard, C., Bremer, E., and Cairney, J. (2020). Evaluation of the move 2 learn program, a community-based movement and pre-literacy intervention for young children. *Physical Education and Sport Pedagogy*, 25(1):101–117.

Berkemeyer, K., Wijndaele, K., White, T., Cooper, A., Luben, R., Westgate, K., Griffin, S., Khaw, K.-T., Wareham, N., and Brage, S. (2016). The descriptive epidemiology of accelerometer-measured physical activity in older adults. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1):2.

Berliner, L. M. (1996). Hierarchical bayesian time series models. In *Maximum entropy and Bayesian methods*, pages 15–22. Springer.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(1):143–155.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association*, 107(497):268–280.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Bilonick, R. A. (1985). The space-time distribution of sulfate deposition in the northeastern united states. *Atmospheric Environment (1967)*, 19(11):1829–1845.

Birkhoff, G. and De Boor, C. R. (1965). Piecewise polynomial interpolation and approximation. *Approximation of functions*, pages 164–190.

Björk, T., Kabanov, Y., and Runggaldier, W. (1997). Bond market structure in the presence of marked point processes. *Mathematical Finance*, 7(2):211–239.

Bochner, S. (1933). Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410.

Bochner, S. (2005). *Harmonic analysis and the theory of probability*. Courier Corporation.

Bradley, J. R., Cressie, N., Shi, T., et al. (2016a). A comparison of spatial predictors when datasets could be very large. *Statistics Surveys*, 10:100–131.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, 13(1):253–310.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association*, 115(532):2037–2052.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016b). Bayesian spatial change of support for count-valued survey data with application to the american community survey. *Journal of the American Statistical Association*, 111(514):472–487.

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear hawkes processes. *The Annals of Probability*, pages 1563–1588.

Breslow, N. E., Day, N. E., and Heseltine, E. (1980). *Statistical methods in cancer research*, volume 1. International agency for research on cancer Lyon.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis*, 50(4):967–991.

Brillinger, D. R. (1997). An application of statistics to meteorology: estimation of motion. In *Festschrift for Lucien Le Cam*, pages 93–105. Springer.

Brillinger, D. R. (2001). *Time series: data analysis and theory*. SIAM.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Comput.*, 14(2):325–346.

Brown, P. E., Roberts, G. O., Kåresen, K. F., and Tonellato, S. (2000). Blur-generated non-separable space–time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):847–860.

Buhmann, M. D. (2003). *Radial basis functions: theory and implementations*, volume 12. Cambridge university press.

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., Carty, C., Chaput, J.-P., Chastin, S., Chou, R., et al. (2020). World health organization 2020 guidelines on physical activity and sedentary behaviour. *British journal of sports medicine*, 54(24):1451–1462.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Cappé, O., Moulines, E., and Rydén, T. (2006). *Inference in hidden Markov models*. Springer Science & Business Media.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Chiu, S.-T. (2000). Boundary adjusted density estimation and bandwidth selection. *Stat. Sin.*, 10(4):1345–1367.

Christakos, G. and Hristopulos, D. (1998). Tractatus stochasticus: The study of spatiotemporal environmental health processes.

Christakos, G., Hristopulos, D., and Bogaert, P. (2000). On the physical geometry concept at the basis of space/time geostatistical hydrology. *Advances in Water Resources*, 23(8):799–810.

Comi, A., Persia, L., Nuzzolo, A., and Polimeni, A. (2018). Exploring temporal and spatial structure of urban road accidents: Some empirical evidences from rome. In *The 4th Conference on Sustainable Urban Mobility*, pages 147–155. Springer.

Cormack, R. (1979). Spatial aspects of competition between individuals. *Spatial and temporal analysis in ecology.*, pages 151–212.

Cottam, G. and Curtis, J. T. (1949). A method for making rapid surveys of woodlands by means of pairs of randomly selected trees. *Ecology*, 30(1):101–104.

Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3):551–563.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157.

Cox, D. R. and Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 415(1849):317–328.

Cressie, N. (1993). Statistics for spatial data. revised edition. *New Yorlg A Wiley-Interscience Publicatiorq*.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339.

Cressie, N. and Lahiri, S. N. (1993). The asymptotic distribution of reml estimators. *Journal of multivariate analysis*, 45(2):217–233.

Cressie, N., Shi, T., and Kang, E. L. (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Cressie, N. A. and Johannesson, G. (2006). Spatial prediction for massive datasets. *Mastering the Data Explosion in the Earth and Environmental Sciences*.

Crouter, S. E., Clowers, K. G., and Bassett Jr, D. R. (2006). A novel method for using accelerometer data to predict energy expenditure. *Journal of applied physiology*, 100(4):1324–1331.

Currie, I. D., Durban, M., and Eilers, P. H. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):259–280.

Daley, D. J. and Vere-Jones, D. (2003). An introduction to the theory of point processes. Vol. I. Probability and its Applications.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016b). Non-separable dynamic nearest-neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Annals of Applied Statistics*, 10:1286–1316.

De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

De Iaco, S., Myers, D. E., and Posa, D. (2001). Space–time analysis using a general product–sum model. *Statistics & Probability Letters*, 52(1):21–28.

De Luna, X. and Genton, M. G. (2005). Predictive spatio-temporal models for spatially sparse enviromental data. *Statistica Sinica*, pages 547–568.

Degroote, L., Hamerlinck, G., Poels, K., Maher, C., Crombez, G., De Bourdeaudhuij, I., Vandendriessche, A., Curtis, R. G., and DeSmet, A. (2020). Low-cost consumer-based trackers to measure physical activity and sleep duration among adults in free-living conditions: Validation study. *JMIR mHealth and uHealth*, 8(5):e16674.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Department for Transport (2019). Reported road casualties in Great Britain: main results 2018. Technical report, UK government.

DiFranzo, A., Sheridan, R. P., Liaw, A., and Tudor, M. (2020). Nearest neighbor gaussian process for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 60(10):4653–4663.

Diggle, P. and Ribeiro Jr, P. (2007). Model-based geostatistics. *Springer series in Statistics*.

Diggle, P. J. (1986). Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. *Journal of neuroscience methods*, 18(1-2):115–125.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

Diggle, P. J., Besag, J., and Gleaves, J. T. (1976). Statistical analysis of spatial point patterns by means of distance methods. *Biometrics*, pages 659–667.

Diggle, P. J. and Cox, T. F. (1983). Some distance-based tests of independence for sparsely-sampled multivariate spatial point patterns. *International Statistical Review/Revue Internationale de Statistique*, pages 11–23.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Dimitrakopoulos, R. and Luo, X. (1994). Spatiotemporal modelling: covariances and ordinary kriging systems. In *Geostatistics for the next century*, pages 88–93. Springer.

Dimitrakopoulos, R. and Luo, X. (1997). Joint space-time modeling in the presence of trends. *Geostatistics wollongong*, 96:138–149.

Dixon, P. M. (2014). Ripley's k function. *Wiley StatsRef: Statistics Reference Online*.

Dobruschin, P. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Van Hees, V. T., Trenell, M. I., Owen, C. G., et al. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one*, 12(2).

Dominici, G. (2015). Per una seconda fase degli open data in italia. *PRISMA Economia-Società-Lavoro*.

Drewnowski, A., Buszkiewicz, J., Aggarwal, A., Rose, C., Gupta, S., and Bradshaw, A. (2020). Obesity and the built environment: A reappraisal. *Obesity*, 28(1):22–30.

Du Rietz, G. E. (1929). *The fundamental units of vegetation*.

Dunton, G. F., Almanza, E., Jerrett, M., Wolch, J., and Pentz, M. A. (2014). Neighborhood park use by children: use of accelerometry and global positioning systems. *American journal of preventive medicine*, 46(2):136–142.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315.

Eilers, P. H., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Eldrandaly, K. and Abdelmouty, A. (2017). Spatio-temporal interpolation: Current practices and future prospects. *International Journal of Digital Content Technology and its Applications*, 11(06):2017.

Faghmous, J. H. and Kumar, V. (2014). Spatio-temporal data mining for climate data: Advances, challenges, and opportunities. In *Data mining and knowledge discovery for big data*, pages 83–116. Springer.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica*, pages 731–761.

Fan, Y., Zhu, X., She, B., Guo, W., and Guo, T. (2018). Network-constrained spatio-temporal clustering analysis of traffic collisions in Jianghan District of Wuhan, China. *PLOS ONE*, 13(4):1–23.

Farooq, A., Martin, A., Janssen, X., Wilson, M. G., Gibson, A.-M., Hughes, A., and Reilly, J. J. (2020). Longitudinal changes in moderate-to-vigorous-intensity physical activity in children and adolescents: A systematic review and meta-analysis. *Obesity Reviews*, 21(1):e12953.

Finley, A., Datta, A., and Banerjee, S. (2017a). spnngp: spatial regression models for large datasets using nearest neighbor gaussian processes. *R package version 0.1*, 1.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spbayes: an r package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of statistical software*, 19(4):1.

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2012). Bayesian dynamic modeling for large space-time datasets using gaussian predictive processes. *Journal of geographical systems*, 14(1):29–47.

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2013). spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *arXiv preprint arXiv:1310.8192*.

Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (2017b). Applying nearest neighbor gaussian processes to massive spatial data sets forest canopy height prediction across tanana valley alaska.". *arXiv preprint arXiv:1702.00434*.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, pages 1–14.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884.

Fisher, R. A. et al. (1960). The design of experiments. *The design of experiments.*, (7th Ed).

Fox, E. W., Schoenberg, F. P., Gordon, J. S., et al. (2016a). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics*, 10(3):1725–1756.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., and Bertozzi, A. L. (2016b). Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Am. Statist. Ass.*, 111(514):564–584.

Freedson, P., Bowles, H. R., Troiano, R., and Haskell, W. (2012). Assessment of physical activity using wearable monitors: Recommendations for monitor calibration and use in the field. *Medicine and science in sports and exercise*, 44(1 Suppl 1):S1.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102(477):321–331.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Furrer, R., Sain, S. R., et al. (2010). spam: A sparse matrix r package with emphasis on mcmc methods for gaussian markov random fields. *Journal of Statistical Software*, 36(10):1–25.

Gall, A. I. and Hall, F. L. (1989). Distinguishing between incident congestion and recurrent congestion: a proposed logic. *Transportation Research Record*, (1232).

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gerrard, D. J. et al. (1969). New measure of the competition affecting individual forest trees.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving mcmc. *Markov chain Monte Carlo in practice*, 6:89–114.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2):493–508.

Gneiting, T., Genton, M. G., and Guttorp, P. (2006). Geostatistical space-time models, stationarity, separability, and full symmetry. *Monographs On Statistics and Applied Probability*, 107:151.

Gneiting, T. and Guttorp, P. (2010). Continuous parameter spatio-temporal processes. *Handbook of Spatial Statistics*, 97:427–436.

Goodman, T. and Hardin, D. (2006). Refinable multivariate spline functions. In *Studies in Computational Mathematics*, volume 12, pages 55–83. Elsevier.

Gramacy, R. B. and Apley, D. W. (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press.

Grenier, I. and Sansó, B. (2019). Distributed implementation of nearest-neighbor gaussian processes.

Grimmet, G. S. et al. (2020). *Probability and random processes*. Oxford university press.

Gu, C. (2002). Smoothing spline anova models springer-verlag. *New York*.

Guennebaud, G., Jacob, B., et al. (2010). Eigen. *URl: http://eigen. tuxfamily. org*.

Guhaniyogi, R. and Banerjee, S. (2018). Meta-kriging: Scalable bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4):430–444.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, 90(432):1189–1199.

Hall, K. S., Howe, C. A., Rana, S. R., Martin, C. L., and Morey, M. C. (2013). Mets and accelerometry of walking in older adults: Standard versus measured energy cost. *Medicine and science in sports and exercise*, 45(3):574.

Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.*, 29(3):624–647.

Hancock, P. A. and Hutchinson, M. (2006). Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. *Environmental Modelling & Software*, 21(12):1684–1694.

Hartman, S. J., Nelson, S. H., and Weiner, L. S. (2018). Patterns of fitbit use and activity levels throughout a physical activity intervention: exploratory analysis from a randomized controlled trial. *JMIR mHealth and uHealth*, 6(2):e29.

Hastie, T., Tibshirani, R., et al. (2000). Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). The elements of statistical learning.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Heidenreich, N.-B., Schindler, A., and Sperlich, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis*, 97(4):403–433.

Hidding, L. M., Chinapaw, M. J., van Poppel, M. N., Mokkink, L. B., and Altenburg, T. M. (2018). An updated systematic review of childhood physical activity questionnaires. *Sports Medicine*, 48(12):2797–2842.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer.

Highways England (2020). How to drive on a smart motorway.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, 42(33):7561–7578.

Hominal, P. and Deheuvels, P. (1979). Estimation non paramétrique de la densité compte-tenu d'informations sur le support. *Revue de statistique appliquée*, 27(3):47–68.

Huang, H.-C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics & Data Analysis*, 22(2):159–175.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Inés, M., González, M., Gutiérrez, C., Martínez, R., Minuesa, C., Molina, M., Mota, M., and Ramos, A. (2016). *Branching Processes and Their Applications*, volume 219. Springer.

INRIX (2019). INRIX Reveals the UK's Worst Traffic Jams Over the Past Year.

James, P., Jankowska, M., Marx, C., Hart, J. E., Berrigan, D., Kerr, J., Hurvitz, P. M., Hipp, J. A., and Laden, F. (2016). "spatial energetics": Integrating data from gps, accelerometry, and gis to address obesity and inactivity. *American Journal of Preventive Medicine*, 51(5):792–800.

Jánossy, L. (1950). On the absorption of a nucleon cascade. In *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, volume 53, pages 181–188. JSTOR.

Jie, L. K., Ramli, A., and Majid, A. A. (2016). A comparison study between b-spline surface fitting and radial basis function surface fitting on scattered points. *Jurnal Teknologi*, 78(6-5).

Johnson, D. H. (1996). Point process models of single-neuron discharges. *Journal of computational neuroscience*, 3(4):275–299.

Johnson, N., Hitchman, A., Phan, D., and Smith, L. (2018). Self-exciting point process models for political conflict forecasting. *European Journal of Applied Mathematics*, 29(4):685–707.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and computing*, 3(3):135–146.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In *Modelling longitudinal and spatially correlated data*, pages 289–298. Springer.

Kalair, K., Connaughton, C., and Alaimo Di Loro, P. (2020). A non-parametric hawkes process model of primary and secondary accidents on a uk smart motorway. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Kamada, M., Shiroma, E. J., Harris, T. B., and Lee, I.-M. (2016). Comparison of physical activity assessed using hip-and wrist-worn accelerometers. *Gait & posture*, 44:23–28.

Kang, E. L., Liu, D., and Cressie, N. (2009). Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics & Data Analysis*, 53(8):3016–3032.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.

Katzfuss, M. and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics*, 23(1):94–107.

Katzfuss, M. and Guinness, J. (2021). A general framework for vecchia approximations of gaussian processes. *Statist. Sci.*, 36(1):124–141.

Katzfuss, M., Guinness, J., Gong, W., and Zilber, D. (2020). Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics*, 25:383–414.

Katzfuss, M. and Hammerling, D. (2017). Parallel inference for massive distributed spatial data using low-rank models. *Statistics and Computing*, 27(2):363–375.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

Kay, S. M. (1993). *Fundamentals of statistical signal processing*. Prentice Hall PTR.

Kelbert, M. Y., Leonenko, N. N., and Ruiz-Medina, M. (2005). Fractional random fields associated with stochastic fractional heat equations. *Advances in Applied Probability*, 37(1):108–133.

Kestens, Y., Wasfi, R., Naud, A., and Chaix, B. (2017). "contextualizing context": Reconciling environmental exposures, social networks, and location preferences in health research. *Current Environmental Health Reports*, 4:51–60.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space–time data: A mixed model approach. *Biometrics*, 62(1):109–118.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Kulldorff, M. (1999). Spatial scan statistics: models, calculations, and applications. In *Scan statistics and applications*, pages 303–322. Springer.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *J. R. Statist. Soc. A.*, 164(1):61–72.

Kuo, H. (2006). Introduction to stochastic integration springer. *Berlin Heidelberg*.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space–time models: a review. *Mathematical geology*, 31(6):651–684.

La Torre, G., Van Beeck, E., Bertazzoni, G., and Ricciardi, W. (2007). Head injury resulting from scooter accidents in rome: differences before and after implementing a universal helmet law. *European journal of public health*, 17(6):607–611.

Lam, N. S.-N. (1983). Spatial interpolation methods: a review. *The American Cartographer*, 10(2):129–150.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lasinio, G. J., Mastrantonio, G., and Pollice, A. (2013). Discussing the "big n problem". *Statistical Methods & Applications*, 22(1):97–112.

Last, G. and Brandt, A. (1995). *Marked Point Processes on the real line: the dynamical approach*. Springer Science & Business Media.

Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes processes. *arXiv preprint arXiv:1507.02822*.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press.

Lee, D.-J. and Durbán, M. (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical modelling*, 11(1):49–69.

Lemos, R. T. and Sansó, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association*, 104(485):5–18.

Lesage, L. (2020). A Hawkes process to make aware people of the severity of COVID-19 outbreak: application to cases in France. Research report, Université de Lorraine ; University of Luxembourg.

Lewis, P. A. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (mars). *Journal of the American Statistical Association*, 86(416):864–877.

Li, L. and Revesz, P. (2002). A comparison of spatio-temporal interpolation methods. In *International Conference on Geographic Information Science*, pages 145–160. Springer.

Li, L. and Revesz, P. (2004). Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems*, 28(3):201–227.

Li, L., Tong, W., and Piltner, R. (2020). Spatiotemporal interpolation methods for air pollution. In *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health*, pages 153–167. Elsevier.

Li, Z., Cui, L., and Chen, J. (2018). Traffic accident modelling via self-exciting point processes. *Reliab. Eng. Syst. Safe.*, 180:312 – 320.

Lim, K. W., Wang, W., Nguyen, H., Lee, Y., Cai, C., and Chen, F. (2016). Traffic flow modelling with point processes. In *Proceedings of the 23rd World Congress on Intelligent Transport Systems*, pages 1–12. ITS.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, 21(3):411–433.

Lippiello, E., Giacco, F., Arcangelis, L. d., Marzocchi, W., and Godano, C. (2014). Parameter estimation in the etas model: Approximations and novel methods. *Bulletin of the Seismological Society of America*, 104(2):985–994.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). When gaussian process meets big data: A review of scalable gps. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

Loader, C. and Switzer, P. (1992). Spatial covariance estimation for monitoring data. *Statistics in the Environmental and Earth Sciences*, pages 52–70.

Loeffler, C. and Flaxman, S. (2018). Is gun violence contagious? a spatiotemporal test. *Journal of quantitative criminology*, 34(4):999–1017.

Lyden, K., Keadle, S. K., Staudenmayer, J., and Freedson, P. S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Medicine and science in sports and exercise*, 46(2):386.

Lyden, K., Kozey, S. L., Staudenmeyer, J. W., and Freedson, P. S. (2011). A comprehensive evaluation of commonly used accelerometer energy expenditure and met prediction equations. *European journal of applied physiology*, 111(2):187–201.

Ma, C. (2002). Spatio-temporal covariance functions generated by mixtures. *Mathematical geology*, 34(8):965–975.

Ma, C. (2008). Recent developments on the construction of spatio-temporal covariance models. *Stochastic Environmental Research and Risk Assessment*, 22(1):39–47.

MacNab, Y. C. and Dean, C. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57(3):949–956.

Maher, C., Virgara, R., Okely, T., Stanley, R., Watson, M., and Lewis, L. (2019). Physical activity and screen time in out of school hours care: an observational study. *BMC pediatrics*, 19(1):1–10.

Mardia, K. V. and Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate environmental statistics*, 6(76):347–385.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146.

Marsan, D. and Lengline, O. (2008). Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079.

Matérn, B. (1960). *Spatial variation: Stachastic models and their application to some problems in forest surveys and other sampling investigations= Stokastiska modeller och deras tillämpning på några problem i skogstaxering och andra samplingsundersökningar*. PhD thesis.

Mauser, W. and Prasch, M. (2015). *Regional Assessment of Global Change Impacts: The Project GLOWA-Danube*. Springer.

McCulloch, C. E. and Searle, S. R. (2001). Generalized, linear, and mixed models (wiley series in probability and statistics).

McKean, H. P. (1969). *Stochastic integrals*, volume 353. American Mathematical Soc.

Meliker, J. R. and Sloan, C. D. (2011). Spatio-temporal epidemiology: principles and opportunities. *Spatial and spatio-temporal epidemiology*, 2(1):1–9.

Mercer, W. and Hall, A. (1911). The experimental error of field trials. *The Journal of Agricultural Science*, 4(2):107–132.

Meyer, S., Elias, J., and Höhle, M. (2012). A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616.

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Nyström, C. D., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J. R., and Ortega, F. B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports medicine*, 47(9):1821–1845.

Miller, N. E., Strath, S. J., Swartz, A. M., and Cashin, S. E. (2010). Estimating absolute and relative physical activity intensity across age via accelerometry in adults. *Journal of aging and physical activity*, 18(2):158–170.

Mingione, M., Alaimo Di Loro, P., Farcomeni, A., Divino, F., Lovison, G., Lasinio, G. J., and Maruotti, A. (2021). Spatial modelling of covid-19 incident cases using richards' curve: an application to the italian regions. *arXiv preprint arXiv:2106.05067*.

Mitchell, J. A., Quante, M., Godbole, S., James, P., Hipp, J. A., Marinac, C. R., Mariani, S., Cespedes Feliciano, E. M., Glanz, K., Laden, F., et al. (2017). Variation in actigraphy-estimated rest-activity patterns by demographic factors. *Chronobiology international*, 34(8):1042–1056.

Mohler, G. (2014a). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast*, 30(3):491–497.

Mohler, G. (2014b). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

Montoye, A. H., Moore, R. W., Bowles, H. R., Korycinski, R., and Pfeiffer, K. A. (2018). Reporting accelerometer methods in physical activity intervention studies: A systematic review and recommendations for authors. *British journal of sports medicine*, 52(23):1507–1516.

Moradi, M. M. and Mateu, J. (2019). First- and second-order characteristics of spatio-temporal point processes on linear networks. *J. Comput. Graph. Stat.*, 0(0):1–21.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neal, P., Roberts, G., et al. (2006). Optimal scaling for partially updating mcmc algorithms. *The Annals of Applied Probability*, 16(2):475–515.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29.

Nürnberger, G. and Zeilfelder, F. (2000). Developments in bivariate spline interpolation. *Journal of Computational and Applied Mathematics*, 121(1-2):125–152.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599.

Oehlert, G. W. (1993). Regional trends in sulfate wet deposition. *Journal of the American Statistical Association*, 88(422):390–399.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. In *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507. Springer.

Omidi, M. and Mohammadzadeh, M. (2016). A new method to build spatio-temporal covariance functions: analysis of ozone data. *Statistical Papers*, 57(3):689–703.

Ott, A. E., Pate, R. R., Trost, S. G., Ward, D. S., and Saunders, R. (2000). The use of uniaxial and triaxial accelerometers to measure children's "free-play" physical activity. *Pediatric Exercise Science*, 12(4):360–370.

Paciorek, C. J., Lipshitz, B., Zhuo, W., Prabhat, ., Kaufman, C. G. G., and Thomas, R. C. (2015). Parallelizing gaussian process calculations in r. *Journal of Statistical Software, Articles*, 63(10):1–23.

Papadakis, J. (1937). Méthode statistique pour des expériences sur champ. *Bull. Inst. Amel. Plantes a Salonique*, 23:13–29.

Pate, R. R., O'neill, J. R., and Lobelo, F. (2008). The evolving definition of" sedentary". *Exercise and sport sciences reviews*, 36(4):173–178.

Payne, H. J., Helfenbein, E. D., and Knobel, H. C. (1976). Development and testing of incident detection algorithms. vol. 2, research methodology and detailed results. Technical Report FHWA-RD-76-20, Federal Highway Adminstration, Washington D.C.

Pearce, M., Strain, T., Kim, Y., Sharp, S. J., Westgate, K., Wijndaele, K., Gonzales, T., Wareham, N. J., and Brage, S. (2020). Estimating physical activity from self-reported behaviours in large-scale population studies using network harmonisation: findings from uk biobank and associations with disease outcomes. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1):1–13.

Pearson, E. S., Gosset, W. S., Plackett, R., and Barnard, G. A. (1990). *Student: a statistical biography of William Sealy Gosset*. Oxford University Press, USA.

Perry, G. L., Miller, B. P., and Enright, N. J. (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant ecology*, 187(1):59–82.

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020a). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 0(0):1–14.

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020b). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *arXiv preprint arXiv:2003.11208*.

Peterson, N. E., Sirard, J. R., Kulbok, P. A., DeBoer, M. D., and Erickson, J. M. (2015). Inclinometer validation and sedentary threshold evaluation in university students. *Research in nursing & health*, 38(6):492.

Pfeifer, P. E. and Deutsch, S. J. (1980). Independence and sphericity tests for the residuals of space-time arma models: independence and sphericity tests for the residuals. *Communications in Statistics-Simulation and Computation*, 9(5):533–549.

Pfeifer, P. E. and Jay Deutsch, S. (1980). Stationarity and invertibility regions for low order starma models: stationarity and invertibility regions. *Communications in Statistics-Simulation and Computation*, 9(5):551–562.

Piercy, K. L., Troiano, R. P., Ballard, R. M., Carlson, S. A., Fulton, J. E., Galuska, D. A., George, S. M., and Olson, R. D. (2018). The physical activity guidelines for americans. *Jama*, 320(19):2020–2028.

Plasqui, G. and Westerterp, K. R. (2007). Physical activity assessment with accelerometers: an evaluation against doubly labeled water. *Obesity*, 15(10):2371–2379.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, 6(1):7–11.

Porter, M. D., White, G., et al. (2012). Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124.

Prévôt, C. and Röckner, M. (2007). *A concise course on stochastic partial differential equations*, volume 1905. Springer.

Rachele, J. N., McPhail, S. M., Washington, T. L., and Cuddihy, T. F. (2012). Practical physical activity measurement in youth: a review of contemporary approaches. *World Journal of Pediatrics*, 8(3):207–216.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Rasmussen, J. G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642.

Reiner, M., Niermann, C., Jekauc, D., and Woll, A. (2013). Long-term health benefits of physical activity–a systematic review of longitudinal studies. *BMC public health*, 13(1):1–9.

Reinhart, A. et al. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318.

Rey, S. J. (2016). Space-time patterns of rank concordance: Local indicators of mobility association with application to spatial income inequality dynamics. *Ann. Am. Assoc. Geogr.*, 106(4):788–803.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Rodriguez-Alvarez, M. X., Boer, M. P., van Eeuwijk, F. A., and Eilers, P. H. (2018). Correcting for spatial heterogeneity in plant breeding experiments with p-splines. *Spatial Statistics*, 23:52–71.

Rodríguez-Iturbe, I. and Mejía, J. M. (1974). The design of rainfall networks in time and space. *Water Resources Research*, 10(4):713–728.

Ross, G. (2016). Bayesian estimation of the etas model for earthquake occurrences. *Preprint*.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Ryan, P. H. and LeMasters, G. K. (2007). A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation toxicology*, 19(sup1):127–133.

Sagelv, E. H., Hopstock, L. A., Johansson, J., Hansen, B. H., Brage, S., Horsch, A., Ekelund, U., and Morseth, B. (2020). Criterion validity of two physical activity and one sedentary time questionnaire against accelerometry in a large cohort of adults and older adults. *BMJ Open Sport & Exercise Medicine*, 6(1):e000661.

Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, pages 681–703.

Sanso, B. and Guenni, L. (1999). Venezuelan rainfall data analysed by using a bayesian space–time model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):345–362.

Santos-Lozano, A., Santin-Medeiros, F., Cardon, G., Torres-Luque, G., Bailon, R., Bergmeir, C., Ruiz, J. R., Lucía Mulas, A., Garatachea, N., et al. (2013). Actigraph gt3x: validation and determination of physical activity intensity cut points.

Sasaki, J. E., John, D., and Freedson, P. S. (2011). Validation and comparison of actigraph activity monitors. *Journal of science and medicine in sport*, 14(5):411–416.

Schoenberg, F. P. (2002). On rescaled poisson processes and the brownian bridge. *Annals of the Institute of Statistical Mathematics*, 54(2):445–457.

Schoenberg, F. P. (2013). Facilitated estimation of etas. *Bulletin of the Seismological Society of America*, 103(1):601–605.

Schoenberg, F. P., Hoffmann, M., and Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Ann. I. Stat. Math.*, 71(5):1271–1287.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. part b. on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):112–141.

Schoenberg, I. J. (1964). On interpolation by spline functions and its minimal properties. In *On Approximation Theory/Über Approximationstheorie*, pages 109–129. Springer.

Schoenberg, I. J. (1988). Spline functions and the problem of graduation. In *IJ Schoenberg Selected Papers*, pages 201–204. Springer.

Schuch, F., Vancampfort, D., Firth, J., Rosenbaum, S., Ward, P., Reichert, T., Bagatini, N. C., Bgeginski, R., and Stubbs, B. (2017). Physical activity and sedentary behavior in people with major depressive disorder: a systematic review and meta-analysis. *Journal of affective disorders*, 210:139–150.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and methods*, 14(5):1123–1136.

Segura, L. J., Zhao, G., Zhou, C., and Sun, H. (2020). Nearest neighbor gaussian process emulation for multi-dimensional array responses in freeze nano 3d printing of energy devices. *Journal of Computing and Information Science in Engineering*, 20(4).

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524.

Shewchuk, J. R. (1996). Triangle: Engineering a 2d quality mesh generator and delaunay triangulator. In *Workshop on Applied Computational Geometry*, pages 203–222. Springer.

Shirota, S., Finley, A. O., Cook, B. D., and Banerjee, S. (2019). Conjugate nearest neighbor gaussian process models for efficient statistical interpolation of large spatial data. *arXiv preprint arXiv:1907.10109*.

Sikka, R. S., Baer, M., Raja, A., Stuart, M., and Tompkins, M. (2019). Analytics in sports medicine: implications and responsibilities that accompany the era of big data. *JBJS*, 101(3):276–283.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Simpson, D., Lindgren, F., and Rue, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23(1):65–74.

Smuck, M., Kao, M.-C. J., Brar, N., Martinez-Ith, A., Choi, J., and Tomkins-Lane, C. C. (2014). Does physical activity influence the relationship between low back pain and obesity? *The Spine Journal*, 14(2):209–216.

Song, J., Wen, R., and Yan, W. (2018). Identification of traffic accident clusters using Kulldorff's space-time scan statistics. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3162–3167.

Spiegel, E., Kneib, T., and Otto-Sobotka, F. (2020). Spatio-temporal expectile regression models. *Statistical Modelling*, 20(4):386–409.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association*, 100(469):310–321.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19.

Stein, M. L., Chen, J., Anitescu, M., et al. (2013). Stochastic approximation of score functions for gaussian processes. *The Annals of Applied Statistics*, 7(2):1162–1191.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

Stoffer, D. S. (1986). Estimation and identification of space-time armax models in the presence of missing data. *Journal of the American Statistical Association*, 81(395):762–772.

Strath, S. J., Kaminsky, L. A., Ainsworth, B. E., Ekelund, U., Freedson, P. S., Gary, R. A., Richardson, C. R., Smith, D. T., and Swartz, A. M. (2013). Guide to the assessment of physical activity: clinical and research applications: a scientific statement from the american heart association. *Circulation*, 128(20):2259–2279.

Stroud, J. R., Müller, P., and Sansó, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):673–689.

Stroud, J. R., Stein, M. L., and Lysen, S. (2017). Bayesian and maximum likelihood estimation for gaussian processes on an incomplete lattice. *Journal of computational and Graphical Statistics*, 26(1):108–120.

Sun, Y., Li, B., and Genton, M. G. (2012). Geostatistics for large datasets. In *Advances and challenges in space-time modelling of natural events*, pages 55–77. Springer.

Sylvia, L. G., Bernstein, E. E., Hubbard, J. L., Keating, L., and Anderson, E. J. (2014). Practical guide to measuring physical activity. *Journal of the Academy of Nutrition and Dietetics*, 114(2):199–208.

Tang, W., Zhang, L., and Banerjee, S. (2019). On identifiability and consistency of the nugget in gaussian spatial process models. *arXiv preprint arXiv:1908.05726*.

Taraldsen, K., Chastin, S. F., Riphagen, I. I., Vereijken, B., and Helbostad, J. L. (2012). Physical activity monitoring by use of accelerometer-based body-worn sensors in older adults: A systematic literature review of current knowledge and applications. *Maturitas*, 71(1):13–19.

Tonellato, S. (1997). Bayesian dynamic linear models for spatial time series. *Rapporto di Ricerca 5/1997*.

Toth, C. D., O'Rourke, J., and Goodman, J. E. (2017). *Handbook of discrete and computational geometry*. CRC press.

Troiano, R. P., McClain, J. J., Brychta, R. J., and Chen, K. Y. (2014). Evolution of accelerometer methods for physical activity research. *Br J Sports Med*, 48(13):1019–1023.

Troped, P. J., Wilson, J. S., Matthews, C. E., Cromley, E. K., and Melly, S. J. (2010). The built environment and location-based physical activity. *American journal of preventive medicine*, 38(4):429–438.

Trost, S. G. (2007). State of the art reviews: measurement of physical activity in children and adolescents. *American Journal of lifestyle medicine*, 1(4):299–314.

Tudor-Locke, C., Brashear, M. M., Johnson, W. D., and Katzmarzyk, P. T. (2010). Accelerometer profiles of physical activity and inactivity in normal weight, overweight, and obese us men and women. *International Journal of Behavioral Nutrition and Physical Activity*, 7(1):60.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Vyas, V. M. and Christakos, G. (1997). Spatiotemporal analysis and mapping of sulfate deposition data over eastern usa. *Atmospheric Environment*, 31(21):3623–3633.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wang, H. and Wang, J. (2009). Estimation of the trend function for spatio-temporal models. *Journal of Nonparametric Statistics*, 21(5):567–588.

West, M. and Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Westerterp, K. R. (2009). Assessment of physical activity: a critical appraisal. *European journal of applied physiology*, 105(6):823–828.

Wikle, C. K. and Berliner, L. M. (2005). Combining information across spatial scales. *Technometrics*, 47(1):80–91.

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86(4):815–829.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatiotemporal hierarchical bayesian modeling tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454):382–397.

Wing, M. G., Eklund, A., and Kellogg, L. D. (2005). Consumer-grade global positioning system (gps) accuracy and reliability. *Journal of forestry*, 103(4):169–173.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

World Health Organization (2018). *Global status report on road safety 2018*. World Health Organization.

Xiao, Y., Gu, X., Yin, S., Shao, J., Cui, Y., Zhang, Q., and Niu, Y. (2016). Geostatistical interpolation model selection based on arcgis and spatio-temporal variability analysis of groundwater level in piedmont plains, northwest china. *SpringerPlus*, 5(1):1–15.

Xu, G., Liang, F., and Genton, M. G. (2015). A bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. *Statistica Sinica*, pages 61–79.

Yanosky, J. D., Paciorek, C. J., Schwartz, J., Laden, F., Puett, R., and Suh, H. H. (2008). Spatio-temporal modeling of chronic pm10 exposure for the nurses' health study. *Atmospheric Environment*, 42(18):4047–4062.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921–936.

Zhu, L. (2013). Central limit theorem for nonlinear hawkes processes. *Journal of Applied Probability*, 50(3):760–771.

Zhuang, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):635–653.

Zhuang, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, Planets and Space*, 63(3):5.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(3):919–942.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, 109(B5).

Zienkiewicz, O. and Taylor, R. (1989). The finite element method, 4th edn., vol. 1. *Basic Formulation and Linear*.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453–470.